**UNIVERSIDAD POLITÉCNICA DE MADRID**

**Escuela Técnica Superior de Ingenieros Informáticos**

# NEW METHODOLOGY FOR THE INTEGRATION OF BIOMETRIC FEATURES IN SPEAKER RECOGNITION SYSTEMS APPLIED TO SECURITY ENVIRONMENTS

## DOCTORAL THESIS

**AUTHOR: LUIS MIGUEL MAZAIRA FERNÁNDEZ**

**DIRECTOR: PROFESSOR DR. AGUSTÍN ÁLVAREZ MARQUINA**

**APRIL, 2014**


**UNIVERSIDAD POLITÉCNICA DE MADRID**

**Escuela Técnica Superior de Ingenieros Informáticos**

# DEPARTAMENTO DE ARQUITECTURA Y TECNOLOGÍA DE SISTEMAS INFORMÁTICOS

# NEW METHODOLOGY FOR THE INTEGRATION OF BIOMETRIC FEATURES IN SPEAKER RECOGNITION SYSTEMS APPLIED TO SECURITY ENVIRONMENTS

**AUTHOR: LUIS MIGUEL MAZAIRA FERNANDEZ**

**DIRECTOR: PROFESSOR DR. AGUSTÍN ÁLVAREZ MARQUINA**

**APRIL, 2014**

Tribunal nombrado por el Sr. Rector Magfco. de la Universidad Politécnica de Madrid, el día 29 de Septiembre de 2014

Presidente: D. Pedro Gómez Vilda

Vocal: D. Carlos García Puntonet

Vocal: D. Daniel Ramos Castro

Vocal: D. Athanasios Tsanas

Secretario: D. Francisco del Pozo Guerrero

Suplente: D. Carlos Enrique Vivaracho Pascual

Suplente: D. Jose Manuel Ferrandez Vicente

Realizado el acto de defensa y lectura de la Tesis el día 14 de Noviembre de 2014 en la E.T.S.I. /Facultad Informáticos

Calificación .....

EL PRESIDENTE LOS VOCALES

EL SECRETARIO

# INDEX

INDEX ..... iv
TABLE INDEX ..... vii
FIGURE INDEX ..... xi
RESUMEN ..... xix
ABSTRACT ..... xx
ACKNOWLEDGEMENTS ..... xxi
1 INTRODUCTION ..... 22
1.1 MOTIVATIONS ..... 22
1.2 OBJECTIVES ..... 25
1.3 A GUIDE TO THE TEXT ..... 26
1.4 BIOMETRIC AUTHENTICATION ..... 27
1.4.1 Biometric characteristics ..... 33
1.4.2 Multibiometrics ..... 62
1.4.3 Multimodal biometric databases ..... 64
1.4.4 Biometric applications ..... 66
1.5 STATE OF THE ART IN SPEAKER RECOGNITION ..... 68
1.5.1 Feature extraction ..... 71
1.5.2 Class modelling and classification methods ..... 76
1.5.3 Class decision ..... 78
1.5.4 Experimental recognition results ..... 80
1.6 SPEAKER RECOGNITION COMMERCIAL APPLICATIONS ..... 87
2 SPEECH PRODUCTION AND BIOMETRIC CHARACTERISATION OF VOICE ..... 91
2.1 SPEECH PRODUCTION ..... 92
2.1.1 3-mass model of the vocal folds (body-cover structure) ..... 98
2.1.2 2-mass model of the cover structure of the vocal folds ..... 100
2.1.3 1-mass model of the body structure of the vocal folds ..... 104
2.2 ACCOUSTIC THEORY OF SPEECH PRODUCTION ..... 104
2.3 SOURCE-TRACT SEPARATION OF THE SPEECH SIGNAL ..... 106
2.3.1 Linear Prediction review ..... 106
2.3.1.1 Properties of the prediction error filters ..... 112
2.3.2 Source-Tract separation proposed algorithm ..... 113
2.3.3 Comments on the source-tract separation algorithm ..... 122
2.4 BIOMETRIC CHARACTERISATION OF VOICE ..... 124
2.4.1 Feature extraction ..... 126
2.4.1.1 Voice signal feature extraction ..... 126
2.4.1.2 Vocal Tract feature extraction ..... 132
2.4.1.3 Glottal source feature extraction ..... 136
2.4.1.3.1 Power Spectral Density Profile of the glottal source ..... 136
2.5 ALTERNATIVE PARAMETERISATIONS ..... 147
2.6 GLOTTAL SOURCE ALTERNATIVE APPLICATION ..... 149
3 CLASSIFICATION METHODS ..... 151
3.1 VECTOR QUANTISATION ..... 151
3.2 HIDDEN MARKOV MODELS ..... 153
3.2.1 Types of HMMs ..... 160

3.2.2 Implementation Issues ..... 162
3.2.3 Available software tools ..... 166
3.3 GAUSSIAN MIXTURE MODELS ..... 166
3.3.1 The expectation maximisation algorithm ..... 168
3.3.2 Background model adaptation ..... 170
3.3.3 Model Order ..... 173
3.3.4 Phonetic GMM variations ..... 175
3.3.5 Available software tools ..... 175
3.4 SUPPORT VECTOR MACHINES ..... 176
3.4.1 Basic SVM theory ..... 177
3.4.2 Kernel functions ..... 181
3.4.3 Training algorithms ..... 183
3.4.4 Multi-class classification in SVMs ..... 186
3.4.5 Available software tools ..... 188
3.5 NEW TRENDS IN CLASSIFICATION METHODS ..... 189
3.5.1 Supervector methods ..... 189
3.5.2 *i*-vectors ..... 199
3.6 FUSION ..... 203
3.6.1 Available software tools ..... 205
4 TEST BENCH ..... 206
4.1 DATABASES ..... 206
4.1.1 The Database HESPERIA ..... 207
4.1.2 The Database ALBAYZIN ..... 208
4.1.3 The Database MOBIO ..... 211
4.1.4 NIST SRE databases ..... 212
4.2 PRACTICAL SCENARIOS ..... 216
4.2.1 Text-Constrained Speaker Recognition ..... 217
4.2.2 Text-Independent Speaker Recognition ..... 220
4.2.3 Text-Independent Speaker Recognition in Mobile Environments ..... 221
4.2.4 NIST SRE Evaluations ..... 223
5 APPLICATION TO SPEAKER RECOGNITION ..... 235
5.1 SYSTEM DESCRIPTION ..... 235
5.1.1 Baseline front-end ..... 235
5.1.2 Gender-Dependent Extended-Biometric front-end ..... 239
5.1.2.1 Glottal Estimate – Vocal Tract Estimate Separation Algorithm ..... 240
5.1.2.2 Algorithm's parameter tuning ..... 240
5.1.2.3 Feature vector composition ..... 241
5.1.3 Speaker Recognition systems ..... 241
5.1.3.1 GMM-UBM ..... 241
5.1.3.2 SV-GMM ..... 243
5.1.3.3 i-vectors ..... 244
5.2 RESULTS ..... 245
5.2.1 Text-Constrained Speaker Recognition ..... 246
5.2.1.1 Scenario 1 (mic-mic) ..... 247
5.2.1.1.1 Brief Conclussions ..... 275
5.2.1.2 Scenario 2 (mic-tel) ..... 276
5.2.1.2.1 Brief Conclussions ..... 302
5.2.2 Text-Independent Speaker Recognition ..... 303
5.2.2.1 Brief Conclussions ..... 331
5.2.3 Text-Independent Speaker Recognition in Mobile Environments ..... 332

5.2.3.1 Brief Conclussions ..... 350
5.2.4 NIST SRE Evaluations ..... 350
5.2.4.1 Brief Conclussions ..... 382
6 CONCLUSIONS ..... 383
6.1 OVERVIEW ..... 383
6.2 CONCLUSIONS ..... 383
6.3 FUTURE WORK ..... 390
6.4 CONTRIBUTIONS ..... 393
I. BIBLIOGRAPHY ..... 399
II. APPENDICES ..... 440
II.1. Software Tools ..... 440
II.2. Glossary ..... 442
II.2.1. Acronyms ..... 442
II.2.2. Terms ..... 443

# TABLE INDEX

TABLE 1-1 DIFFERENT EAR REPRESENTATION'S: (1. [IANNARELLI,1989] 2. [BURGE,2000] 3. [MORENO,1999] 4. [HURLEY,2000]) ..... 37
TABLE 1-2 FVC2006 DATASETS ..... 49
TABLE 1-3 VASCULAR PATTERNS USED FOR INDIVIDUAL RECOGNITION ..... 52
TABLE 1-4 SUMMARY OF MULTIMODAL DATABASES (FA-FACE, IR-IRIS, FP-FINGERPRINT, HP-HAND PRINT, HW-HANDWRITING, KS-KEYSTROKE) ..... 66
TABLE 1-5 DIFFERENT RECOGNITION RESULTS IN SPEAKER RECOGNITION ..... 86
TABLE 1-6 INDUSTRY VENDORS FOR SPEAKER RECOGNITION SYSTEMS ..... 90
TABLE 3-1 SOME EXAMPLES OF MAPPING FUNCTIONS ..... 194
TABLE 4-1 MAIN CHARACTERISTICS OF THE DIFFERENT DATABASES USED IN THE EXPERIMENTS CARRIED OUT IN THE COURSE OF THIS THESIS (*: ONLY ENGLISH). ... 207
TABLE 4-2 NUMBER OF UTTERANCE PER SPEAKER FOR EACH SESSION AND CHANNEL IN THE HESPERIA DATABASE. ..... 208
TABLE 4-3 UTTERANCE DISTRIBUTION FOR FEMALE SPEAKERS REGARDING AGE AND SUB-CORPUS ..... 210
TABLE 4-4 UTTERANCE DISTRIBUTION FOR MALE SPEAKERS REGARDING AGE AND SUB-CORPUS ..... 210
TABLE 4-5 UTTERANCE, SESSION AND SPEAKER DISTRIBUTION FOR EACH OF THE SITES PARTICIPATING IN THE MOBIO DATABASE ..... 212
TABLE 4-6 NUMBER OF RECORDINGS AND SPEAKERS IN THE NIST SRE 2004 DATABASE ..... 213
TABLE 4-7 NUMBER OF RECORDINGS AND SPEAKERS IN THE NIST SRE 2006 DATABASE ..... 214
TABLE 4-8 NUMBER OF RECORDINGS AND SPEAKERS IN THE NIST SRE 2008 DATABASE ..... 215
TABLE 4-9 NUMBER OF RECORDINGS AND SPEAKERS IN THE NIST SRE 2010 DATABASE ..... 216
TABLE 4-10 NUMBER OF RECORDINGS AND SPEAKERS IN THE NIST SRE 2012 DATABASE ..... 216
TABLE 4-11 DESCRIPTION OF THE CONTENTS OF THE DIFFERENT SUBSETS OF THE HESPERIA DATABASE FOR SCENARIO 1 ..... 218
TABLE 4-12 DESCRIPTION OF THE CONTENTS OF THE DIFFERENT SUBSETS OF THE HESPERIA DATABASE FOR SCENARIO 2 ..... 220
TABLE 4-13 DESCRIPTION OF THE CONTENTS OF THE DIFFERENT SUBSETS OF THE ALBAYZIN DATABASE ..... 221
TABLE 4-14 DESCRIPTION OF THE CONTENTS OF THE DIFFERENT SUBSETS OF THE MOBIO DATABASE ..... 223
TABLE 4-15 NUMBER OF MODELS, SPEAKERS AND TRAILS TO BE PROCESSED ..... 226
TABLE 4-16 NUMBER OF MODELS, SPEAKERS AND TRAILS TO BE PROCESS IN THE DIFFERENT CONDITIONS OF CORE TASK ..... 227
TABLE 4-17 *EER* RANGES FOR THE DIFFERENT TASKS ON NIST SRE 2010 ..... 229
TABLE 4-18 FEATURES AND CLASSIFIERS USED BY THE TOP 6 SYSTEMS ..... 230
TABLE 4-19 DESCRIPTION OF THE CONTENTS OF THE DIFFERENT SUBSETS OF THE NIST SRE 2012 CORE-CORE TASK ..... 233
TABLE 5-1 BASELINE FRONT-END BASED SR SYSTEM PROVIDING BETTER *HEER* IN A GENDER-INDEPENDENT CONFIGURATION ..... 250
TABLE 5-2 GDC VS. GIC FOR HESPERIA'S DEVELOPMENT SET ON SCENARIO 1 ..... 254

TABLE 5-3 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET WHEN EXTRA PARAMETERS ARE INCLUDED ON THE FEATURE VECTORS FOR GIC AND GDC ..... 256
TABLE 5-4 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET COMPARING DIFFERENT GIC MFCC+Δ+ΔΔ CONFIGURATIONS ..... 257
TABLE 5-5 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON DEVELOPMENT SET (NO SCORE NORMALISATION), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS. (PO: PREDICTION ORDER; FF: FORGETTING FACTOR; FB: FILTER BANK) ..... 259
TABLE 5-6 $EER_M$, $EER_F$, AND *HEER* OBTAINED FOR THE DEVELOPMENT SET (ZNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 264
TABLE 5-7 $EER_M$, $EER_F$, AND *HEER* OBTAINED FOR THE DEVELOPMENT SET (TNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 265
TABLE 5-8 $EER_M$, $EER_F$, AND *HEER* OBTAINED FOR THE DEVELOPMENT SET (ZTNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 266
TABLE 5-9 $HTER_X$ OBTAINED FOR SELECTED CONFIGURATIONS ON THE EVALUATION SET, APPLYING DIFFERENT SCORE NORMALISATIONS ..... 269
TABLE 5-10 BASELINE FRONT-END BASED SR SYSTEM PROVIDING BETTER *HEER* IN A GENDER-INDEPENDENT CONFIGURATION ..... 279
TABLE 5-11 GDC VS. GIC FOR HESPERIA'S DEVELOPMENT SET ON SCENARIO 2 ..... 282
TABLE 5-12 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET WHEN EXTRA PARAMETERS ARE INCLUDED ON THE FEATURE VECTORS FOR GIC AND GDC ..... 284
TABLE 5-13 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET COMPARING DIFFERENT GIC MFCC+Δ+ΔΔ CONFIGURATIONS ..... 285
TABLE 5-14 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET (NO SCORE NORMALISATION), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 287
TABLE 5-15 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET (ZNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 292
TABLE 5-16 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET (TNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 293
TABLE 5-17 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET (ZTNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 294
TABLE 5-18 $HTER_X$ PRODUCED FOR SELECTED CONFIGURATIONS ON EVALUATION SET, APPLYING DIFFERENT SCORE NORMALISATIONS ..... 297
TABLE 5-19 BASELINE FRONT-END PRODUCING THE BEST *HEER* IN A GENDER-INDEPENDENT CONFIGURATION ..... 306
TABLE 5-20 GDC VS. GIC FOR THE ALBAYZIN DEVELOPMENT SET (RR – RELATIVE REDUCTION) ..... 310
TABLE 5-21 $EER_M$, $EER_F$ AND *HEER* OBTAINED ON DEVELOPMENT SET WHEN EXTRA PARAMETERS ARE INCLUDED ON THE FEATURE VECTORS FOR GIC AND GDC ..... 312
TABLE 5-22 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET COMPARING DIFFERENT GIC MFCC+Δ+ΔΔ CONFIGURATIONS ..... 313

TABLE 5-23 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON DEVELOPMENT SET (NO SCORE NORMALISATION), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 316
TABLE 5-24 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET (ZNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 319
TABLE 5-25 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET (TNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 320
TABLE 5-26 $EER_M$, $EER_F$, AND *HEER* OBTAINED ON THE DEVELOPMENT SET (ZTNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 321
TABLE 5-27 $HTER_X$ OBTAINED FOR SELECTED CONFIGURATIONS ON EVALUATION SET, APPLYING DIFFERENT SCORE NORMALISATIONS ..... 326
TABLE 5-28 GDC VS. GIC FOR MOBIO DEVELOPMENT SET (RR – RELATIVE REDUCTION) ..... 338
TABLE 5-29 *EER* OBTAINED FOR THE TESTS WHICH INCORPORATE E, ΔE, F0 AND F3 IN THE FEATURE VECTORS (BEST RESULTS HIGHLIGHTED IN GREEN)..... 342
TABLE 5-30 $EER_M$ OBTAINED ON DEVELOPMENT SET (ZTNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS..... 343
TABLE 5-31 $EER_F$ OBTAINED ON DEVELOPMENT SET (NO NORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS..... 344
TABLE 5-32 $HTER_X$ OBTAINED FOR THE SELECTED CONFIGURATIONS ON EVALUATION SET, APPLYING ZTNORM – MALE – AND NO NORM – FEMALE. ..... 347
TABLE 5-33 EER % ON THE DEVELOPMENT (DEV) SET AND HALF TOTAL ERROR RATE (*HTER* %) ON THE EVALUATION (EVAL) SET FOR THE SYSTEMS PARTICIPATING IN 2013 SRE IN MOBILE ENVIRONMENTS. ..... 349
TABLE 5-34 RESULTS OBTAINED IN TERMS OF *ACER* BY THE DIFFERENT CLASSIFICATION METHODS FOR EACH GENDER ..... 353
TABLE 5-35 $EER_M$ *AND* *ACER* OBTAINED DEPENDING ON THE *BASELINE* FRONT-END SETUP NIST SRE 2010 (BEST RESULTS ARE HIGHLIGHTED IN GREEN)..... 360
TABLE 5-36 $EER_F$ *AND* *ACER* OBTAINED DEPENDING ON THE *BASELINE* FRONT-END SETUP NIST SRE 2010 (BEST RESULTS ARE HIGHLIGHTED IN GREEN)..... 361
TABLE 5-37 $EER_M$ COMPARISON FOR THE DIFFERENT CONDITIONS IN WHICH DEVELOPMENT SET IS DIVIDED ..... 363
TABLE 5-38 $EER_F$ COMPARISON FOR THE DIFFERENT CONDITIONS IN WHICH DEVELOPMENT SET IS DIVIDED ..... 365
TABLE 5-39 $EER_M$ OBTAINED ON THE DEVELOPMENT SET (SNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS..... 367
TABLE 5-40 $EER_F$ OBTAINED ON THE DEVELOPMENT SET (SNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS..... 368
TABLE 5-41 $EER_F$ OBTAINED ON THE DEVELOPMENT SET (SNORM+ZTNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 369
TABLE 5-42 *EER* RANGES FOR THE DIFFERENT TASKS ON NIST SRE 2010..... 375

TABLE 5-43 $EER_M$ OBTAINED ON EVALUATION SET (SNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS 376
TABLE 5-44 $EER_F$ OBTAINED ON EVALUATION SET (SNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS 377
TABLE 5-45 $EER_F$ OBTAINED ON EVALUATION SET (SNORM+ZTNORM), COMPARING CLASSICAL PARAMETERS WITH EXTRA PARAMETERS AND EXTENDED BIOMETRIC PARAMETERS ..... 377

# FIGURE INDEX

FIGURE 1-1 OPERATION MODES OF A BIOMETRIC RECOGNITION SYSTEM..... 29
FIGURE 1-2 SCORE DISTRIBUTION CURVES OF A HYPOTHETICAL BIOMETRIC RECOGNITION SYSTEM..... 30
FIGURE 1-3 ROC CURVE OF A HYPOTHETICAL BIOMETRIC RECOGNITION SYSTEM..... 30
FIGURE 1-4 DET CURVE OF A HYPOTHETICAL BIOMETRIC RECOGNITION SYSTEM..... 31
FIGURE 1-5 TWO ARTIFICIAL EPC CURVES COMPARING TWO MODELS USING A PERFORMANCE MEASURE V (EXTRACTED FROM [BENGIO,2005])..... 31
FIGURE 1-6 RIGHT EYE DESCRIPTION..... 38
FIGURE 1-7 EXAMPLE OF IRIS LOCALISATION (WHITE CIRCLE LINES) [DAUGMAN,2004]. 39
FIGURE 1-8 IRIS NORMALISATION PROCESS (EXTRACTED FROM[ROY,2007])..... 40
FIGURE 1-9 GENERAL VOICE IDENTIFICATION SYSTEM..... 44
FIGURE 1-10 HAND ACQUISITION SURFACE WITH PEGS (EXTRACTED FROM [JAIN,1999-B])..... 50
FIGURE 1-11 DIFFERENT APPROACHES FOR HUMAN GAIT MODEL. A) [GUO,2008] B) [CUNADO,2003] C) [LEE,2002]..... 59
FIGURE 1-12 VOICE BIOMETRY AND ITS RELATION TO BIOMETRICS AND SPEECH PROCESSING..... 68
FIGURE 1-13 SPEAKER RECOGNITION..... 69
FIGURE 1-14 GENERAL SPEAKER IDENTIFICATION SYSTEM..... 69
FIGURE 1-15 GENERAL SPEAKER VERIFICATION SYSTEM..... 70
FIGURE 1-16 TRAIN/ENROLMENT PHASE..... 71
FIGURE 1-17 GENERAL FILTER BANK-BASED CEPSTRAL PARAMETERISATION..... 72
FIGURE 1-18 GENERAL LPC-BASED CEPSTRAL PARAMETERISATION..... 73
FIGURE 2-1 A SIMPLIFIED DIAGRAM OF SPEECH COMMUNICATION (WITHOUT TAKING INTO ACCOUNT BRAIN INTERACTION). V: VELUM. G: TONGUE. L: LIPS, P: HARD PALATE. A: ALVEOLUS. N: NASAL CAVITY. T: TEETH..... 91
FIGURE 2-2 VOCAL FOLDS: ENDOSCOPIC VIEW OF VOCAL FOLDS AT INSPIRATION (LEFT-HAND SIDE) AND IN PRE-PHONATION POSITION (RIGHT-HAND SIDE)..... 92
FIGURE 2-3 EQUIVALENT DYNAMIC MASS OF THE VOCAL FOLD. THE SECTION THAT MAINLY CONTRIBUTES TO THE VIBRATION OF THE FOLD CORRESPONDS TO THE WEDGE ON THE RIGHT, CORRESPONDING TO AN INERTIAL MASS [...]..... 94
FIGURE 2-4 CROSS SECTION OF THE LEFT VOCAL FOLD: (A) BODY-COVER STRUCTURE, FOLLOWING THE BLUE DASH-LINE OF THE ENDOSCOPIC VIEW IN FIGURE 2-2 RIGHT, (B) K-MASS (BODY MASS + K-1 COVER MASSES) EQUIVALENT MECHANICAL MODEL..... 94
FIGURE 2-5 PHONATION CYCLE SEQUENCE FROM GLOTTAL CLOSURE TO GLOTTAL CLOSURE (1-10), DESCRIBING THE MUCOSAL WAVE. [...]..... 96
FIGURE 2-6 SIMPLEST VIBRATION MODE OF THE VOCAL FOLDS. TOP: SOLID LINE REPRESENTS THE POSITION OF THE RIGHT BODY MASS, WHILE DASH LINE REPRESENTS THE POSITION OF THE LEFT BODY MASS. [...]..... 97
FIGURE 2-7 GLOTTAL SOURCE IDEAL MODEL. A) GLOTTAL SOURCE, IN WHICH WE CAN HIGHLIGHT THE FOLLOWING INSTANTS: [...]..... 97
FIGURE 2-8 SIMPLIFIED VERSION OF THE WELL-KNOWN STORY-TITZE MODEL, WHICH IS AGREED TO GIVE A MORE ACCURATE EXPLANATION OF THE *MUCOSAL WAVE PHENOMENON* [...]..... 98
FIGURE 2-9 2-MASS AND 3 SPRING BY VOCAL FOLD MODEL. SCHEMATIC STRUCTURE OF ONE-SECTION OF THE RIGHT (R) AND LEFT (L) VOCAL FOLDS. [...]..... 99

FIGURE 2-10 EQUIVALENT ELECTROMECHANICAL MODEL OF THE 2-MASS AND 3-SPRING MODEL OF THE VOCAL FOLDS DEPICTED IN FIGURE 2-9. [...] ..... 101
FIGURE 2-11 POWER SPECTRAL DENSITY (PSD) OF MALE VOICE SEGMENT SYNCHRONOUSLY EVALUATED IN A PHONATION CYCLE, WHICH MATCH THE HARMONIC ENVELOPE OR THE PSD PROFILE. [...] ..... 101
FIGURE 2-12 MODULUS AND PHASE FREQUENCY RESPONSE OF THE TRANS-ADMITTANCE BETWEEN THE FORCE THAT AFFECTS ONE MASS ANT THE VELOCITY THAT APPEARS IN THE OPPOSITE MASS. [...] ..... 103
FIGURE 2-13 GUNNAR FANT'S PRODUCTION MODEL. AN EXCITATION SOURCE *E(N)*, EITHER A GLOTTAL EXCITATION FOR VOICED VOICE OR A TURBULENT EXCITATION FOR UNVOICED VOICE[...] ..... 105
FIGURE 2-14 ACOUSTIC TUBE MODEL OF SPEECH PRODUCTION ([CAMPBELL,1997]) ..... 106
FIGURE 2-15 LINEAR PREDICTION OVER A DISCRETE SIGNAL WHICH RESULTS FROM SAMPLING A CONTINUOUS SIGNAL AT REGULAR INTERVALS ..... 107
FIGURE 2-16 LATTICE STRUCTURE THAT GENERATES FORWARD AND BACKWARD PREDICTOR ERRORS FOR ALL OPTIMAL PREDICTORS WITH ORDER $1 \leq K \leq K$, AND DETAILS OF THE $K^{TH}$ LATTICE ..... 111
FIGURE 2-17 GEOMETRICAL INTERPRETATION OF THE *ORTHOGONALITY PRINCIPLE* IN A PREDICTION ERROR FILTER OF ORDER 2 ..... 112
FIGURE 2-18 GENERAL FILTERING MODEL FOR THE INVERSION OF FANT'S VOICE PRODUCTION MODEL (SEE FIGURE 2-13) ..... 114
FIGURE 2-19 LIP RADIATION CANCEL FILTER IMPLEMENTED AS A FIRST ORDER LATTICE. A) PARCOR LATTICE IMPLEMENTATION. B) FIRST ORDER TRANSVERSAL FILTER ..... 115
FIGURE 2-20 BLOCK DIAGRAM METHOD FOR THE RECONSTRUCTION OF THE GLOTTAL RESIDUAL CORRELATE BY COMPLEMENTED INVERSE FILTERING[...] ..... 117
FIGURE 2-21 ITERATIVE ESTIMATION OF THE VOCAL TRACT TRANSFER FUNCTION $F_V(Z)$AND THE GLOTTAL RESIDUAL $V_G(N)$. BLOCKS $F_V(Z)$ AND $H_V(Z)$ OR $F_G(Z)$ AND $H_G(Z)$ ARE IMPLEMENTED BY SUCCESSIVE CHAINS OF ADAPTIVE LATTICE FILTERS .. 118
FIGURE 2-22 PAIRED LATTICE JOINT ESTIMATOR THAT COMBINES A MODELLING FILTER SECTION WITH AN INVERSE FILTERING SECTION, JOINING IN A SINGLE STRUCTURE THE TWO BLOCKS OF EACH DIAGONAL IN FIGURE 2-21 ..... 118
FIGURE 2-23 SPEECH GLOTTAL TRACES THAT RESULT FROM THE DESCRIBED SEPARATION ALGORITHM[...] ..... 119
FIGURE 2-24 REAL GLOTTAL SOURCE MODELS RECONSTRUCTED ACCORDING TO THE PRESENTED METHOD FOR A MALE SPEAKER[...]. ..... 121
FIGURE 2-25 REAL GLOTTAL SOURCE MODELS RECONSTRUCTED ACCORDING TO THE PRESENTED METHOD FOR A FEMALE SPEAKER [...]. ..... 122
FIGURE 2-26 ANALYTIC DESCRIPTION OF VOICE BIOMETRY FROM THE PRODUCTION MODEL IN TERMS OF VOCAL (MAINLY MESSAGE DEPENDENT) AND GLOTTAL (MAINLY BIOMETRIC) CHARACTERISTICS ..... 126
FIGURE 2-27 BLOCK DIAGRAM REPRESENTATION OF THE MFCC PARAMETERISATION .. 127
FIGURE 2-28 EXTRACTION OF SPECTRAL ENVELOPE USING CEPSTRAL ANALYSIS (BLUE SOLID LINE) AND LINEAR PREDICTION (GREEN SOLID LINE) FROM THE FFT SPECTRUM WITH N=512 ..... 129
FIGURE 2-29 MEL-SCALE FILTER BANK WITH 24 TRIANGLE-SHAPE FILTERS ..... 129
FIGURE 2-30 ENVELOPE SPECTRUM USING A 24-BAND MEL FILTER BANK ..... 130
FIGURE 2-31 LINEAR SCALE FILTER BANK FOR TELEPHONE CHANNEL VOICE PROCESSING ..... 131
FIGURE 2-32 VOCAL TRACT (MIDDLE) AND GLOTTAL SOURCE (LOWER) ESTIMATES FOR A FEMALE VOWEL AN UTTERANCE (UPPER) ..... 132

FIGURE 2-33 FFT POWER SPECTRUM OF THE ORIGINAL VOICE SIGNAL (SOLID BLUE LINE) AND TRANSFER FUNCTION OF THE VOCAL TRACT ESTIMATE (RED SOLID LINE) ..... 133
FIGURE 2-34 BLOCK-DIAGRAM REPRESENTATION OF AN LPCC EXTRACTION PROCEDURE ..... 133
FIGURE 2-35 BLOCK-DIAGRAM REPRESENTATION OF AN LSP COEFFICIENTS EXTRACTION PROCEDURE ..... 134
FIGURE 2-36 ZEROS OF THE LSP-PROCESS INVOLVED POLYNOMIALS. WHITE CIRCLES: ORIGINAL LP POLYNOMIAL ZEROES. SQUARES: *P(Z)'S* ZEROES. FILLED CIRCLES: *Q(Z)'S* ZEROES (EXTRACTED FROM [ONSHAUNJIT,2008]). ..... 135
FIGURE 2-37 POWER SPECTRAL DENSITY OF THE GLOTTAL SOURCE EVALUATED OVER A TEMPORAL WINDOW WHICH INCLUDES MULTIPLE GLOTTAL CYCLES. UPPER – MALE VOICE. LOWER – FEMALE VOICE. [...] ..... 137
FIGURE 2-38 AAW AND MWC REAL CASE ESTIMATION: A) LEVELLED GLOTTAL SOURCE. B) AVERAGE ACOUSTIC WAVE ALSO KNOWN AS AVERAGE GLOTTAL SOURCE. C) LEVELLED GLOTTAL FLOW. D) MUCOSAL WAVE CORRELATE ..... 140
FIGURE 2-39 APPROXIMATION OF THE PSD PROFILE OF THE AAW (SOLID LINE) THROUGH THE TRANSFER FUNCTION OF A SECOND ORDER ELECTROMECHANICAL EQUIVALENT (RLC: DOTTED LINE). [...] ..... 141
FIGURE 3-1 A MARKOV CHAIN WITH 4 STATES. THE DISCRETE STATES, LABELLED AS $S_I$ ARE REPRESENTED BY NODES, AND THE TRANSITION PROBABILITIES, $A_{IJ}$, ARE REPRESENTED BY LINKS BETWEEN NODES. ..... 154
FIGURE 3-2 3-STATE HMM; WHERE $A_{IJ}$ REPRESENTS THE STATE TRANSITION PROBABILITIES, $V_K$ DENOTES THE OBSERVATION SYMBOLS ALSO KNOWN AS EMITTING OR VISIBLE STATES AND $B_J(K)$ REPRESENTS THE PROBABILITY OF THE EMISSION OF A VISIBLE STATE. ..... 155
FIGURE 3-3 ILLUSTRATION OF 5-STATE LEFT-TO-RIGHT HMM ..... 155
FIGURE 3-4 SVM TRAINING STRATEGY FOR SPEAKER RECOGNITION ..... 177
FIGURE 3-5 MAXIMUM MARGIN HYPERPLANE THAT SEPARATES TWO LINEARLY SEPARABLE CLASSES ..... 178
FIGURE 3-6 COMPARATIVE OF THREE STEPS OF THE CHUNKING, OSUNA AND SMO TRAINING ALGORITHMS (EXTRACTED FROM [PLATT,1999-A]). THE HORIZONTAL SOLID LINE REPRESENTS THE TRAINING SET, WHILE THE BLACK BOXES CORRESPOND TO THE LAGRANGE MULTIPLIERS BEING OPTIMISED. ..... 186
FIGURE 3-7 DAG FOR A 4-CLASS CLASSIFICATION PROBLEM, WHERE EACH NODE REPRESENTS A BINARY CLASSIFIER. (EXTRACTED FROM [PLATT,1999-B]). ..... 187
FIGURE 3-8 BLOCK DIAGRAM DESCRIPTION SHOWING THE MAIN CONCEPTS INVOLVED IN THE SEQUENCE KERNEL SVM MODELLING APPROACH. ..... 190
FIGURE 4-1 NOISY TELEPHONIC RECORDING (TIME DOMAIN – LEFT – AND FREQUENCY DOMAIN – RIGHT) IN NIST SRE 2010 DATABASE ..... 224
FIGURE 4-2 LOW QUALITY RECORDING (TIME DOMAIN – LEFT – AND FREQUENCY DOMAIN – RIGHT) IN NIST SRE 2010 DATABASE ..... 224
FIGURE 4-3 INCORRECT LABELLED RECORDING (TIME DOMAIN – LEFT – AND FREQUENCY DOMAIN – RIGHT) IN NIST SRE 2010 DATABASE. NO SPEAKER INFORMATION AVAILABLE. ..... 224
FIGURE 4-4 SET OF DET CURVES OBTAINED FOR THE DIFFERENT CONDITIONS (1 TO 6) OF THE CORE-CORE TASK (PROVIDED BY NIST) ..... 231
FIGURE 4-5 SET OF DET CURVES OBTAINED FOR THE DIFFERENT CONDITIONS (7 TO 9) OF THE CORE-CORE TASK (PROVIDED BY NIST) ..... 232
FIGURE 5-1 CLASSICAL FEATURE EXTRACTION PROCESS ..... 237
FIGURE 5-2 FEATURE VECTOR COMPOSITION WITH COMPENSATION TECHNIQUES ..... 238

FIGURE 5-3 SEPARATION ALGORITHM WITH LIP RADIATION COMPENSATION USING FIRST ORDER PREDICTION LATTICE ..... 240
FIGURE 5-4 PARAMETERISATION SCHEME USED FOR MALE SPEAKERS ..... 241
FIGURE 5-5 GMM-UBM SPEAKER VERIFICATION SYSTEM IMPLEMENTED ..... 243
FIGURE 5-6 SV-GMM SPEAKER VERIFICATION SYSTEM IMPLEMENTED TO RUN THE NIST EXPERIMENTS ..... 244
FIGURE 5-7 *I*-VECTOR SPEAKER VERIFICATION SYSTEM IMPLEMENTED TO RUN THE NIST EXPERIMENTS ..... 245
FIGURE 5-8 *HEER* OBTAINED DEPENDING ON THE NUMBER OF MFCCS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 248
FIGURE 5-9 *HEER* OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 249
FIGURE 5-10 *HEER* OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 250
FIGURE 5-11 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF MFCCS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 251
FIGURE 5-12 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF MFCCS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 251
FIGURE 5-13 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 252
FIGURE 5-14 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 252
FIGURE 5-15 $EER_M$ (BLUE) AND $EER_F$ (RED) OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 253
FIGURE 5-16 DET CURVE FOR CLASSIC PARAMETERS ON HESPERIA'S MALE DEVELOPMENT SET FOR GIC AND GDC ..... 255
FIGURE 5-17 DET CURVE FOR CLASSIC PARAMETERS ON HESPERIA'S FEMALE DEVELOPMENT SET FOR GIC AND GDC ..... 255
FIGURE 5-18 DET CURVES COMPARING DIFFERENT SET OF PARAMETERS UNDER GIC AND GDC WITHOUT EXTENDED BIOMETRICS ON MALE'S DEVELOPMENT SET ..... 258
FIGURE 5-19 DET CURVES COMPARING DIFFERENT SET OF PARAMETERS UNDER GIC AND GDC WITHOUT EXTENDED BIOMETRICS ON FEMALE'S DEVELOPMENT SET ..... 258
FIGURE 5-20 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON HESPERIA'S DEVELOPMENT SET FOR MALE SPEAKERS ..... 260
FIGURE 5-21 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON HESPERIA'S DEVELOPMENT SET FOR FEMALE SPEAKERS ..... 261
FIGURE 5-22 INFLUENCE OF GSE CONFIGURATION ON THE $EER_M$ (DEVELOPMENT SET) 262
FIGURE 5-23 INFLUENCE OF GSE CONFIGURATION ON THE $EER_F$ (DEVELOPMENT SET). 262
FIGURE 5-24 DET CURVES FOR *BASELINE* AND GDEB FRONT-END, APPLYING DIFFERENT SCORE NORMALISATION TECHNIQUES (MALE'S DEVELOPMENT SET) ..... 267
FIGURE 5-25 DET CURVES FOR *BASELINE* AND GDEB FRONT-END, APPLYING DIFFERENT SCORE NORMALISATION TECHNIQUES (FEMALE'S DEVELOPMENT SET) ..... 268
FIGURE 5-26 MALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, WITHOUT APPLYING ANY SCORE NORMALISATION TECHNIQUE ..... 271
FIGURE 5-27 MALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, APPLYING ZNORM ..... 271
FIGURE 5-28 MALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, APPLYING TNORM ..... 272

FIGURE 5-29 MALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, APPLYING ZTNORM ..... 272
FIGURE 5-30 FEMALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, WITHOUT APPLYING ANY SCORE NORMALISATION TECHNIQUE ..... 273
FIGURE 5-31 FEMALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, APPLYING ZNORM ..... 273
FIGURE 5-32 FEMALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, APPLYING TNORM ..... 274
FIGURE 5-33 FEMALE'S DET CURVES ON THE EVALUATION SET FROM HESPERIA, APPLYING ZTNORM ..... 274
FIGURE 5-34 *HEER* OBTAINED DEPENDING ON THE NUMBER OF MFCCS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 277
FIGURE 5-35 *HEER* OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 277
FIGURE 5-36 *HEER* OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 278
FIGURE 5-37 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF MFCCS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 279
FIGURE 5-38 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF MFCCS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 280
FIGURE 5-39 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 280
FIGURE 5-40 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 281
FIGURE 5-41 $EER_M$ (BLUE) AND $EER_F$ (RED) OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 281
FIGURE 5-42 DET CURVE FROM CLASSIC PARAMETERS ON HESPERIA MALE DEVELOPMENT SET FOR GIC AND GDC ..... 283
FIGURE 5-43 DET CURVE FROM CLASSIC PARAMETERS ON HESPERIA FEMALE DEVELOPMENT SET FOR GIC AND GDC ..... 283
FIGURE 5-44 DET CURVES COMPARING DIFFERENT SETS OF PARAMETERS UNDER GIC AND GDC WITHOUT EXTENDED BIOMETRICS ON MALE'S DEVELOPMENT SET ..... 286
FIGURE 5-45 DET CURVES COMPARING DIFFERENT SETS OF PARAMETERS UNDER GIC AND GDC WITHOUT EXTENDED BIOMETRICS ON FEMALE'S DEVELOPMENT SET ..... 286
FIGURE 5-46 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON HESPERIA DEVELOPMENT SET FOR MALE SPEAKERS ..... 288
FIGURE 5-47 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON HESPERIA DEVELOPMENT SET FOR FEMALE SPEAKERS ..... 289
FIGURE 5-48 INFLUENCE OF GSE CONFIGURATION ON THE $EER_M$ (DEVELOPMENT SET) 290
FIGURE 5-49 INFLUENCE OF GSE CONFIGURATION ON THE $EER_F$ (DEVELOPMENT SET). 290
FIGURE 5-50 DET CURVES FOR *BASELINE* AND GDEB FRONT-END, APPLYING DIFFERENT SCORE NORMALISATION TECHNIQUES (MALE'S DEVELOPMENT SET) ..... 295
FIGURE 5-51 DET CURVES FOR *BASELINE* AND GDEB FRONT-END, APPLYING DIFFERENT SCORE NORMALISATION TECHNIQUES (FEMALE'S DEVELOPMENT SET) ..... 295
FIGURE 5-52 MALE'S DET CURVES ON HESPERIA EVALUATION SET, WITHOUT APPLYING ANY SCORE NORMALISATION TECHNIQUE ..... 298
FIGURE 5-53 MALE'S DET CURVES ON HESPERIA EVALUATION SET, APPLYING ZNORM ..... 298

FIGURE 5-54 MALE'S DET CURVES ON HESPERIA EVALUATION SET, APPLYING TNORM ..... 299
FIGURE 5-55 MALE'S DET CURVES ON HESPERIA EVALUATION SET, APPLYING ZTNORM ..... 299
FIGURE 5-56 FEMALE'S DET CURVES ON HESPERIA EVALUATION SET, WITHOUT APPLYING ANY SCORE NORMALISATION TECHNIQUE ..... 300
FIGURE 5-57 FEMALE'S DET CURVES ON HESPERIA EVALUATION SET, APPLYING ZNORM ..... 300
FIGURE 5-58 FEMALE'S DET CURVES ON HESPERIA EVALUATION SET, APPLYING TNORM ..... 301
FIGURE 5-59 FEMALE'S DET CURVES ON HESPERIA EVALUATION SET, APPLYING ZNORM ..... 301
FIGURE 5-60 *HEER* OBTAINED DEPENDING ON THE NUMBER OF MFCCS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 304
FIGURE 5-61 *HEER* OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 305
FIGURE 5-62 *HEER* OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF Δ AND ΔΔ (GIC – DEVELOPMENT SET) ..... 306
FIGURE 5-63 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF MFCC AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 307
FIGURE 5-64 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF MFCC AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 307
FIGURE 5-65 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 308
FIGURE 5-66 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 308
FIGURE 5-67 $EER_M$ (BLUE) AND $EER_F$ (RED) OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF Δ AND ΔΔ (GDC – DEVELOPMENT SET) ..... 309
FIGURE 5-68 DET CURVE FOR CLASSIC PARAMETERS ON ALBAYZIN MALE DEVELOPMENT SET FOR GIC AND GDC ..... 311
FIGURE 5-69 DET CURVE FOR CLASSIC PARAMETERS ON ALBAYZIN FEMALE DEVELOPMENT SET FOR GIC AND GDC ..... 311
FIGURE 5-70 DET CURVES COMPARING DIFFERENT SETS OF PARAMETERS UNDER GIC AND GDC WITHOUT EXTENDED BIOMETRICS ON THE MALE DEVELOPMENT SET ..... 314
FIGURE 5-71 DET CURVES COMPARING DIFFERENT SETS OF PARAMETERS UNDER GIC AND GDC WITHOUT EXTENDED BIOMETRICS ON THE FEMALE DEVELOPMENT SET ..... 314
FIGURE 5-72 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON ALBAYZIN DEVELOPMENT SET FOR MALE SPEAKERS ..... 317
FIGURE 5-73 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON ALBAYZIN DEVELOPMENT SET FOR FEMALE SPEAKERS ..... 317
FIGURE 5-74 INFLUENCE OF THE GSE CONFIGURATION ON THE $EER_M$ (DEVELOPMENT SET) ..... 318
FIGURE 5-75 INFLUENCE OF THE GSE CONFIGURATION ON THE $EER_F$ (DEVELOPMENT SET) ..... 318
FIGURE 5-76 DET CURVES FOR *BASELINE* AND GDEB FRONT-END, APPLYING DIFFERENT SCORE NORMALISATION TECHNIQUES (MALE DEVELOPMENT SET) ..... 322
FIGURE 5-77 DET CURVES FOR *BASELINE* AND GDEB FRONT-END, APPLYING DIFFERENT SCORE NORMALISATION TECHNIQUES (FEMALE DEVELOPMENT SET) ..... 323

FIGURE 5-78 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON ALBAYZIN DEVELOPMENT SET FOR MALE SPEAKERS AND ZTNORM ..... 324
FIGURE 5-79 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON ALBAYZIN DEVELOPMENT SET FOR FEMALE SPEAKERS AND TNORM ..... 324
FIGURE 5-80 MALE DET CURVES ON ALBAYZIN EVALUATION SET, WITHOUT APPLYING ANY SCORE NORMALISATION TECHNIQUE ..... 327
FIGURE 5-81 MALE DET CURVES ON ALBAYZIN EVALUATION SET, APPLYING ZNORM ..... 327
FIGURE 5-82 MALE DET CURVES ON ALBAYZIN EVALUATION SET, APPLYING TNORM ..... 328
FIGURE 5-83 MALE DET CURVES ON ALBAYZIN EVALUATION SET, APPLYING ZTNORM ..... 328
FIGURE 5-84 FEMALE DET CURVES ON ALBAYZIN EVALUATION SET, WITHOUT APPLYING ANY SCORE NORMALISATION TECHNIQUE..... 329
FIGURE 5-85 FEMALE DET CURVES ON ALBAYZIN EVALUATION SET, APPLYING ZNORM ..... 329
FIGURE 5-86 FEMALE DET CURVES ON ALBAYZIN EVALUATION SET, APPLYING TNORM ..... 330
FIGURE 5-87 FEMALE DET CURVES ON ALBAYZIN EVALUATION SET, APPLYING ZTNORM..... 330
FIGURE 5-88 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF MFCCS (GDC – DEVELOPMENT SET) AND FOR DIFFERENT SCORE NORMALIZATION ALGORITHMS .... 333
FIGURE 5-89 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF MFCCS (GDC – DEVELOPMENT SET) AND FOR DIFFERENT SCORE NORMALIZATION ALGORITHMS .... 334
FIGURE 5-90 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS IN THE FILTER BANK (GDC – DEVELOPMENT SET) AND FOR DIFFERENT SCORE NORMALIZATION ALGORITHMS ..... 335
FIGURE 5-91 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF FILTERS IN THE FILTER BANK (GDC – DEVELOPMENT SET) AND FOR DIFFERENT SCORE NORMALIZATION ALGORITHMS ..... 335
FIGURE 5-92 $EER_M$ OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF DIFFERENT SCORE NORMALIZATIONS (GDC – DEVELOPMENT SET)..... 336
FIGURE 5-93 $EER_F$ OBTAINED DEPENDING ON THE NUMBER OF GAUSSIANS AND THE USE OF DIFFERENT SCORE NORMALIZATIONS (GDC – DEVELOPMENT SET)..... 336
FIGURE 5-94 DET CURVE FOR CLASSIC PARAMETERS ON MOBIO MALE DEVELOPMENT SET FOR GIC AND GDC ..... 339
FIGURE 5-95 DET CURVE FOR CLASSIC PARAMETERS ON MOBIO FEMALE DEVELOPMENT SET FOR GIC AND GDC ..... 339
FIGURE 5-96 $EER_M$ OBTAINED FOR THE TESTS WHICH INCORPORATE E, ΔE, F0 AND F3 IN THE FEATURE VECTORS..... 340
FIGURE 5-97 $EER_F$ OBTAINED FOR THE TESTS WHICH INCORPORATE E, ΔE, F0 AND F3 IN THE FEATURE VECTORS..... 341
FIGURE 5-98 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON MOBIO'S DEVELOPMENT SET FOR MALE SPEAKERS AND ZTNORM ..... 345
FIGURE 5-99 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON MOBIO'S DEVELOPMENT SET FOR FEMALE SPEAKERS AND NO NORM..... 345
FIGURE 5-100 INFLUENCE OF GSE CONFIGURATION ON THE $EER_M$ (DEVELOPMENT SET) ..... 346
FIGURE 5-101 INFLUENCE OF GSE CONFIGURATION ON THE $EER_F$ (DEVELOPMENT SET) ..... 346

FIGURE 5-102 MALE DET CURVES ON MOBIO EVALUATION SET, APPLYING ZTNORM 348
FIGURE 5-103 FEMALE DET CURVES ON MOBIO EVALUATION SET, APPLYING ZTNORM ..... 348
FIGURE 5-104 $EER_M$ OBTAINED DEPENDING ON THE CLASSIFIER - NIST SRE 2010, COND-1 TO COND-9 ..... 352
FIGURE 5-105 $EER_F$ OBTAINED DEPENDING ON THE CLASSIFIER - NIST SRE 2010, COND-1 TO COND-9 ..... 352
FIGURE 5-106 DET CURVES FOR MALE SPEAKERS UNDER DIFFERENT CONDITIONS COMPARING THE USE $\Delta\Delta$ COEFFICIENTS ..... 356
FIGURE 5-107 DET CURVES FOR FEMALE SPEAKERS UNDER DIFFERENT CONDITIONS COMPARING THE USE $\Delta\Delta$ COEFFICIENTS ..... 358
FIGURE 5-108 $EER_M$ *AND* *ACER* OBTAINED DEPENDING ON THE *BASELINE* FRONT-END SETUP NIST SRE 2010 ..... 360
FIGURE 5-109 $EER_F$ *AND* *ACER* OBTAINED DEPENDING ON THE *BASELINE* FRONT-END SETUP NIST SRE 2010 ..... 361
FIGURE 5-110 *ACER* OBTAINED FOR GDC WHEN EXTRA PARAMETERS ARE INCORPORATED INTO THE FEATURE VECTOR AND SNORM IS APPLIED FOR MALE SPEAKERS ..... 363
FIGURE 5-111 $EER_M$ COMPARISON FOR THE DIFFERENT CONDITIONS IN WHICH DEVELOPMENT SET IS DIVIDED ..... 363
FIGURE 5-112 *ACER* OBTAINED FOR GDC WHEN EXTRA PARAMETERS ARE INCORPORATED INTO THE FEATURE VECTOR AND SNORM IS APPLIED FOR FEMALE SPEAKERS ..... 364
FIGURE 5-113 *ACER* OBTAINED FOR GDC WHEN EXTRA PARAMETERS ARE INCORPORATED INTO THE FEATURE VECTOR AND SNORM IS COMBINED WITH ZTNORM FOR FEMALE SPEAKERS ..... 364
FIGURE 5-114 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON THE NIST SRE10 DEVELOPMENT SET FOR MALE SPEAKERS AND SNORM ..... 371
FIGURE 5-115 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON THE NIST SRE10 DEVELOPMENT SET FOR FEMALE SPEAKERS AND SNORM ..... 372
FIGURE 5-116 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON THE NIST SRE10 DEVELOPMENT SET FOR FEMALE SPEAKERS AND SNORM+ZTNORM . 374
FIGURE 5-117 NIST SRE12 FILES WITH FAKE ADDITIVE NOISE (UP - TIME DOMAIN, DOWN FREQUENCY DOMAIN) ..... 378
FIGURE 5-118 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON THE NIST SRE12 EVALUATION SET FOR MALE SPEAKERS AND SNORM ..... 379
FIGURE 5-119 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON THE NIST SRE12 EVALUATION SET FOR FEMALE SPEAKERS AND SNORM ..... 380
FIGURE 5-120 DET CURVES COMPARING CLASSICAL PARAMETERS AND GDEB ON THE NIST SRE12 EVALUATION SET FOR FEMALE SPEAKERS AND SNORM+ZTNORM .... 381

# RESUMEN

La cuestión principal abordada en esta tesis doctoral es la mejora de los sistemas biométricos de reconocimiento de personas a partir de la voz, proponiendo el uso de una nueva parametrización, que hemos denominado parametrización biométrica extendida dependiente de género (GDEBP en sus siglas en inglés). No se propone una ruptura completa respecto a los parámetros clásicos sino una nueva forma de utilizarlos y complementarlos. En concreto, proponemos el uso de parámetros diferentes dependiendo del género del locutor, ya que como es bien sabido, la voz masculina y femenina presentan características diferentes que deberán modelarse, por tanto, de diferente manera. Además complementamos los parámetros clásicos utilizados (MFFC extraídos de la señal de voz), con un nuevo conjunto de parámetros extraídos a partir de la deconstrucción de la señal de voz en sus componentes de fuente glótica (más relacionada con el proceso y órganos de fonación y por tanto con características físicas del locutor) y de tracto vocal (más relacionada con la articulación acústica y por tanto con el mensaje emitido).

Para verificar la validez de esta propuesta se plantean diversos escenarios, utilizando diferentes bases de datos, para validar que la GDEBP permite generar una descripción más precisa de los locutores que los parámetros MFCC clásicos independientes del género. En concreto se plantean diferentes escenarios de identificación sobre texto restringido y texto independiente utilizando las bases de datos de HESPERIA y ALBAYZIN. El trabajo también se completa con la participación en dos competiciones internacionales de reconocimiento de locutor, NIST SRE (2010 y 2012) y MOBIO 2013. En el primer caso debido a la naturaleza de las bases de datos utilizadas se obtuvieron resultados cercanos al estado del arte, mientras que en el segundo de los casos el sistema presentado obtuvo la mejor tasa de reconocimiento para locutores femeninos.

A pesar de que el objetivo principal de esta tesis no es el estudio de sistemas de clasificación, sí ha sido necesario analizar el rendimiento de diferentes sistemas de clasificación, para ver el rendimiento de la parametrización propuesta. En concreto, se ha abordado el uso de sistemas de reconocimiento basados en el paradigma GMM-UBM, supervectores e *i*-vectors.

Los resultados que se presentan confirman que la utilización de características que permitan describir los locutores de manera más precisa es en cierto modo más importante que la elección del sistema de clasificación utilizado por el sistema. En este sentido la parametrización propuesta supone un paso adelante en la mejora de los sistemas de reconocimiento biométrico de personas por la voz, ya que incluso con sistemas de clasificación relativamente simples se consiguen tasas de reconocimiento realmente competitivas.

**Palabras clave:** biometría, voz, reconocimiento de locutor, caracterización de locutor, Procesamiento digital de señal, fuente glótica, tracto vocal, separación fuente-tracto, GMM-UBM, Supervectores, *i*-vectors, Competiciones de reconocimiento de locutor.

# ABSTRACT

The main question addressed in this thesis is the improvement of automatic speaker recognition systems, by the introduction of a new front-end module that we have called Gender Dependent Extended Biometric Parameterisation (GDEBP). This front-end do not constitute a complete break with respect to classical parameterisation techniques used in speaker recognition but a new way to obtain these parameters while introducing some complementary ones. Specifically, we propose a gender-dependent parameterisation, since as it is well known male and female voices have different characteristic, and therefore the use of different parameters to model these distinguishing characteristics should provide a better characterisation of speakers. Additionally, we propose the introduction of a new set of biometric parameters extracted from the components which result from the deconstruction of the voice into its glottal source estimate (close related to the phonation process and the involved organs, and therefore the physical characteristics of the speaker) and vocal tract estimate (close related to acoustic articulation and therefore to the spoken message). These biometric parameters constitute a complement to the classical MFCC extracted from the power spectral density of speech as a whole.

In order to check the validity of this proposal we establish different practical scenarios, using different databases, so we can conclude that a GDEBP generates a more accurate description of speakers than classical approaches based on gender-independent MFCC. Specifically, we propose scenarios based on text-constrain and text-independent test using HESPERIA and ALBAYZIN databases. This work is also completed with the participation in two international speaker recognition evaluations: NIST SRE (2010 and 2012) and MOBIO 2013, with diverse results. In the first case, due to the nature of the NIST databases, we obtain results closed to state-of-the-art although confirming our hypothesis, whereas in the MOBIO SRE we obtain the best simple system performance for female speakers.

Although the study of classification systems is beyond the scope of this thesis, we found it necessary to analise the performance of different classification systems, in order to verify the effect of them on the propose parameterisation. In particular, we have addressed the use of speaker recognition systems based on the GMM-UBM paradigm, supervectors and *i*-vectors.

The presented results confirm that the selection of a set of parameters that allows for a more accurate description of the speakers is as important as the selection of the classification method used by the biometric system. In this sense, the proposed parameterisation constitutes a step forward in improving speaker recognition systems, since even when using relatively simple classification systems, really competitive recognition rates are achieved.

**Key Words:** Biometry, Voice, Speaker Recognition, Speaker characterisation, Digital Signal Processing, Glottal Source, Vocal Tract, Source-Tract Separation, GMM-UBM, Supervectors, *i*-vectors, Speaker Recognition Evaluation

# ACKNOWLEDGEMENTS

It is quite difficult to compile a list that gathers all the people who in one way or another have played an important role in the achievement of this thesis. For this reason, I want to thank all of them for their continued support and especially for the shared time. However, special mention deserves:

Professor Pedro Gómez Vilda, for its confidence at the beginning of this long journey, for his unconditional and continuous support after so many years of collaboration and in particular for sharing his extensive knowledge with me.

Professor Agustín Álvarez Marquina, for agreeing to lead this thesis, for putting up with multiple delays during the development of the thesis, for his master lessons on classification systems and last but not least for the great moments at the office.

All members of GIAPSI, the list is large as it is my appreciation and admiration for all of you.

My parents who have understood and encouraged my dedication to study, as well as my sister for being there whenever I needed her.

Coral who is certainly the one that most has suffered the effort and dedication devoted to this thesis. Thank you for your time, for your support, for your patience and for Manuel.

# 1 INTRODUCTION

## 1.1 MOTIVATIONS

Since its introduction in the 1980s, the term *globalisation* and all that it entails has become part of our everyday lives. More specifically, globalisation has had a considerable impact in the following areas:

- Trade: Countries have increased their share of world trade, varying not only the type of products exported but also the countries to which these products are exported.
- Capital movements: The increasing integration of economies around the world has also led to an increase in the capital flows, especially caused by the decentralisation of investments.
- Movement of people: There are several reasons that explain the increase in the movement of people. First of all, workers move from one area, city or country to another, trying to find better employment opportunities. In addition, there has also been an increase in the amount of people moving for cultural or touristic reasons. Last but not least, there are also movements of people on grounds of crime and war.
- Spread of knowledge and technology: Closely related with the three previous aspects is the spread of information, knowledge and technology. For instance capital movements in the form of foreign investments result in an exchange of knowledge and technical innovation, especially across international borders. Furthermore, social networks have proliferated mainly due to the human need to share their experiences, and also to keep in touch with known people anywhere in the world.

Although some people regard this process not only as inevitable and irreversible but also beneficial, others show some rejection and even fear of this new scenario. No matter whether we consider globalisation as a positive or negative, the fact is that all the changes that it has caused have led to new needs as far as security is concerned. Given the decentralisation of business facilities, buildings and services, the increase of international trade, the exchange of information and technology, it is essential to have systems that allow more precise control over the access to these resources. The main goal is to more precisely control who can access certain areas or facilities, who can access certain information or who is allowed to interact with a particular system. Additionally, the set of attacks that have taken place in the Western World (9/11 2001 NY, 11/03 2004 Madrid, 7/7 2005 London) has also triggered a demand in homeland security, not only in border control, but also in governmental facilities and buildings, public transports (airports, railway, bus and metro stations), leisure centres, civil engineering works, etc.

In this scenario, which can be characterised as “security-obsessed world”, biometrics and more specifically biometric recognition has become an area of great research activity, on the one hand since it turns out to be an important technique for human-machine interaction in applications with security considerations, on the other hand due to its application to multiple everyday security situations.

One of the most important goals of security systems is to be able to reliably establish the identity of the people that is in a place under surveillance or the people with which a system is interacting. In this sense, biometric authentication defined as the automatic recognition of individuals based on some specific characteristics that can be physiological or behavioural [Dugelay,2002], should provide more accurate recognition results that the ones provided by classical security systems. In other words, biometric recognition systems are able to identify a person based on what he/she is rather than on something he/she possesses (e.g. ID card) - and can be stolen or misplaced – or something he/she remembers (e.g. password) – and can be forgotten. As physiological characteristics we can cite, among others, DNA, fingerprint, palm print, iris, odour, retina, etc. In the case of behavioural characteristics, the most popular are: gait, signature, and keystroke. Finally voice, can be considered as a combination of physiological and behavioural characteristics, because as it will be seen later, the production of voice is closely related to specific organs and tracts (lungs, vocal cords, vocal tract, etc.) but at the same time, the speech of a person changes over the time due to social relations, age, emotional or medical state, etc. [Jain,2004-B], [Weaver,2006].

In the last years, biometric systems applied to security environments have been successfully installed (for instance, US-VISIT Biometric Identification Services). However, this does not imply that biometric recognition applied to security is a fully solved problem. More specifically, the obsession for security has also sparked off a reject on people, because most times security systems bother them. In the case of biometric systems, rejection can be even stronger for different reasons: they can be regarded as a violation of civil liberties (for instance, video surveillance for face or gait recognition), they can be regarded as unsafe (for instance, in the case of iris and retina scanners) or unhealthy (use of touchable sensors to acquire fingerprints) due to the type of sensor that is used to capture the specific biometric, or they can be unaffordable due to its deployment costs. For these reasons, researchers are making an effort in developing biometric security systems less intrusive or “more friendly” for people using them without reducing their capacity of surveillance and recognition.

As briefly noted, each biometric characteristic has its strengths and weaknesses, and voice is not an exception. The major strengths of voice biometric recognition systems are manifold:

- Acceptance → Voice is considered as the quintessential vehicle of communication between humans, so people use it almost at every time and in every place. For this reason, people do not consider speaking in front of a microphone to determine their identity as an exceptional act or a meddling in their everyday lives as opposed to the use of other biometric features.
- Safety and health → As already said the acquisition of this biometric characteristic can be carried out using a simple microphone or via telephonic devices. As a result, its acquisition is considered safe and hygienic by users as sensors are passive and not need to be touched.
- Cost → Due to the technology used to capture the biometric characteristic and the low requirements in terms of computation and specific hardware, voice biometric recognition systems are preferred over other biometric systems.

However, taking into account that voice reflects both physiological and behavioural characteristics of the speaker, it suffers from several limitations that are still fully unsolved. As far as physiological aspects are concerned, speaker recognition is affected

for instance by the process of aging (due to changes in organs and tracts involved in voice production or to hormonal changes), or/and by the medical state (no matter whether temporal – flu, cold – or permanent – tracheotomy, Alzheimer, etc.), and even by the gender of the speakers. In the case of behavioural aspects, voice can be affected by the emotional state (euphoric, sad, angry) of the speaker, or by social relations (differences between working environment and leisure environment). With all these variables affecting the voice of a person, no security system has been yet designed that completely relies on a voice biometry, because its performance is not good enough and still needs to be improved. Nevertheless, speaker recognition plays an important role in different scenarios such as: control access to facilities, infrastructures and restricted areas by the use of biometric signature of voice (speaker identification), speaker tracking by the identification of specific vocal features, access to information systems and databases with multimodal registries (video linked to voice) addressable by vocal context (speaker verification), automatic indexing, search and information retrieval in audio sources, etc.

Taking into account all the information provided so far, we have considered the need to improve speaker recognition systems applied to security environments for the following reasons:

- Security → Despite all the advances that have been seen in recent years related to security, this is still an open issue for which no global solution has been achieved. From this point of view, improvements in a particular system such as speaker recognition will always be welcomed as they ultimately help to improve the entire security system.
- Cost → As previously pointed out, the development of a speaker recognition system does not require large outlay of funds due to limited computational and hardware specific requirements if compared with biometric systems based on other biometric characteristic.
- Acceptability → As noted before, although most users tends to reject biometric security systems for multiple reasons, the ones based on voice are the most accepted among the others. Additionally, due to cost restrictions they are also more accepted by companies that want to implement security measures but have a limited budget.
- Usage → another important reason to improve speaker recognition systems is the fact that any improvement in this area can be applied to multiple scenarios.
- Open problem → Despite the great advances that the area has experimented in the last years, speaker recognition is not a fully solve problem. Major advances in the area have resulted from improvements in classification systems, since the parameters used to characterise a speaker have not changed significantly since early studies in the area. Due to the number of variables that can affect the voice, it seems necessary to find a new set of parameters that allow for a better characterisation of speakers limiting the effect of these variables, instead of - or complementary to - the classical parameters originally intended to model the speech instead of the speaker as it will be shown later on.

The search for a new set of parameters that allows a better characterisation of speakers from a biometric point of view should result in a significant improvement in the performance of speaker recognition systems, and therefore it would be possible to reliably use them in security environments.

## 1.2 OBJECTIVES

The starting point of the present research is the proven fact that voice can be used for automatic recognition of people, not only in laboratory scenarios but also in security environments as derived from the existence of commercial applications. Additionally, due to new social and macro-economic scenarios voice-biometrics have suffered a growing demand especially in security environments

Since the early researches in the late 1970's [Atal,1976] there have been numerous publications that have proposed speaker recognition systems. A comprehensive review of the advances in the area can be found in [Campbell,1997], [Kinnunen,2009]. However, the performance of current speaker recognition applications is not as robust and reliable as other biometric feature applications mainly due to the fact that main improvements in the area derive from improvements in the area of classification and normalisation rather than improvements in the parameterisation that allows a better characterisation of speakers.

Starting from the initial hypothesis and the prior consideration on the reasons for progress in the area, the hypothesis underpinning this thesis can be stated as follows:

> *It is possible to design a more reliable biometric recognition system based on voice, by defining a set of gender-dependent biometric features extracted from voice that allows for a better speaker characterisation, without significantly affecting computational requirements.*

Without taking into account classification methods, although different alternatives will be explored, current trends in speaker recognition deals with the fusion of high and low level features. However, classical parameters derived from the power spectral density of speech without considering the gender of speakers are still the most used ones. The objective of the present research is on incorporating additional information extracted from glottal source (more related to biometry) and vocal tract (more related with message) into classical speaker recognition systems. According to a simplified speech production model, voice can be regarded as the result of filtering an excitation signal with the transfer function of the vocal tract and the lip radiation model. Lot of attention has been paid to feature extraction from the vocal tract characteristics, mainly because they are also useful to speech characterisation. However, the excitation signal has not been deeply studied and used although it has been proven to be very effective for example in voice pathology and gender detection [Gómez,2007-A], [Gómez,2004-B], [Gómez,2005-B], or even in speaker characterisation [Plumpe,1999], [Zheng,2006].

Another interesting aspect is the one related to gender of speakers. It is well-known that male and female voices exhibit not only acoustic-phonetic differences but physiological variations as well [Whiteside,2001]. However, to date and to the author's knowledge, not any speaker recognition system has proposed a gender-dependent parameterisation approach.

The main objective is to demonstrate that a gender-dependent parameterisation which incorporates biometric information about the speaker, mainly extracted from the glottal source, will improve the performance of speaker recognition systems. To achieve this we set the following specific objectives:

- Selection of a reference algorithm set (both for classification and parameterisation) in biometric recognition systems based on voice.

- Construction of a biometric recognition system based on voice which implements the previously selected algorithms, and which provides reliable recognition data for a comparative analysis with other systems.
- Definition, implementation and calibration of a *voice deconstruction* algorithm to split voice signal in its glottal source and vocal tract estimations. Previous works on this topic are [Alku,1992], [Akande,2005].
- Selection of the set of features that best represent the previously cited components of voice that allow a more precise characterisation of speakers depending on its gender. LPC and MFCC are classical parameters for vocal tract characterisation, but for the vocal source estimation an efficient set of parameters has not been proposed yet. For example, there is a first attempt to used a time-domain characterisation of the glottal source in [Plumpe,1999]; others propose a frequency characterisation as in [Gudnason,2008] or a time-frequency feature extraction approach as in [Zheng,2004].
- Establishing a framework for reliably comparing the performance of different speaker recognition systems.
- Evaluating the performance of gender-dependent extended biometric recognition systems based on deconstruction of voice and the state-of-the-art speaker recognition system.
- Establishing the benefits and cost of using the proposed parameterisation compared to classical approaches.
- Applying the developed system in practical scenarios such as speaker recognition evaluations.

## 1.3 A GUIDE TO THE TEXT

The thesis is organised as follow:

- Chapter 1: The remainder of this chapter is devoted to make a brief review of the state of the art in the field of biometric recognition of people. The review will cover the following aspects: basic structure of biometric recognition systems, biometric characteristics mostly used in specific implementations, and biometric databases available for research purposes and commercial applications. As the aim of the present thesis is to define a new parameterisation scheme that provides a better characterisation of speakers, a specific review of the state of the art in the area of speaker recognition will be also performed. It will cover the different parts that comprise a recognition system of this type: feature extraction, class modelling and classification methods, and class decision. Additionally, the recognition rates achieved by state-of-the-art speaker recognition systems will be provided and somehow compared, taking into account that it is not possible to directly compare the results obtained by different systems. Finally, we present some commercial solutions that implement voice recognition systems.
- Chapter 2: The first contribution of the thesis is presented in this chapter. First, we will introduce the speech production model and the corresponding characterisation by physical models. A brief review of the acoustic theory of speech production will be presented before moving on to describe the proposed algorithm that allows us to perform what we have called *deconstruction of voice*,

i.e. the separation of the voice signal into the glottal and vocal tract component. Finally, we present a gender-dependent parameterisation that incorporates both components to characterise the speakers more accurately.

- Chapter 3: This chapter is entirely devoted to present classification methods that have been successfully applied to solve the speaker classification problem. A review of classical generative (HMMs and GMMs) and discriminative models (SVMs) will be covered as well as new trends in the area such as supervectors and *i*-vectors. Normalisation techniques applied in some scenarios will also be presented. Fusion strategies are also presented at the end of the chapter.
- Chapter 4: This chapter will be used to describe the test bench available to determine the technological level of biometric recognition systems using voice, analyzing its strengths and weaknesses. We will also describe a specifically defined test bench that use a specific database collected under strict conditions as far as quality is concerned. This test bench will allow us to prove the validity of the proposed parameterisation under, among others, two different scenarios: spontaneity (NIST SRE) and standardisation (HESPERIA). We will also present some test using ALBAYZIN and MOBIO databases.
- Chapter 5: The practical experiments carried out throughout the thesis are included in this chapter. We will describe the official results of the NIST SRE 2010, along with the system developed to participate in it and the results achieved with the improvements applied thanks to the acquired experience. The results obtained in the set of tests defined in chapter 4 will also be presented.
- Chapter 6: The last chapter is dedicated to present the conclusions that can be drawn from the work done throughout this thesis, and to suggest new lines for future research. A complete review of the contributions to advance the scientific state of the art in speaker recognition is included as well as the published papers and congress contributions that have been made in the course of the research leading to this thesis are listed.

As appendices we have included a glossary of terms and acronyms, as well as a summary of software tools developed during the thesis. Finally the bibliographic references cited throughout this document are listed in alphabetic order by first author's family name.

## 1.4 BIOMETRIC AUTHENTICATION

Traditionally, the identification process between a person and a system, has been mainly based on something the person knows (e.g. password) or something he/she owns (e.g. smartcard), leaving person's biometrics out for forensics or law enforcement area (e.g. criminal identification, claim of parental relations, etc.). However, the use of passwords or ID cards entails some problems to person recognition. They can be more or less easily guessed, shared, misplaced or copied, compromising the security of the entire system (*"The security of the entire system is only as good as the weakest password"* [Jain,2004-B]). Therefore, as biometrics offer a natural and reliable way to identify a person based on some physical and/or behavioural characteristic (e.g. based on who you are); they are increasingly playing an important role in human-machine interaction for security purposes, both in civilian and military applications. [Ross,2007], [Weaver,2006], [Jain,2004-B]. Moreover, it is usual to combine these three different types of authentication especially in security-sensitive applications.

The design of an automatic or semi-automatic system to recognised individuals, based on biometrics, relies mainly on signal processing techniques and pattern recognition. Thus, a biometric system can be build using the following modules:

- Data acquisition module: This module includes a sensor which captures the biometric data of an individual. The physical information acquired by the sensor is transformed into a digital format which allows a computational treatment.
- Feature extraction module: The aim of this module is to extract a set of discriminatory features of the trait for recognition captured by the data acquisition module. This set of features must show high inter-class variability while keeping the intra-class variability quite low. Additionally, the number of features in the feature set must be kept low as the complexity of the system increases as the number of attributes rise. The compact but expressive representation of an individual based on this set of features is also known as template.
- Database module: This module can be implemented as a central database or as a personal device (such as a smartcard) in which the templates of the enrolled users of the system (typically, individuals with access) are stored.
- Matcher module: During the recognition step the features, extracted from the input, are compared against the templates stored in the database, providing a matching score. The matching score quantifies the similarity between the input and the template. The higher the score, the more certain is the system that the two templates come from the same person. This module also encapsulates a decision-making phase, in which a user's claimed identity is verified or a user's identity is established based on the matching scores.

A biometric recognition system can work in two different modes, verification or identification. In the verification mode, a one-to-one comparison is performed to determine whether the captured data corresponds to the stored biometric template of the person's claimed identity. In the case of identification mode, a one-to-all comparison is performed to try to establish the identity of a person, by searching the templates of all enrolled users. Regardless the operation mode, it is necessary to perform an enrolment phase in which a template or model is generated for all the users to whom we want to grant access to the system. Figure 1-1 summarises the different stages involved in a biometric recognition system.

Regardless the biometric characteristic used for biometric recognition, there are some important research issues that should be addressed:

- How to acquire the biometric data: Type of sensors, controlled/uncontrolled acquisition conditions, etc.
- Feature selection: How to extract intra-class invariant features while keeping the inter-class variance in such a level that allows class separation, dimensionality reduction to allow an internal representation computationally affordable, etc.
- How to implement the matching metric: A matching algorithm should provide a measure of similarity between two patterns. The pattern recognition problem can be addressed using two different approaches: generative classifiers and discriminative classifiers. Generative methods learn a density distribution for each of the classes from 'a priori' information. On the other hand discriminative methods model the decision boundary between classes.

- The representation scheme and the matching metric determine the accuracy performance of the system.

The system's performance is a critical issue as it provides a quantitative measure of system reliability. Through this measure, we are able to establish a comparison between different biometric recognition systems. Unfortunately, the performances of different systems are only comparable in the context of a given database in a specific train/test environment, that's the reason of the existence of different biometric recognition competitions, which provides a common test framework for all the tested systems.

The performance of the system is evaluated in relation with its ability to classify a template as belonging to a genuine or known individual or belonging to an impostor. Given an input template, we can build two different hypotheses:

- $H_0$ → The input template does not belong to a known individual (belongs to an impostor)
- $H_1$ → The input template belongs to a known individual

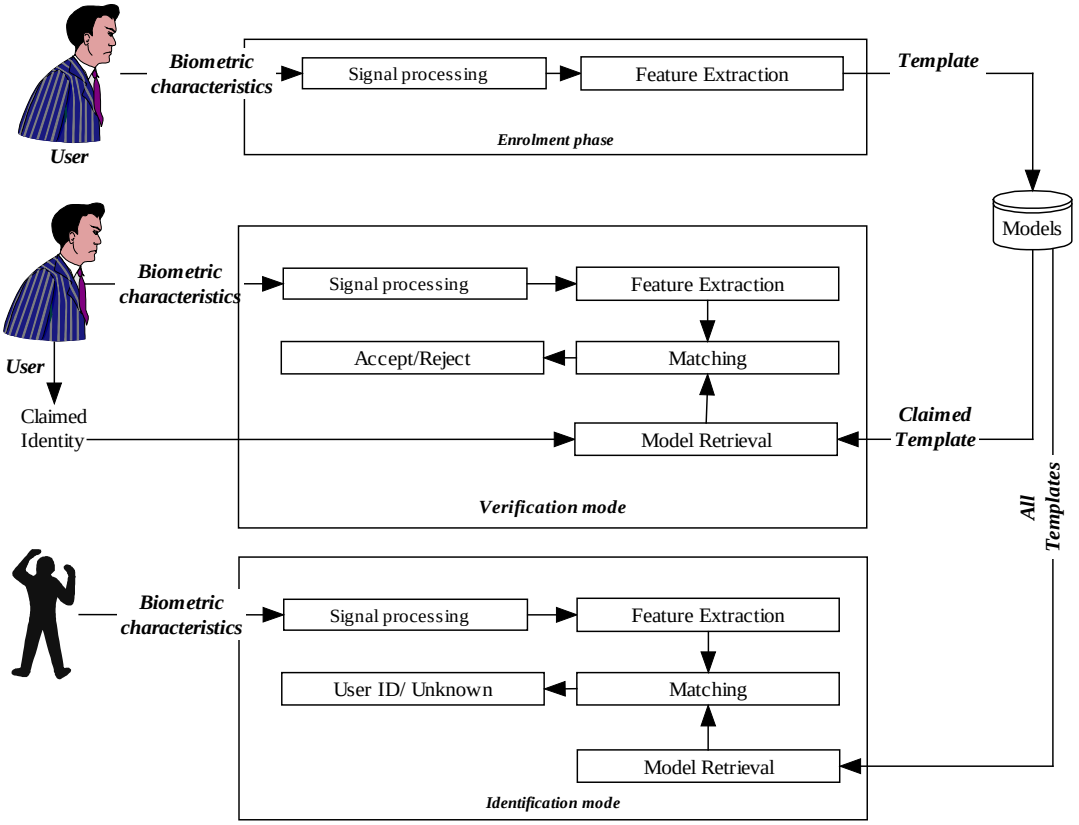



**Figure 1-1** Operation modes of a biometric recognition system

Given these two hypotheses, the system can generate four possible outcomes: (a) a genuine user is accepted, (b) a genuine user is not correctly recognised, (c) an impostor is accepted as a genuine user and (d) an impostor is rejected. Cases (a) and (d) are considered correct outputs whereas (b) and (c) are considered recognition errors. Type (b) errors are also known as false non-match errors or false rejection (*FR*), while type (c) errors are known as false match or false acceptance (*FA*) error. The error rates (*FAR* and *FRR*) depend upon the threshold, θ, fixed for accepting or rejecting a hypothesis, thus there is always a trade-off between false match rates and false non-match rates in every security system.

These error rates can be used as a measure of performance and can be represented using different curves:

- *Score Distribution Curves*: This graph represents the probability density function (pdf) for genuine users (true acceptance), and for impostors (true rejection). The overlapping area between both distributions characterises the false-rejections and false acceptance errors. Selecting a threshold in the intersection between both pdf gives the minimum recognition error. In an ideal situation, in which no classification errors occur, there is no intersection between both distributions.

  As it can be seen in Figure 1-2, *FRR* and *FAR* are function of the operating threshold. That is to say, if the system's threshold is increased to reduce the *FR* errors, then the *FA* errors will be increased, thus reducing the security of the system.

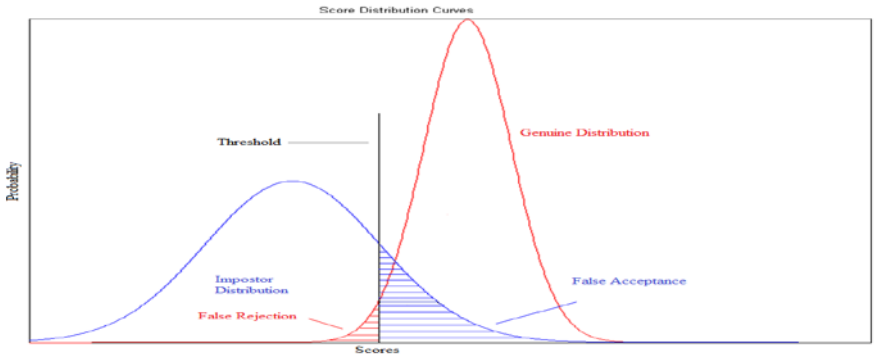



**Figure 1-2** Score Distribution Curves of a hypothetical biometric recognition system

- *ROC Curves*: The original Receiver Operating Characteristic curve plots the true acceptance rate with respect to the false acceptance rate, thus providing an empirical assessment of the system performance at different operating points. However, several researches used it in different ways. For example, false non-match errors are plotted on the horizontal axis, while the false match errors are plotted on the vertical axis, in relation with the selected decision threshold. In other words, the ROC curve is a graph that represents *FRR* and *FAR* for all possible values of the threshold. An interesting introduction to ROC analysis can be found in [Fawcett,2006].

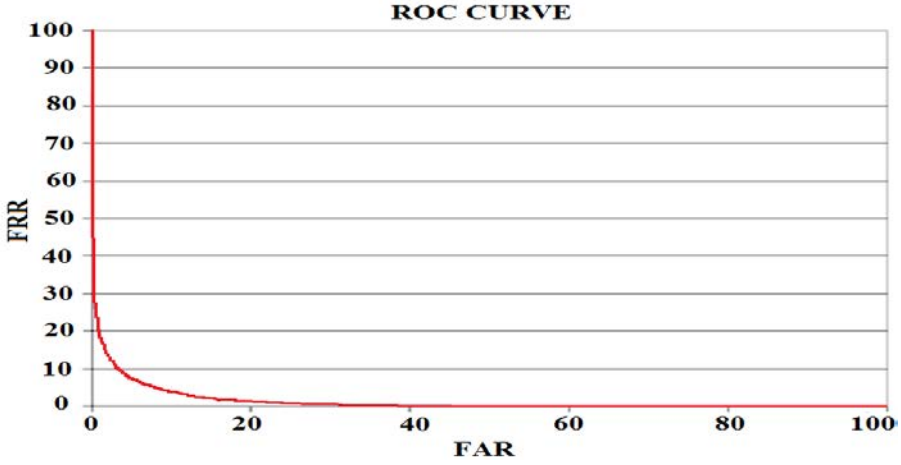



**Figure 1-3** ROC Curve of a hypothetical biometric recognition system

- *DET Curves*: The Detection Error Trade-off curve represents *FRR* and *FAR* on a non-linear scale (normal deviate scale). This type of analysis usually produces plots that are close to linear. The circle marker corresponds to the *EER* point.

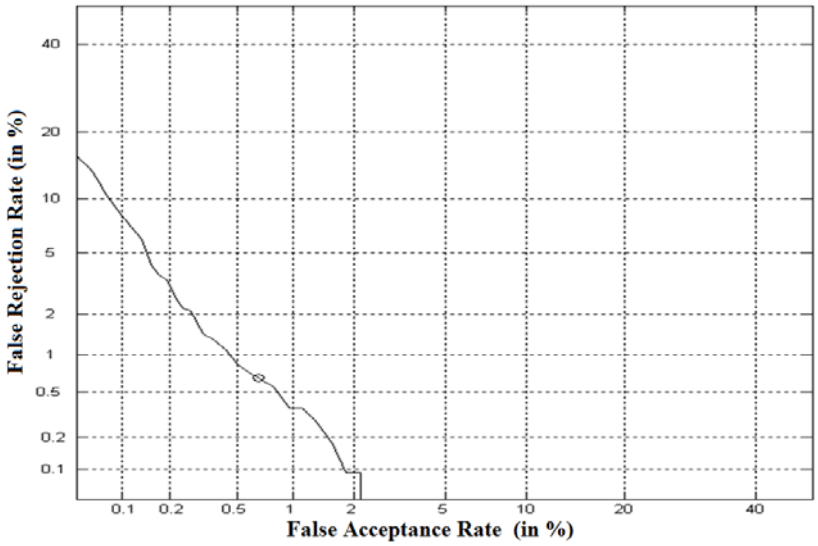



**Figure 1-4** DET Curve of a hypothetical biometric recognition system

- EPC: The Expected Performance Curve [Bengio,2005], provides a method to compare several systems according to a specific performance criterion (*HTER*, *DCF* or *EER*). Having defined the performance measure (which depends on the *FAR(θ)*, *FRR(θ)* and a parameter $\alpha$), the EPC is obtained varying $\alpha$ inside a reasonable range (which depends on the specific criterion). For each $\alpha$, we estimate $\theta$ that minimises the performance criterion on a train set. Finally, the obtained $\theta$ is used to compute the performance measure on the test set.

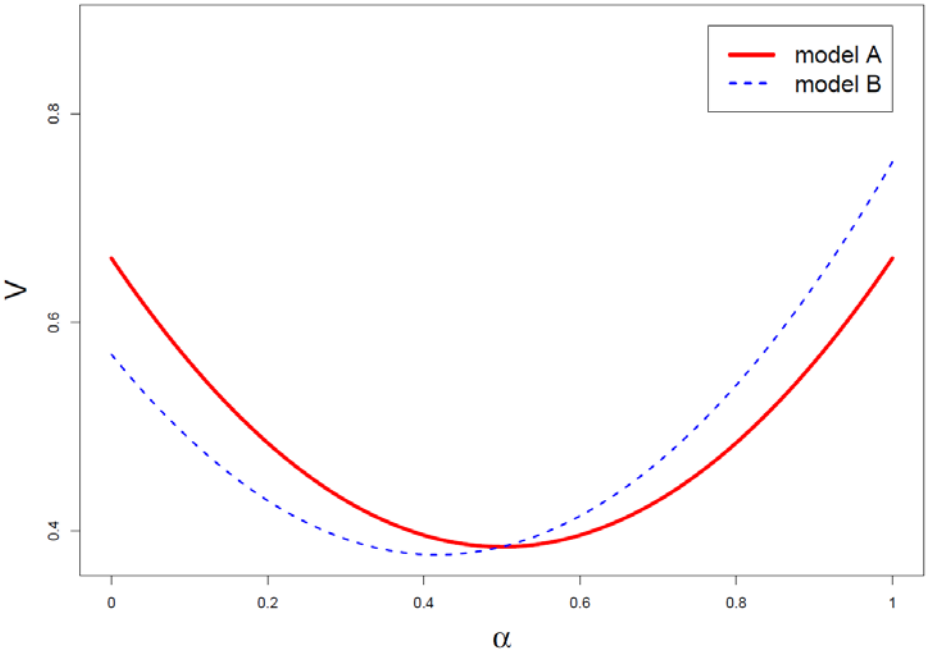



**Figure 1-5** Two artificial EPC curves comparing two models using a performance measure V (Extracted from [Bengio,2005])

Additionally, we can generate single quality measures that allow a quick comparison of performances of different systems:

- *AUC*: This metric represents the Area Under the ROC Curve. According to [Fawcett,2006], *the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance*. An algorithm to compute the AUC can also be found in this study.
- *EER*: The Equal Error Rate corresponds to the threshold for which the FAR and the FRR are equal. It can be mathematically defined as:

$$\theta_{EER} = \underset{\theta}{\operatorname{argmin}} |FAR(\theta) - FRR(\theta)| \quad \textbf{Eq. (1-1)}$$

- *HTER*: Stands for Half Total Error Rate, and can be defined as:

$$HTER = \frac{FAR(\theta) - FRR(\theta)}{2} \quad \textbf{Eq. (1-2)}$$

- *DCF*: Referred also as $C_{Det}$ Cost Function, is a basic metric in the NIST evaluations computing a linear combination of the error rate. $C_{DET}$ can be view as a cost function based on assigned costs for false acceptance and false rejections, and assumed target richness. Cost and richness are arbitrarily chosen parameters. $C_{Det}$ can be defined as:

$$C_{DET} = Norm_{Fact} * \Big( \big( C_{FR} * P_{FR|Target} * P_{Target} \big) + \big( C_{FA} * P_{FA|NonTarget} \big) \Big) \quad \textbf{Eq. (1-3)}$$

  where:
  - $C_{FR}$=10 → Cost of a false rejection.
  - $C_{FA}$=1 → Cost of a false alarm.
  - $P_{Target}$ = 0.01 → Probability of a target
  - $P_{NonTarget}$ = 1 – $P_{Target}$ = 0.99 → Probability of a non-target or impostor
  - $Norm_{Fact}$. → Normalisation factor defined to make 1.0 the score of a knowledge-free system that always decide false.

The system performance is also affected by the inherent differences in the recognisability of different users. Although originally defined in the speaker recognition area [Doddington,1998], this hypothesis can be extended for other biometric systems. Based on the research presented by Doddington, users of a biometric recognition system can be categorised in four different classes:

- Sheep: Represent the users with low intra-class variation and with sufficiently distinctive features, so the recognition system will perform significantly well for them. In other words, the *FAR* and *FRR* will be negligible if only sheep users are supposed to use the system.
- Goats: Represent the users who are particularly difficult to recognise due to its high intra-class variation. This kind of user adversely affects the performance of the recognition system as they tend to increase the false rejection rates.
- Lambs: This class encloses those users particularly easy to imitate. A set of lamb users will show very low inter-class variability between each user, thus making

it extremely difficult to differentiate between them. The rate of false acceptance will be increased with this kind of users.

- Wolves: This group represents the users who are skilful imitators. As in the case of lambs, the performance of the system can be affected by this kind of users as they increment false acceptance errors.

The existence of these different groups of users, which provokes the increase of error rates, shows that biometrics, is not a recognition panacea. Additionally, biometric recognition systems can cause a negative reaction in users due to privacy concerns, cleanliness and security of devices, etc. In this context, [Elliott,2007] present a research on how people perceive biometric technology. However, biometrics-based solutions provide powerful tools for increasing security access systems, especially when combined with classical identification approaches.

### 1.4.1 Biometric characteristics

As we have already established, biometric characteristics can be divided into two different main groups: physiological characteristics (i.e. based on physical traits) and behavioural characteristics (i.e. based on behaviour traits). Regardless this classification, there are some requirements that a biometric characteristic should meet to be used in a biometric application. In 1994, [Clarke,1994], enumerate a list of requirements: universality, uniqueness, permanence, indispensability, collectability, storability, exclusivity, precision, cost, convenience and acceptability. Later on, [Jain,1998], added two additional requirements for practical considerations: performance (i.e. resources required to achieve the desired performance) and circumvention (i.e. how easily the system can be mislaid).

The most commonly implemented or studied biometrics is: fingerprint, face, iris, voice, signature, and hand geometry. However, in what follows, we will introduce not only this biometrics but emerging biometrics as well.

- **Face**

  The aim of a face recognition system is to identify or verify faces present in an image or video scene, analyzing facial features. The main problems that face recognition systems meet are related with lighting conditions of the captured images, face variability (due to facial expression, aging, use of cosmetics, facial hair and hair styling), pose variability. On the other side, it is a non-intrusive method, with no need to touch any device for feature extraction. Although early face recognition systems operate with visible light images, current researches are able to deal with infrared [Kong,2007], [Huang,2007] and 3D images [Zhong,2008]. In the 2006 Face Recognition Vendor Test (URL: FRTV) the verification performance achieved by the participants was quite good, especially under high resolution and controlled illumination conditions [Phillips,2009].

  Over the last years different algorithms have been proposed to address the face recognition problem that can be classified mainly into two categories: geometric feature-based methods, also known as model-based methods or structural methods, and appearance based methods, also known as holistic methods. Obviously, some approaches combine both methods in order to try reaching a better performance.

  - Appearance based methods: These methods use the whole face region as input to the recognition system. PCA (Principal Component Analysis)

commonly referred to as the use of eigenfaces [Turk,1991-A], is a technique applied to most of appearance based methods. The PCA approach is used to reduce the dimension of the data. Face images are projected onto a feature space that encodes the variation among known face images. The face space is defined by the eigenvectors of the initial training set of face images. From the projection coefficients, we can perform a face classification based on a nearest-neighbour approach [Turk,1991-B], Neural Networks, Support Vector Machines or a multi-class one-against-all binary Minimax Probability Machine (MPM) [Hoi,2004]. Another statistical approach that aims to maximise the between-class variance and minimise the within-class variance is LDA (Linear Discriminant Analysis). LDA is used to find a linear combination (Fisherfaces) of features which best separate two or more classes. The resulting combinations are used as a linear classifier. Like in the PCA case, different classification techniques can be applied: Multilayer Feed-forward Neural Network, Euclidean Distance or Normalised Correlation [Nazeer,2007]. The ICA (Independent Component Analysis) statistical approach has also been applied to face recognition systems. There are two architectures of ICA in face recognition, ICA1 which treats the images as random variables and ICA2 in which pixels of images are treated as random variables. [Liu,2004] apply Nearest Neighbour classifier on the ICA2 version. [Lei,2006] apply a modified ICA1 approach, using an image sequence instead of a single image as operation unit.

As stated in [Sharkas,2008], none of the ICA, fisherface or eigenface algorithms significantly outperformed the others. Other approaches have also been proposed to solve the face recognition problem. [Zhao,1998] combined PCA and LDA in a two-step process: PCA is used to obtain a face subspace and then LDA is used to obtain the best linear classifier. The classification is performed in the classification space based on a distance measure criterion. [Xiao,2008] applied a SVM classifier to the raw images of the frontal face area, while the sigmoid function is used to estimate the posterior probability output in binary SVM classifier. [Kim,2004-A] proposed the use of GMM as modelling technique but instead of the raw image, DCT (Discrete Cosine Transform) coefficients were used in face feature vectors.

- Model based methods: These methods employ shape and texture of the face, along with 3D depth information (when available). For instance, areas, distances or angles between eyes, nose or mouth. Probably one of the most popular structural methods is the EBGM (Elastic Bunch Graph Model) which relies on the fact that face images have many non-linear characteristics. In this case a graph is used to represent faces [Wiskott,1997]. The Gabor wavelet transform plays an important role in this type of representation, as a set of complex Gabor wavelet coefficients (also known as jets) are used to designate each node of the elastic grid that defines a face. These nodes are located at facial landmarks. The similarity between graphs is computed as an average over the similarities between pairs of the corresponding jets. Gabor wavelet representations have been also used in [Liu,2003] and

[Wang,2006]. However, [Liu,2003] use PCA to reduce the dimensionality of the data and then use ICA to define the independent Gabor features used in a Bayes linear classifier. [Wang,2006] combine PCA and LDA. [Amira,2005] compare the use of 3 types of wavelet transforms (Gabor, Haar and Biorthogonal 9/7) as a preprocessing step for 14 ratio calculations between automatically extracted landmarks. [Husken,2005] apply a Hierarchical Graph Matching (HGM) technique based on the EBGM approach to model faces using both 2D (texture) and 3D (shape) information.

- Hybrid Methods: Prior to the application of PCA for recognition purposes, [Mittal,2008], for example, use a Fast Fuzzy Edge Detection to determine the nose-width and mouth-with and perform a categorisation of faces in subgroups. Thus reducing the number of faces to test during the test phase.

A comprehensive review of different face recognition algorithms and techniques used to solve different open problems related to image preprocessing can be found in [Zhao,2003]. State-of-the-art recognition rates can be found in [Phillips,2006], [Phillips,2009].

There are a quite big number of databases available for face recognition applications [Phillips,2009], [Gross,2005]. Among them, the most cited are:

- *M2VTS/XM2VTS*: The M2VTS database contains five 286x350 pixels colour images from 37 different individuals. These five images where extracted over a 3 months period in a highly cooperative scenario: good picture quality, indoor shooting, nearly constant lighting and uniform grey background. For each person motion sequence (head rotation from 0º to -90º, back to 0º and then to 90º) and glasses off sequence images where extracted. (URL: M2VTS). The XM2VTS differs from the M2VTS in the number of enrolled individuals, the resolution and in the inclusion of a 3D model of each individual head acquired using a high-precision stereo-based camera. This database contains four 720X576 resolution images taken over a four month period from 295 individuals, in a speaker and rotation head manner. (URL: XM2VTS). Both databases include voice recordings from each person uttering numbers or a specific phrase.
- *FERET*: The FERET database includes a grey-scale and a colour dataset. The grey-scale FERET dataset consists of 14051 greyscale images from 1209 subjects with views ranging from frontal to left and right profiles, and a resolution of 256x384 pixels. The images were collected over the course of 15 sessions and under different illumination conditions and facial expressions. The colour FERET database contains 11338 facial images from 994 subjects with a 512x768 pixels resolution.(URL: COLORFERET)
- *ORL*: The ORL Database of Faces, collected at the AT&T Laboratories Cambridge, contains images from 40 different subjects taken between 1992 and 1994. The ages of the subjects, 4 females and 36 females, ranges from 18 to 81. For each subject 10 images of 92x112 pixels with 256 grey levels per pixel are extracted in different situations: different lighting conditions, different facial expressions and different facial

details (where possible). However, all the images were taken with a dark homogeneous background with the subjects in an upright, frontal position. (URL: ORL)

- *NIST Mugshot Identification Database*: This database contains a total of 3248 segmented 8-bit greyscale images of variable size using lossless compression from 1573 individuals (1495 males and 78 females). The database contains both front and side views of the individuals. More precisely, 1333 cases with both front and side views, 131 cases with two or more front views and 89 cases with two or more profiles.
- *Equinox HID Face Database*: The Equinox database consists of 1622 pairs of visible and calibrated thermal IR faces from 90 individuals. The visible and thermal IR images have a spatial resolution of 320 × 240 pixels, and a greyscale resolution of 8 bits (visible) and 12 bits (IR). The registered images are extracted under different lighting conditions, different pose of the individual and, with and without glasses when available. (URL: EQUINOX)
- *CASIA 3D Face Database*: This database contains 4624 scans of 123 subjects using a non-contact 3D digitiser. For each person in the database a 2D colour image and a 3D facial triangulated surface are extracted, under different conditions of poses, expressions and illuminations. For subjects wearing glasses two different scans are generated: with and without glasses.
- *FRAV3D*: The database contains 2D images (texture information in BMP format) and 3D images (VRML file) from 106 subjects. For each person 16 captures where taken under different conditions covering: gesture variations, lighting changes and pose variations. (URL: FRAV3d).

- **Ear**

  The use of ears as a biometric characteristic for identification purposes, like other biometric characteristics, comes from the forensic field. Furthermore, earmarks (also known as earprints) are still used by police and forensic specialist as a proof of identity [Kasprzak,2003], especially in the absence of valid fingerprints. [Iannarelli,1989] carried out different studies and concluded that the ear contains unique physiological features and what is more important its structure does not change significantly over time. Ear is also suitable for human identification, because unlike face recognition or voice, it is not affected by emotional state or health of people. Moreover, as state in [Choras,2007-A], ear is one of our sensors, therefore it is usually visible to enable good hearing. Other advantage of ear biometric systems is the high acceptability from the sanitary and safety point of view, as far as during both enrolment and test people do not need to touch any device, and the feature acquisition can be performed from an image capture with a simple camera. On the other side, the use of ear entails some problems or disadvantages. Gravity can cause the ear to undergo stretching, especially in the lobe of the ear [Iannarelli,1989], [Burge,2000]; illumination, occlusion and head rotation are classical reported problems during feature extraction process; finally, to obtain a good model of the ear, cooperation of the person is compulsory [Pun,2004].

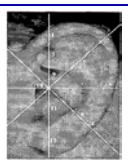
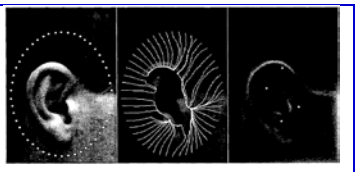
| | | | |
|---|---|---|---|
|  |  |  |  |
| 1 | 2 | 3 | 4 |

**Table 1-1** Different ear representation's: (**1.** [Iannarelli,1989] **2**. [Burge,2000] **3.** [Moreno,1999] **4.** [Hurley,2000])

Although [Jain,2004-B] state that features of an ear are not expected to be very distinctive in establishing the identity of an individual, different studies report a recognition performance higher than 90%.

[Burge,2000] introduce a geometrical method in which the ear is modelled as a Voronoi neighbourhood graph of the curves extracted from the detected edges of the ear. To deal with erroneous curve segments, due to changes in lighting, shadowing and occlusion, while preserving the ear structure, they introduce an *Error Correcting Graph Matching* algorithm. However, as reported in [Pun,2004] and [Yuan,2005], this method is unstable as it is affected by relative small changes in camera-to-ear orientation and lighting, and also, because it will not be able to differentiate real edges and non-edge curves [Sana,2007]. Additionally, [Burge,2000] propose the use of the width of the curve corresponding to the upper Helix rim to improve the false acceptance rate, and the use of thermogram images to deal with partial occlusion problems. [Yuan,2005] and [Mu,2005] also propose a geometrical approach to ear modelling based on shape and structural features (LABSSFE – long axis based shape and structural feature extraction), reporting an 85% of recognition accuracy. The representation used is invariant to ear's image parallel move, scale and rotation. In [Choras,2007-B] and [Choras,2006], five different methods are presented also based in geometrical parameters of ear contours (i.e. distinctive information about shape and geometrical properties): Concentric Circles based Method (CCM), Contour Tracing Method (CTM), Angle-Based contour representation Method (ABM), Geometrical Parameters Method – Triangle Ratio Method (GPM-TRM) and Geometrical Parameters Method – Shape Ratio Method (GPM-SRM). Experiments show a FRR lower than 10% in all cases, and 0% FRR is reported for GPM method but, no FAR is reported.

An important aspect common to all these methods is the need of a preprocessing stage of the image, which entails: image normalisation and ear detection. On the other hand [Hurley,2000], proposed a method that do not need any preprocessing or ear extraction. A Force Field Transformation is carried out, in which the image is treated as an array of mutually attracting particles that act as the source of a Gaussian force field. Then a Force Field Feature Extraction is performed in which the ear is described as a set of energy lines, wells and channels. The technique is highly invariant to initialisation, translation and scaling, and has an excellent noise tolerance, providing a recognition rate higher than 95% [Hurley,2005].

Other approaches avoid feature extraction, and propound a holistic approach using appropriate tools. [Moreno,1999] proposed the use of neural networks in two ways: to extract the macro features of the image (Compression Network), and to perform identification tasks (perceptron), reporting identification results

of 93%, without considering rejection thresholds. [Sana,2007] developed a new approach base on discrete Haar Wavelet transform. After a preprocessing to detect ear on the image, wavelet coefficients are computed from Haar Wavelet decomposed image, to create a template. Using Hamming distance approach for decision making, an accuracy of about 96% is reported. [Victor,2002] used a partially automated algorithm based on Principal Component Analysis (PCA), both for face and ear recognition.

Last but not least, 3D representation of the ear is been under extensively research in the last years. [Cadavid,2008] create a 3D ear model from a single video frame using Shape from Shading (SFS) technique. Subsequently, the Iterative Closes Point (ICP) algorithm is used for model alignment and RMSD values is used for model recognition, producing recognition rates of 95%. The main problems with this method are sensitiveness to pose and light variations. [Chen,2007], also used the ICP algorithm for model alignment, but in this case, the 3D ear model is obtained fusing range images and colour images, producing two different representations: the ear helix/antihelix, and the Local Surface patch (LSP). The root mean square (RMS) registration error is used as the matching criterion, obtaining recognition rates around 95%. [Yan,2007] and [Yan,2006] use an improved-ICP for model alignment and an Active Contour algorithm for edge based ear model extracted from 2D+3D information. They achieved recognition results of 97.6%, even in the case of partially occluded ears or ears with earrings.

- **Iris**

The iris is a pigmented membrane that controls how much light reaches the sensory tissue of the retina, through the pupil, thanks to the sphincter and the dilator muscle. The amount of pigment contained in the iris determines eye's colour, and its visual texture formed during prenatal growth and developed during the first years of life, carries unique and distinctive patterns useful for personal recognition. Iris recognition is used in high security areas due to its accuracy, reliability and speed of recognition. Moreover, its accuracy is not especially affected by pupil dilation (pupil size can vary from 10% to 80% of the iris diameter) or the use of eyeglasses or contact lenses [Weaver,2006]. However, it requires users collaboration because partially occlusion by eyelids or eyelashes, or completely occlusion of iris due to blink will affect adversely recognition results. Another problem that affects iris recognition systems is the specular reflection within the iris region [Padma Polash,2007].

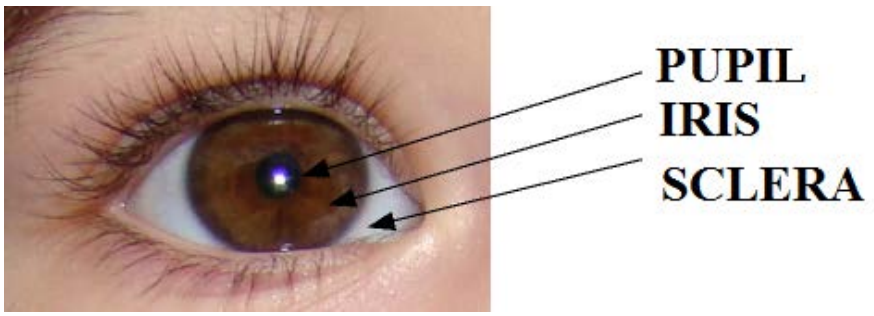



**Figure 1-6** Right eye description

The iris recognition system can be broken up into five main stages:

- *Iris localisation*: This stage consists in processing of eye image to detect the iris region (inner and outer boundaries of iris). It is important to remove eyelids or eyelashes from eye image.

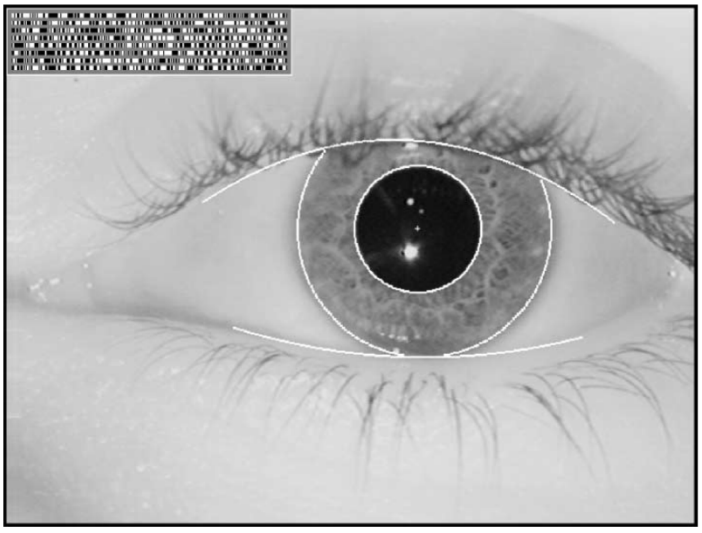


**Figure 1-7** Example of iris localisation (White circle lines) [Daugman,2004]

Different algorithms have been proposed to solve this problem. The first to be used in a commercial system is the one by [Daugman,1993], [Daugman,2004], which implements a coarse-to-fine strategy to locate outer and inner limits of iris in an image using an *integro-differential operator*. A similar approach, changing the integration contour from circular to arc, is used to locate eyelid boundaries. This approach is also used in [Ko,2006]. [Wildes,1997], [Zhang,2004] and [Padma Polash,2007] use the circular and parabolic *Hough Transform* to perform iris detection after a binary edge-map is built via gradient-based edge detection. [Grabowski,2006] use the *black hole method* to estimate the pupil area and then the Daugman's method to estimate the outer limit of iris. Other approaches fuse different mentioned-above approaches, like [Roy,2007], in which a first step for pupil detection is proposed based on *Canny edge detection* and *Sobel edge detection*. The second step consists in the application of Daugman's method. However, the first step is rejected due to high computational cost. Finally, [Miyazawa,2005] model the inner boundary of the iris as an ellipse for which a sudden change in luminance summed around its perimeter takes place. The outer limit is modelled in a similar way but using a circle path of contour instead of an ellipse.

In order to reduce the effect of eyelashes in iris recognition, different approaches have been implemented. The simpler approach consists in removing the areas in which eyelashes may occur from the iris template. [Miyazawa,2005] use only the lower arc of the iris to create an eyelash-free iris template before normalisation. [Ma,2003] remove the lower part of the iris pattern after normalisation, getting a similar result. [Roy,2007] use a pixel threshold to determine whether a pixel corresponds to an eyelash or not. More complex eyelash detection algorithms are present in

[Huang,2002] and [Kong,2001] based on line detection or Gabor filter, and variance of intensity over a small window in the iris image

- *Iris normalisation*: This step is done to compensate for iris deformation, converting iris image from Cartesian *(x,y)* coordinates to Polar coordinates *(r,θ)*, where $r \in [0,1]$ and $\theta \in [0,2\pi]$. At the end of this stage we get an unwrapped representation of iris in a rectangle image with angular and radial resolution. The most use method is the rubber sheet model described in [Daugman,1993].

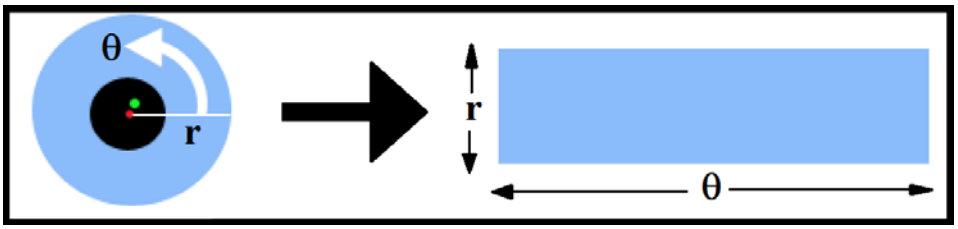



**Figure 1-8** Iris normalisation process (extracted from[Roy,2007])

However, some variations have been proposed in which only some ROI (regions of interest) areas are used in this processing stage [Szewczyk,2007], [Miyazawa,2006]. Daugman's model do not account for rotational inconsistencies, but in the matching phase.

- *Image enhancement*: Image enhancement algorithms are used to deal with specular reflections and/or non-uniform illumination in iris image.

- *Feature extraction*: This stage consists in creating an iris template which includes the most relevant and discriminating features from the normalised image, for identification purposes. As is the case of iris normalisation, most of the experiments in iris recognition use the feature extraction method presented in [Daugman,1993]. A 256-byte Iris Code is created using a 2-D Gabor filters and a coarse phase quantisation of the local texture signal by approximating it as one vertex of the logical unit square. [Kong,2001], [Liu,2005], [Roy,2007] also used Gabor filters, while [Padma Polash,2007] or [Zhang,2004] use 1D Log-Gabor filters for feature extraction. An alternative method evaluated for this stage is wavelet transformation that decomposes the iris image into components with different resolutions. Haar, Daubechies, Biorthogonal, Coiflet, Symlet and Meyer have been tested in different studies [Szewczyk,2007], [Elsherief,2006], [Thornton,2007]. [Miyazawa,2005] use phase-based components in 2D DFTs (Discrete Fourier Transform) of iris image. [Chu,2005] create the iris feature vector using LPCC (Linear Prediction Cepstral Coefficients) and LDA (Linear Discriminant Analysis) for dimensionality reduction. [Huang,2002] adopt ICA to extract it is texture features. [Ko,2006] generate iris codes by analyzing the changes of grey values of iris patterns using cumulative sums over each group of basic cell regions in which an iris image is decomposed. Finally, [Wildes,1997] use an isotropic band pass decomposition derived from application of Laplacian of Gaussian filters to the image.

- *Template matching*: A target iris template is compared with all the patterns in the database using a matching metric. Finally, a decision of identification or verification is done. Hamming Distance is used as a test

of statistical independence between two iris representations, so two templates are supposed to be from the same iris if the test fails. This method proposed by [Daugman,1993] is also used in [Liu,2005], [Ko,2006] or [Padma Polash,2007]. Other matching methods applied to iris recognition are: Band-Limited Phase-Only Correlation [Miyazawa,2005], Euclidean distance [Huang,2002], average cosine distance [Thornton,2007], Learning Vector Quantisation (LVQ) [Elsherief,2006], Probabilistic Neural Networks (PNN) [Elsherief,2006], [Chu,2005] or Support Vector Machine (SVM ) [Gu,2005].

The first iris recognition patented system, [Daugman,1993], is able to perfectly identify an individual, given millions of possibilities. Moreover, most of the implemented systems report recognition rates over 90%.

The most popular databases for recent studies in iris recognition are:

- ICE database developed in the Iris Challenge Evaluation project [Phillips,2007], which includes 132 different individuals, and more or less 20 iris images from both left and right eyes.
- UBIRIS database [SOCIALab,2014] which in its 2.0 version includes more 11000 images with realistic noise factors.
- CASIA database [CASIA,2014] includes a total of 22051 iris images from more than 700 subjects and 1500 eyes. It also includes the first publicly available twins' iris image dataset from 200 subjects.
- University of Bath database [Monro,2008], which includes in its commercially available dataset 20 images from each eye of 1600 subjects.
- Palacký University database [Dobeš,2014], with 348 iris images captured from 64 subjects.

- **Retina**

As stated in one of the earliest works on this biometric trait [Hill,1998], retinal recognition is a misnomer, because these systems actually deal with the blood vessel patterns of the choroid, located behind the retina. However for the sake of simplicity, the term retina recognition has been coined. The main advantage of this biometric trait is that it is not exposed to the external environment, thus it is almost impossible to change or replicate the retinal vasculature pattern in an impostor. However, it suffers from some disadvantages. Up to some time ago, retina image acquisition equipment was quite expensive, thus limiting somehow the number of people working in the area as well as the databases available for research purposes. Additionally, sample acquisition process involves cooperation and a conscious effort on the part of the subject, as retina is small, internal and therefore difficult to capture [Delac,2004], [Jain,2004-B]. Last but not least, blood vessel patterns can be affected by some diseases, thus revealing some medical conditions of individuals that result in problems of user acceptability [Bowyer,2004], [Deriche,2008].

In retinal recognition systems, a preprocessing step to enhance vessel pattern is carried out. In most cases, this preprocessing step involves the location of the optic disc, which is the region on the retina at which optic nerve axons enter and leave the eye, and which can work as a landmark for other features in retinal

image. For optic disc location, different approaches have been implemented. For instance, [Tabatabaee,2006] used the Haar wavelet and active contours or snake models, [Rahman,2008] introduced a method based on intensity variance of adjacent pixels, and [Ortega,2006] implemented a fuzzy circular Hough transform to localise the optical disk.

Different features have also been tested, with none preferred among the others. [Tabatabaee,2006] used a feature vector which consists of central Fourier-Mellin Transform harmonics length and complex moment magnitudes up to order (2 2). In this approach the centre of the optic disc, in conjunction with the image centre of mass is used for rotation compensation that can occur during scanning process. [Rahman,2008] built a 32 component feature vector, which consists of 12 colour moments extracted from each of the 3 colour space (HSI – Hue, Saturation and Intensity) in which the retina image is decomposed, and 20 texture features extracted from the Grey-level Co-occurrence matrices. In this approach the optic disc is used to extract a region of interest, from which the feature vectors are extracted. [Ortega,2006] used an approach similar to some works on the area of fingerprint recognition, which consist in regarding the retina scanned image as a set of ridges (vessels) and valleys, thus using the whole retinal vessel tree. As in [Rahman,2008], they used the optic disc to define a region of interest (ROI) from which they extract a set of features based on the ridges endings and bifurcations inside the ROI. [Mariño,2006] used the whole retinal vessel tree to detect crests and valleys through the use of a geometry-based method known as MLSEC (multilocal level set extrinsic curvature). [Xu,2005] also extracted the blood vessel skeleton, but in this research, the vector curve of blood vessel skeleton is obtained and a feature vector is defined which includes feature points, directions and zoom factor.

Classification methods applied to retinal recognition found in the literature are the classical Euclidean Distance proposed by [Rahman,2008], a fuzzy C-means clustering proposed by [Tabatabaee,2006], and image registration [Ortega,2006], [Mariño,2006], [Xu,2005]. In these last three cases, the classification method consists in finding a transformation from the target retina-image into an objective retina-image. Once a transformation is performed, a measure of similarity is applied to determine whether the target image and the objective image belong to the same individual.

In [Farzin,2008] a retina recognition system similar to the ones related for iris recognition and different to the previous ones is presented. It is similar in the sense that a ring of the vessel pattern is extracted and transformed to polar coordinates to generate rotation invariant features. This region of interest, centred at the optical disc, is extracted after a binarisation process is performed to enhance the vessel pattern. After polar transformation, a multi-scaled analysis is conducted using a discrete stationary biorthogonal wavelet transform, creating 3 different feature vectors according to the diameter size of vessels. Finally, a similarity index is computed using a scale-weighted summation of a correlation function evaluated over each scale of the feature vector.

Performance results on classification are quite promising as in most cases they reach more than 95% recognition rates. However, results are not comparable due to the reduce number of users used in the experiments and the lack of a reference database.

Some databases used in retinal identification research, primarily collected for medical purposes, are:

- DRIVE: The 40 images included in the DRIVE database were collected from a diabetic retinopathy screening program in the Netherlands, thus containing both signs of mild early diabetic retinopathy and no sign of retinopathy. The retina images were captured using 8 bits per colour and with a resolution of 768x584 pixels. Moreover, a single manual segmentation of the vasculature is also available. (URL: DRIVE)
- STARE: One additional output of the STARE project, aimed at designing a system that automatically diagnoses diseases of the human eye, is the STARE database, that can be accessed on-line (URL: STARE).
- VARIA: This Spanish database (URL: VARIA) contains a set of 233 retinal images from 139 different individuals used for authentication purposes. The images have a resolution of 768x584 pixels. The images are mostly provided by the *Complejo Hospitalario de la Universidad de Santiago* (CHUS) and consist of optic disc centred images and macula centred ones.
- MESSIDOR: This French dataset contains 1200 colour eye images, captured using 8 bits per colour plane at different resolutions (1440x960, 2240x1488 or 2304x1536 pixels). As this database was collected for computer assisted disease diagnosis, it contains images with pupil dilation and without dilation. From each image an expert diagnosis is provided. (URL: MESSIDOR).

- **Voice**

  Voice is a biometric characteristic influenced both by physical and behavioural characteristics of an individual. The physical component refers to the shape and size of the vocal tract, the vocal folds and in general all the appendages (lungs, trachea, vocal folds, tongue, oral/nasal cavities, lips, teeth) that are involved in the voice production process. These characteristics are almost invariant for an individual, however, the behavioural characteristics of speech changes over time due mainly to age, health conditions, emotional state, etc.

  Besides these intra-speaker variations and other problems such as speaker mimicking or replication of previous recorded voice to circumvent a voice recognition system, this biometric characteristics is becoming popular for biometric recognition thanks to: people acceptability as non-intrusive methods are used to collect voice information, availability of relative low-price high quality devices for collecting voice samples (computer microphones, telephone network, etc.), and last but not least, is the only biometric trait that can be used for identification over telephone. A good introduction to speaker recognition systems can be found in [Campbell,1997], [Nickel,2006].

  Although a deeper review will be made later, a generic speaker recognition system can be divided into the blocks shown on Figure 1-9.

  The signal processing block allows the system to convert the voice into a digital signal that can be processed. This processed signal acquired is extremely influenced by the microphone quality, the communication channel and the

digitiser characteristics. The next step to be performed is the extraction of specific features from this speech signal that provides high inter-speaker variability and low intra-speaker variability. The most popular features extracted from the speech signal to be used in speaker recognition systems are Mel Frequency Cepstral Coefficients (MFCC) or Linear Prediction Coefficients (LPC), both of them conveying low-level spectral information. However, there are other low-level and high-level feature characteristics that have been used in this kind of systems, such as: formant estimation, pitch estimation, periodicity/prosodic measures, semantics, idiolects, pronunciations, etc. These features can be combined to improve the performance and make the speaker recognition system more robust.

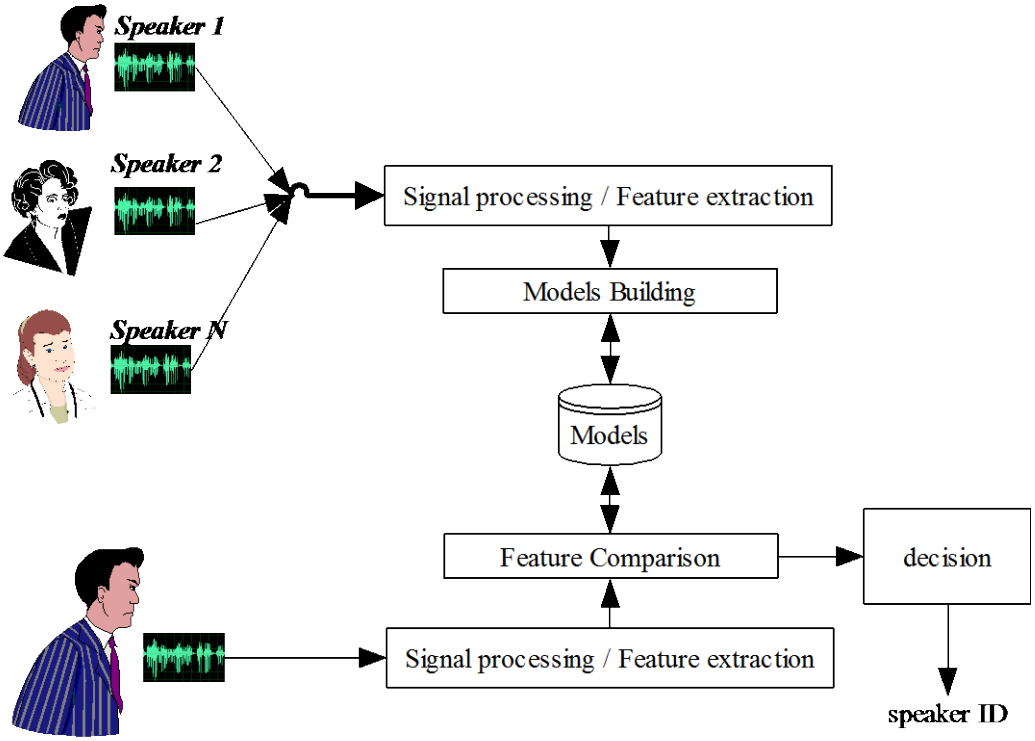



**Figure 1-9** General voice identification system

For the model building block, different approaches have been tested. Among them, the most popular are HMMs (Hidden Markov Models) in the case of text-dependent systems and GMMN (Gaussian Mixture Models) for text-independent systems. Both methods provide a statistical representation of the speech, however, HMMs represent underlying variations and temporal changes over time found in the speech, while GMMs represents the different sound classes uttered by a speaker.

Finally a decision over a target input speech to the system has to be made. For this purpose, the features extracted from the input utterance are compared with existing models of speakers to measure the similarity between them. If there is enough confidence that the utterance belongs to a known speaker previously modelled then, the system will recognise the speaker as an authorised user, otherwise the system will treat the input utterance as belonging to a no-authorised user.

In the state-of-the-art of speaker recognition systems performance is tested in the Speaker Recognition Evaluation (SRE) carried out by NIST (National Institute

of Standards and Technology). The evaluation results can be found at the NIST SRE web page: (URL: NIST-SRE)

Like in other biometric characteristics, there are quite a lot number of databases for voice authentication research. Most of these databases are available through the Linguistic Database Consortium (URL: LDC) which collects and distributes speech databases for research and development purposes and the European Language Resources Association (URL: ELRA). We can divide the available databases into different subsets regarding the origin of the recordings, and also regarding the utterances performed by the speakers:

- Telephone recordings:
  - Text-constrained speech:
    - The ***Idiologos/Bootstrap*** French database, developed under the Neologos project, contains French fixed telephone recordings of 1000 persons (470 males and 530 females) balance across age and regional (12 distinct French regions were used) characteristics. The speakers uttered different information: digit sequences, a spelling of a directory assistance city name and 45 phonetically rich sentences. [Pinto, 2004] (URL: IDIOLOGOS)
  - Conversational/spontaneous speech:
    - ***Switchboard*** is a collection of about 2400 American English telephone conversations among 543 speakers (almost gender balanced) talking about 70 different topics. Speakers' age range from 20 to 60 and the educational level was: 14 persons for less than high school, 39 for less than college, 309 for college, 176 for more than college, and 4 for unknown. (URL: switchboard)
  - Both:
    - ***POLYCOST*** [Hennebert,2000] is an English-spoken-by-foreigners database. 17 different mother-tongue languages were present among the 114 speakers (74 males and 60 females). The recordings were performed in an average of 9 sessions per speaker and there is almost no phone-set intra-speaker variability, as about 80% of subjects called from the same phone in all sessions. The recordings include: 10 prompts with connected digits uttered in English, 2 prompts with sentences uttered in English and 2 prompts in mother tongue (One dedicated to free speech).

- Desktop/microphone recordings:
  - Text-constrained speech:
    - ***HESPERIA*** is a Spanish database recorded for the HESPERIA (Homeland sEcurity: tecnologíaS Para la sEguridad integRal en espacIos públicos e infrAestructuras) project (URL: proyecto-hesperia.org) by the GIAPSI (Grupo de Investigación en Informática Aplicada al Procesado de Señal e Imagen) research group. At the moment, this database recorded for speaker recognition purposes in high security applications consist of 202 speakers (109 male and 93 female), recorded in one session, under controlled conditions with different microphones uttering different pre-established speech (five Spanish vowels, name of speakers, a phonetically balance phrase, 1 minute reading of a text and different combinations of 4-digit pins).
    - ***YOHO*** database contains a large scale, high-quality speech corpus from 138 subjects (106 males and 32 females) recorded in 4 different sessions. The recordings comprise the utterance of a sequence of three two-digit numbers. (URL: yoho)
  - Conversational/spontaneous speech:
    - ***Mixer Corpora***: The databases generated under the Mixer Project (URL: MIXER), include speech from a large pool of speakers. In the case of Mixer 5 corpora, 300 speakers were recorded on 6 different sessions using 14 different microphones and in different communicative situations. The datasets recorded under the Mixer Project have been used in the NIST Speaker Recognition Evaluations (URL: NIST-SRE) and in the Language Recognition Evaluations (URL: LRE). A description of some of the recorded corpora can be found in [Cieri,2007] which includes also other corpora with telephone call recordings.
    - ***NIST Meeting Pilot Corpus Speech*** [Garofolo,2004] consists of recordings of meetings in which an average of 6 people take part. There are 61 unique speakers (41 males and 20 females) in the 19 meetings which contain a mix of real meetings and scenario-driven meetings.
  - Both:
    - ***Ahumada*** is a set of Spanish speech databases recorded under different conditions. In Ahumada I, up to 100 male speakers were recorded in 6 different sessions using different microphones and uttering both spontaneous speech and specific text. Ahumada II, contains the same speakers but recorded through the Spanish GSM network and 10 years later [Ortega-Garcia,2000]. Ahumada III

Release 1 consists of 61 male speakers extracted from authorised GSM conversational speech from real forensic cases [Ramos,2008].

- Broadcast resources: Although not specifically collected for speaker authentication purposes, these databases can be useful for speaker segmentation research.
  - In the ***2002 Rich Transcription Broadcast News and Conversational Telephone Speech*** there is a subset of broadcast news data. This data is composed of six approximately 10-minute excerpts from six different broadcasts. The broadcasts were selected from programs from MNB, PRI, NBC, CNN, VOA and ABC, all collected in 1998. (URL: RTBNCTS)
  - The ***ESTER*** corpus includes near 2000 hours of French broadcast news and among them, 100 hours with information about speaker turns, identities, genre of the speaker, accent if any and if he/she is French native speaker. The acoustic resources come from: France Inter, France Info, Radio France International (RFI), Radio Télévision Marocaine (RTM), France Culture and Radio Classique. The records contain a total of 2172 speakers (744 female, 1398 males and about 20 children) with about 20% of them being non-native speakers [Galliano,2006].

A special mention is to be made in relation with databases for speaker recognition. TIMIT (compiled in the early 1980's) was one of the early corpora widely used for research in this area. TIMIT is a corpus of read speech, containing 10 phonetically diverse sentences spoken by each of 630 speakers chosen to represent 8 major dialect regions of the United States. The records were made using high quality microphones, lower quality microphones and different types of telephone channels.

- **Fingerprint & Palmprint**

A fingerprint is the pattern of ridges and grooves on the surface of a fingertip. Fingerprints have been used for identification purposes for many centuries providing high matching accuracy. So far, no known pair of fingerprints has the same pattern, even between identical twins or between each fingertip of the same person.

The fingerprint acquisition process has changed over the years. The oldest method consists in the use of ink and paper to obtain a fingerprint. However, in the last years, live-scan acquisition systems have been developed making the acquisition process faster and cleaner, although the fingerprint area captured is usually smaller. These live-scan systems include optical methods, thermal imaging, electromagnetic field imaging or ultrasound imaging. A complete review of the different acquisition technologies can be found in [Adhami,2001]. Regardless the acquisition's technology, there are some factors that prevent them from registering a good pattern, for example, skin disease, scars, sweat, dirt, etc.

Previously to the matching phase for identification purposes, a set of features should be extracted from the fingerprint pattern. The fingerprint pattern can be analysed at different levels providing different types of features. A coarse analysis may provide a set of features that can be seen with the naked eye. The

flow of the ridge lines delineates a set of singular points that can be classified as: loop, whorl, arch and tented arch. These kinds of features are suitable for classification and indexing but do not provide enough distinctiveness for accurate recognition. External fingerprint shape, image orientation and image frequency are also coarse level features [Maltoni,2009]. A fine analysis of the fingerprint pattern will provide a set of characteristics called minutiae that are unique to the individual. Typically minutiae features are ridge bifurcations (points where a ridge splits to form a Y shape) and ridge terminations. However, a more complete set of minutiae points can be defined, including: lakes or enclosures, short or independent ridges, dots, spurs, and crossovers or bridges [Rahal,2006]. Minutiae information can be characterised based on orientation, spatial frequency, curvature and position [Weaver,2006]. At a very fine level, intra-ridge details, such as sweat points, can be detected. However, a high-quality fingerprint acquisition is required making it unsuitable for most applications. Based on the idea of IrisCode described by Daugman, [Jain,1999-A] introduce the idea of FingerCode for classification and matching. In this approach, the FingerCode feature vector is obtained after applying a Gabor filter bank over a ROI of the finger extracted after applying a Poincaré-based algorithm (a method to locate singular points in a flow field).

According to [Maltoni,2009], fingerprint classification methods can be distributed into one of the following categories:

- Correlation-based matching, in which correlation between input and target fingerprint images (at the intensity level) are computed for different alignments. Correlation matching can be performed both in the spatial and in the frequency domain.
- Minutiae-based matching involves comparing the two-dimensional minutiae patterns extracted from the user's print with those in the template. i.e., finding an alignment between input and target fingerprint so that the number of coincident minutiae features are maximised.
- Ridge feature-based matching: On these methods, fingerprints are compared in terms of local orientation and frequency of the ridge pattern, ridge shape or texture information.

Palm recognition have become an area of interest mainly due to the increasing interest of law enforcement agencies, as at least 30% of prints lifted from crime scenes are of palms. Palm print recognition inherently implements many of the same matching characteristics of finger print recognition systems, as in both cases features are extracted from the information present in the ridge impression. Palms of the human hands contain a pattern of ridges and valleys like finger prints. However, palmprint scanners need to capture a large area, but on the other hand palmprints are expected to hold more distinctive features.

Like in fingerprint feature extraction, a multi-resolution analysis can be performed. At low-resolution processing, principal lines, wrinkles and creases [Chen,2001] can be extracted. At high-resolution minutiae features from ridges existing in the palm and singular points can be characterised. Additionally, texture features of the skin can be used in palmprint recognition systems [Choras,2007-B]. Texture analysis can be performed using FFT, DCT, Gabor Filters [Shi,2007] or Wavelet transform [Butt,2008]. All features can be combined in order to build a more accurate biometric system.

Matching algorithms can be classified in the same way as in the case of palmprint recognition systems. Some works use Euclidean [Butt,2008] or Hamming distance [Shi,2007] classifiers, Neural Networks [Shang,2006] and Support Vector Machines [Zhou,2006] have also been used in palmprint recognition systems.

The major drawbacks of both fingerprint and palmprint biometric recognition, are: a small fraction of the population may be unsuitable for automatic identification due to genetic factors, aging, environmental or occupational reasons. Although it is probably the most accepted biometric it is also associated with forensic recognition, especially for crime solving and police records. Additionally, fingerprints are not very private. We leave fingerprints everywhere and every time we touch something, thus the probability of a fingerprint being stolen is very high, and what is more problematic, once a fingerprint is stolen, it is stolen for live. Some work has been done on the problematic of using fingerprints for recognition purposes [Matsumoto,2002], [Uludag,2004], [Palka,2007].

The most known fingerprint database is the one owned by the FBI and used in the IAFIS (Integrated Automated Fingerprint Identification System). The IAFIS maintains the largest biometric database in the world and contains information (fingerprints and criminal history information) for more than 55 million subjects. (URL: Iafis). However, it is not publicly available. Fortunately, there are other fingerprint and palmprint databases available for research activities such as:

- FVC databases: The FVC (Fingerprint Verification Competition) has been evaluating the advances in fingerprint recognition technologies every two years since 2000. For each of the competitions, they have provided different fingerprint databases. For instance, in the FVC2006 (URL: fvc2006), four disjoint databases were collected with different sensor/technology where each database contains 1800 fingerprint images (150 fingers x 12 samples/finger) in 256 grey-level BMP format, with different size and resolution.

| | *Sensor Type* | *Image Size* | *Resolution* |
|---|---|---|---|
| ***DB1*** | Electric Field Sensor | 96x96 (9Kpixels) | 250 dpi |
| ***DB2*** | Optical Sensor | 400x560 (224Kpixels) | 569 dpi |
| ***DB3*** | Thermal Sweeping Sensor | 400x500 (200Kpixels) | 500 dpi |
| ***DB4*** | SFinge v3.0 | 288x384 (108Kpixels) | About 500 dpi |

**Table 1-2** FVC2006 Datasets

- SFinGe: SFinGe stands for Synthetic Fingerprint Generator, and is a method developed at BioLab (URL: BIOLAB) for the generation of synthetic fingerprint images. (URL: Biolab-Fingerprint)
- NIST Databases: The National Institute of Standards and Technology (NIST), has collected different databases for fingerprint recognition research (URL: Biomet), for example:
  - SD4: NIST Special Database 4 contains 2000 8-bit greyscale images of fingerprint image groups. Each image is 512x512 pixels and is classify according to its pattern: arch, tented arch, left loop, right loop or whorl.

        - SD9: NIST Special Database 9 contains 27000 8-bit greyscale images of mated fingerprint card pairs. Each image is 832x768 pixels in size.
        - SD27: NIST Special Database 27 contains fingerprint Minutiae from Latent and Matching Tenprint Images. Each image is 832x768 pixels in size, quantised to 256 levels of grey.
    - CASIA Palmprint Image Database V1.0: Contains 5239 palmprint images captured from 301 different subjects. All palmprint images are 8 bit grey-level JPEG files collected in only one session. (URL: CASIA)
    - NAFIS (Australian National Automated Fingerprint Identification system): The NAFIS houses one of the largest repositories of palm prints in the world with 4.8 million palm prints.

- **Hand &Finger geometry**

  Hand geometry refers to the use of dimensions of the hand for identification purposes. The information extracted can include, among others, size of palm, lengths and widths of fingers, shape, angles between landmarks, and ratios of those quantities. Hand geometry does not provide very distinctive characteristics (for example, hand size changes over time), however it is widely used for security application all over the world due to the user's acceptability if compared with other biometric characteristics such as iris or fingerprint recognition systems.

  A recognition system based on hand geometry presents the same basic steps as other systems based on different biometrics: sample acquisition, biometric template building, and classification of a target sample into a known class (template). In a typical system [Jain,1999-B], hand image is acquired using a camera or a scanner. Additionally, a mirror can be used to project the side-view of the hand, resulting in hand thickness information acquisition. In order to standardise the acquisition process, pegs can be used in the acquisition surface to fix the position of hand and finger (see Figure 1-10). However, there is a trend to use a pegless acquisition process as long as all the fingers are placed apart during the acquisition process.

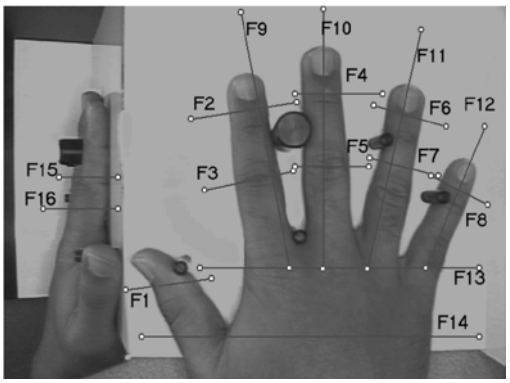



**Figure 1-10** Hand acquisition surface with pegs (extracted from [Jain,1999-B])

Usually silhouette is extracted from the hand image, thus high resolution images are not necessary. Moreover, in [Fouquier,2007] different resolutions have been tested yielding better results with lower resolution images. After image binarisation is performed, geometric features are extracted.

[Jain,1999-B] used a peg-image obtaining 16 features from the hand. These features include the width of the fingers at different positions, length of fingers, thickness of the hand and width of the palm. In [Sanchez-Reillo,2000-B] 31 features have also been extracted from a peg-image, including 21 widths of fingers and palm, heights of the little finger, middle finger and palm, four finger deviations and 3 angles of the interfinger valleys with the horizontal. A feature selection based on the f-ratio process has been performed, resulting in a 25 feature vector. [Covavisaruch,2005] built a 21 feature vector made up of the lengths of each finger, the widths of each finger at 3 different locations and the width of the palm. Contrary to previous works, [Faundez-Zanuy,2005] used a hand image acquired without the use of pegs or templates during the image acquisition process. In this work, some landmarks are previously detected: finger tips, valleys between the fingers and 3 additional points of the palm. Using these landmarks 8 points are extracted to build the feature vector representing the hand: length of the 5 fingers and 3 measures related to the palm landmarks previously extracted. [Kumar,2003] also built a 1x16 feature vector, which includes 4 finger lengths, 2 widths per finger, the palm width and length, hand area, and hand length. Other approaches which do not involve the use of detailed geometric measures extracted from the hand, but the hand silhouette itself after binarisation, is for example [Yöruk,2006], which compare two hand silhouette shapes in terms of the Hausdorff distance.

In the classification or verification step, the use of different distance metrics are the most common approach [Jain,1999-B], [Covavisaruch,2005], [Fouquier,2007], although other pattern recognition methods have been tested. For instance, [Sanchez-Reillo,2000-B], have tested the Euclidean Distance, Hamming Distance, GMM or RBF-NN (Radial basis function neural networks) as classification methods. Among all these methods, GMM yields the best experimental results with a 97% of success. General Regression Neural Networks (GRNN) [Polat,2008], and Multi-Layer Perceptron (MLP) [Faundez-Zanuy,2007] are also tested in hand biometric recognition systems.

Although there is an increasing number of hand biometric commercial products, and the publications in this topic keep increasing, the number of public available databases is negligible if compared to other biometric characteristics. Most of the researches in this area are carried out using ad hoc databases, due to the low cost image acquisition systems. Moreover, we can divide the databases into two subsets depending on the kind of image acquired: fix-placement hand image using pegs to determine the right position of the hand [Sanchez-Reillo,2000-A], [Jain,1999-B] or free-placement hand image [Yöruk,2006], [Boreki,2005], [Polat,2008]. Despite this tendency to build ad hoc databases, we can find some public available databases for research purposes:

- GPDS150hand database: This database contains 1500 right-hand images from 150 different individuals, resulting in 10 different images per person. The images were acquired over a scanning surface without peg limitations, and were stored in jpg format, 256 grey levels and 120 dpi of resolution [Ferrer,2007-B]. (URL: GPDS).
- Biosecure database: Biosecure is a multimodal database which includes, apart from hand information, voice, face, signature, fingerprint and iris data. Hand information was acquired in two different sessions from 750 individuals. For a subset of 642 individuals ambidextrous recordings are

available. For another subset, different resolution images were taken, resulting in a database of 4700 hand images.

- **Vascular Patterns**

Vascular pattern recognition, also referred as vein-pattern recognition consists in the use of blood vessel structure of hand or fingers for individual identification. Different researches have established that the vascular pattern of the human body is unique and does not change over time. Although this biometric trait shares some common points with retina identification, as it uses vein patterns of the human body, it presents additional advantages over it: higher user acceptability and hand/finger vein sensors are near contact-less and are more easy to use than retinal scans. Moreover, it also presents additional advantages over fingerprint or hand geometry biometrics. First of all, the sensors look for information below the skin; therefore it is not affected by finger/hand moisture, cuts or dirt. Secondly, vein patterns are difficult to forge because they are inside the hand and blood needs to flow to register the image.

Table 1-3 shows some hand (a) and finger (b) vein IR images used for vascular pattern recognition [Shahin,2008-A], [Miura,2004], and FIR (far-infrared) image (c).

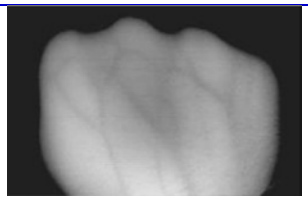
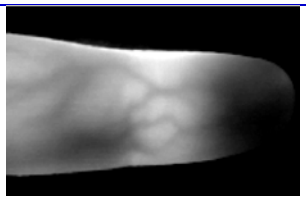
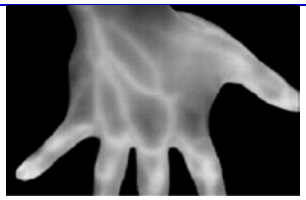
| | | |
|---|---|---|
|  |  |  |
| (a) | (b) | (c) |

**Table 1-3** Vascular patterns used for individual recognition

Although different vein pattern recognition systems have come to market, there are few published research studies addressing the issue of feature selection in vein recognition systems. Almost all the researches on this biometric agree on the procedure for personal identification using vein patterns: (1) Image Acquisition, (2) Image Normalisation, (3) Pattern Extraction, (4) Matching.

Different approaches have been tested for the first step. For instance, [Wang,2008] capture the vein pattern in the back of the hand, but imposing some restrictions in the position of the hand. [Ding,2005] and [Shahin,2008-A] use the palm-dorsa vein patterns but captured with the fist clenched. [Mulyono,2008], [Miura,2004] and, [Hashimoto,2006] restrict the problem of vein recognition to finger area.

After vein structure is captured a normalisation step is accomplished to extract an emphasised vascular pattern. In most cases, this normalisation step involves the selection of a region of interest (ROI) from which features are extracted later. Furthermore, local thresholding [Shahin,2008-A], [Wang,2008], [Im,2001] or adaptive thresholding [Mulyono,2008], are used to highlight the vein structure, and a median filter is also used for noise reduction [Im,2001]. Additionally, some authors apply an extra step to accurately determine the vein structure. For example, [Wang,2008] and [Ding,2005] proposed the use of a thinning algorithm to reduce the size of veins; whereas [Miura,2004] proposed a line tracking algorithm for vein detection.

Most researches used the whole image after preprocessing for pattern matching [Shahin,2008-A], [Miura,2004], [Mulyono,2008], whilst others used minutiae points on the image (bifurcations and ending points) like [Wang,2008] or [Ding,2005]. A completely different approach is the one presented by [Lin,2004], who apply a multi-resolution analysis over selected dominant points, represented as the x and y coordinates, the grey values, the temperature gradient, and the gradient direction inside the vein skeleton of the hand.

Finally, the matching step is accomplished using different approaches. [Ding,2005] use a structural matching computing distances between minutiae feature points. In a similar way, [Wang,2008] propose the use of the modified Hausdorff distance algorithm to evaluate the discriminant power of the set of feature points. However, this algorithm is quite sensitive to geometrical transformations on the data. On the other hand [Miura,2004], [Mulyono,2008] and, [Hashimoto,2006] use an improved template matching to compute the correlations between the input pattern and all the registered patterns. [Shahin,2008-A] perform a kind of registration process, in which a 2D transformation is carried out over the target image until a maximum correlation percentage between the target and the reference vein pattern is achieved.

To the author's knowledge there is no finger vein database available for the public research community. Therefore all the researches on this biometric trait build up their own vein database to evaluate the accuracy of their methods. However, some promising results are achieved regardless the small size of the databases used in the experiments. For example, [Shahin,2008-A] report and *EER* of 0.695% in personal identification, while [Miura,2004] reduce it to 0.145%, and in the case of [Wang,2008] an *EER* of 0% is reported.

- **Signature / Handwriting**

  Signature has been widely accepted for verification purposes, especially in commercial transactions, or in financial and contractual matters. It is considered a biometric trait showing both anatomic and behavioural characteristics of an individual influencing the way he or she signs. However, it is extremely conditioned by the individual behavioural component, as it changes over time and even between consecutive realisations, making it difficult to be used in high-security systems due to its intra-class high variability. Moreover, signature can be easily forged (especially by professionals) as it does not rely on any physical characteristic. On the other hand signature verification enjoys public acceptance as it has been extensively used in bank checks or bank card transactions.

  Signature verification can by classified into two different categories, according to the acquisition process used to collect the data: off-line signature verification and on-line signature verification also known as dynamic signature verification. In the former, signature is represented as an image while in the later, not only the shape of the signed name is used but velocity, acceleration, pressure and direction of the signature strokes are acquired using adequate equipment as functions of time. Dynamic signature verification approaches are more robust to skilled forgery, as they present almost the same shape but handwriting motion details will probably be quite different.

  Different approaches have been tested both for classification and for the extraction of discriminative features from off-line signature data. Since these methods deal with image data, most of them perform a preprocessing stage in

order to remove useless and noisy information from the image. This preprocessing step usually includes binarisation of the signature image, noise reduction and/or image resizing. The set of features extracted from the signature data and the classification methods vary from approach to approach. For instance, [Han,1995] use geometric (horizontal bars, vertical bars and loops) and topological features (end points, branch points, crossing points, convex points, and concave points) from which he builds a 2-dimensional symbol vector. In this case the matching criterion used is the 2D string longest common subsequence (LCS) shared by two 2D strings. Fuzzy and Neural Network classifiers were used in verification systems [Zhou,1996]. The system input was a set of four features referred as: reference pattern based features, global baseline, pressure features and slant features. [Porwik,2007] extract a set of four features (proportion factor, vertical and horizontal projections, centre of gravity and the Hough transform) and computed a weighted global similarity coefficient to determine the likelihood between two signatures. [Özgündüz,2005] propose the use of SVM's one-against-all method to verify and classify the signatures. In this case, each signature is represented as a set of features extracted from signature densities, signature-line directions and specific characteristics of the signature shape.

In the case of on-line verification systems, we can classify them regarding the type of features they used to build a signature representation. In other words, we can find systems based on global features (features are extracted for the whole signature, for instance average signing speed or Fourier descriptors of the signatures trajectory), local features (features are extracted for each sampled point, for instance distance and curvature change between successive points on the signature trajectory), segmental features (features are extracted from each segment in which the signature is divided) or a combination of some/all of them. [Jain,2008] present a set on 100 features for on-line signature verification, while [Richiardi,2005] review the most classically used global and local features (velocity in 2D, acceleration, pressure or pen azimuth among others). Additionally, [Richiardi,2005] present a feature dimensionality reduction method based on Fisher ratios and GMM classifiers assuming that the feature probability densities are unimodal Gaussians (which is not always true for every cited feature). The main problem on-line methods have to face is the variability in signing speed, which may lead to different signing trajectory lengths even for signatures belonging to the same person. Therefore, a method for normalising or aligning two signatures is commonly used in almost all systems, among which dynamic time warping (DTW) is the most popular. [Guru,2009] propose a system based on a set of 11 global features which represents each signature, and used Fuzzy C-means algorithm to classify the signatures, thus providing a semantic representation for each class. Additionally they provide a comparison between this method and other 14 state-of-the-art classification methods using the same database. [Kholmatov,2005] and [Alonso-Fernandez,2005] present two different systems using local features. The first one implements a Bayes classifier, an SVM classifier and a linear classifier in conjunction with PCA using three local features of the points on the signature trajectory: x–y coordinates relative to the first point of signature trajectory, the x and y coordinate differences, and the curvature differences between two consecutive points. [Alonso-Fernandez,2005] use a set of 7 discrete-time functions and first order derivatives of all of them but in this case, the classification method used

was HMM. [Kiran,2001], [Igarza,2003] and [Liu,2008] prosose different systems based on the combination of global and local features using GMMs, HMMs and a two-stage classifier based on a majority classifier and Neural Network classifier, respectively. Finally [Li,2006] propose an online signature verification using null component analysis (NCA) and PCA applied on segmental features computed using segmentation algorithm proposed in [Brault,1993].

Depending on the method used for signature recognition, i.e. static or dynamic, different databases are used. In the case of static methods only bitmap image analysis is performed, and they can be acquired using a scanner to acquire both signatures and handwritten text from white paper. However, in the case of dynamic methods, additional devices are needed which provide information about position, pressure time stamps, etc. of a signature. The first databases are known as off-line databases and the later ones as on-line databases. Due to its easy to collect properties of handwriting information and the popularisation of tablet PC's for data acquisition (in the case of online databases), most research groups used their own database both for handwriting and signature recognition [Nalwa,1997]. However, there are some public available databases for research activities:

- NLPR: The on-line handwriting database consists of both English and Chinese handwritten texts acquired from 149 writers in two sessions. Handwriting data is collected using a Wacom Intuos2 tablet. For each writer, 6 sentences are stored: 1 Chinese sentence of about 50 words, 1 English sentence of about 50 words, 2 additional and different from the 1 Chinese sentence of 50 words, and 2 additional and different from the 1 Chinese sentence of 50 words. In each writer file, the signature is represented as a sequence of points. The first line stores a single integer which is the total number of points in the writer file. Each of the following lines corresponds to one point characterised by features listed in the following order: x-coordinate, y-coordinate, time stamp, button status, azimuth, altitude, and pressure. More information can be found in (URL: NPRL).
- SUSIG: Is an online signature database collected in two different sessions using two different hardware acquisition devices. It contains over 5000 signatures and is divided into two different datasets. The Blind sub-corpus was collected using a Wacom's Graphire2 pressure-sensitive tablet in one session, and no visual feedback of the signature was provided. For each of the 100 individuals (25 women and 65 men) in the sub-corpus, eight or ten genuine signatures and a total of ten skilled forgeries from the same person were collected. Signatures in the Visual sub-corpus were collected using an Interlink Electronics's ePad-ink tablet which is able to display people's signatures while signing. For each of the 100 individuals (29 women and 71 men) in the sub-corpus, 20 genuine signatures (in two different sessions one week apart) and 5 skilled and 5 highly skilled forgeries were collected. Each signature in the database is saved as a text file, containing the x and y coordinates, time stamp, and pressure level for each point on the signature trajectory [Kholmatov,2009].

- BIOMET: This multimodal database includes among other biometric information, hand and online handwritten signatures collected from 327 individual, gender balanced. The database was registered in three different sessions. In the first session, 15 genuine online signatures and 17 impostor online signatures were collected for 130 individuals using a Wacom Intuos2 tablet. The x and y-coordinates, pressure, azimuth and altitude for each point on the trajectory were recorded. In the second and third recording campaign the same tablet was used, but a Wacom's Ink Pen was used, thus providing visual feedback and both online and hand handwritten signatures for the 197 left individuals. Forgery signatures were provided by five different impostors performing the imitation signatures [Garcia-Salicetti,2003].
- MCYT: Is a bimodal database that comprises both fingerprint and handwritten signatures as biometric features. Both on-line information and off-line information are considered in the database. 330 individuals were involved in the acquisition process. For each one, 25 genuine signatures and 25 additional highly skilled forgeries (provided by 5 different people) were collected. For the online handwritten signature acquisition process a WACOM pen tablet, model INTUOS A6 USB had been used, thus providing, for each point on the signature's trajectory, the x and y-coordinates, pressure, azimuth and altitude. No timestamps are provided [Ortega-Garcia,2003].
- SVC2004 / NISDCC-09: The main strength of these two different databases is the availability of benchmark results of signature verification systems that have participated in the SVC2004 competition or will participate in the sigcomp09. The SVC2004 database [Yeung,2004] was acquired in two different sessions using the WACOM Intuos tablet. The database contains 4000 signatures from 100 individuals, with 20 genuine signatures from each contributor (but not using his/her real daily used signature) and 20 skilled forgeries from at least four other contributors. The NISDCC database contains simultaneously acquired online and offline signatures. The collection contains 5 genuine signatures for each of the 12 genuine contributors (i.e. 60 authentic signatures) and 5 forgeries per authentic signature from 31 impostors. (I.e. 31 x 12 x 5= 1860 forgery signatures). The online signature is captured using an inking digitiser pen on a Wacom A4-oversized tablet, resulting in both digital information and green ink signatures on white paper. And the offline signature collection was scanned at 600dpi from the original paper registered signatures. The online dataset provides information about pen position, pressure, pen tilt and azimuth (URL: SIGCOMP09-NISDCC).
- Artist's Signatures: This is a curious hand-copied signature database as far as it contains 55000 signature examples by 50000 artists active from 1800 to the present. It is available on line at URL: artists' signatures.

- **Keystroke dynamics**

  Keystroke dynamics refers to the way a person types in a keyboard when using a computer. It is supposed that each person types in a specific way, thus allowing the recognition of a person identity while using a keyboard and without the use

of additional hardware unlike in other biometric recognition systems. Moreover, the used of an ID/Password combination is universally accepted as a method for individual verification when using computer equipments, so the incorporation of keystroke dynamics as an additional security step will be straight forward and will not produce people's rejection. Additionally, with this method, constant authorisation of a person using a system can be performed if constant unobtrusively monitoring of the keyboard is performed while the user is typing [Rybnik,2008]. As keystroke dynamics is a particular instance of behavioural biometrics, it can also be used to perform emotion recognition [Lv,2008].

Designing a keystroke recognition system involves different stages that have to be analysed:

- Database acquisition: Up to now, no standards have been defined to establish a protocol on data acquisition for keystroke recognition. As reported later, some authors acquired user data in one session [Chen,2004], [Jin,2008], [Monrose,1997], [Chang,2005], others used multiple sessions [Revett,2007], [Bergadano,2002], [Obaidat,1997], [Araujo,2005]. The number of samples collected from each user for training purposes is also variable, ranging from four samples (a fixed text of 683 characters) in [Bergadano,2002], 20 samples (username/password/first name /last name) in [Chen,2004] or 30 samples of their keystroke pattern for a particular string [Hosseinzadeh,2008].
- Feature extraction: Hold-time (also known as keystroke duration or dwell time, i.e. the time interval a particular key is held down) and keystroke latency (the time interval between two successive keystrokes) are the most popular features used for keystroke recognition [Obaidat,1997], [Araujo,2005], [Hosseinzadeh,2008]. Additionally, [Saevanee,2009] and [Monrose,2000] used the pressure applied to the key being held down and in the case of [Monrose,2000] the finger placement is also used as a feature. Another approach relies in the use of n-graphs, which refers to the elapsed time between the first key is pressed until the *n*th occurrence. [Bergadano,2002] used trigraphs in their experiments. [Shiv Subramaniam,2007] ruled out the used of graphs of order higher than three when using username/password recognition.
- Classification methods: Different approaches have been proposed for classification purposes. In [Bleha,1990], probably one of the earliest and most known researches in this area, a Bayesian statistical classifier and a minimum distance classifier (after a dimensionality reduction based on Fisher's Linear Discriminant) were used. Other works using statistical classifiers are, for example: [Araujo,2005], [Rybnik,2008], [Boechat,2007], [Monrose,2000], [Hu,2008]. Stochastic modelling using GMMs or HMMs have also been used in [Hosseinzadeh,2008] or [Chen,2004], respectively. Fuzzy logic has also been applied using a user's categorisation as output [Jin,2008], [Herbst,2008]. Finally, neural networks have also been used in some experiments: [Saevanee,2009] propose the use of a Probabilistic Neural Network (PNN) as the classification method, [Loy,2007] use the ARTMAP-FD and they also provide a comparison between this method and other machine learning systems in terms of performance. [Cho,2006] test different classification methods: k-Nearest Neighbour, back-propagation neural networks, and

Bayesian classification. [Obaidat,1997] provide a comprehensive review of different classification methods already cited.

The performance results presented by the different researches are, in most cases not comparable, mainly by the lack of a standard database for testing purposes. However, some works report promising results of this biometric characteristic for identification purposes and also compare different features and classification methods [Cho,2006], [Hosseinzadeh,2008].

Contrary to other biometric characteristics, no public database is available for keystroke dynamics research. Authors have compiled their own databases for their specific research and in most cases the size of these databases are quite small. The lack of a reference database does not allow performing a comparison between different approaches to the keystroke recognition problem. Moreover, there is not a standard procedure to collect a keystroke database resulting in different approaches to data acquisition. Some authors have collected the train and test data in the same session, [Chen,2004], [Jin,2008], [Chang,2005], which means that no much variation is going to be found among all the realisations of each user. This is quite important because keystroke biometrics is a behavioural characteristic that can vary over time due, for instance, to emotional state. Other authors have used data collected over different sessions. [Revett,2007] record logging information at 3 different periods of the day (morning, noon and early evening) over a 7-day period. [Bergadano,2002] collect information from 44 individuals over a one-month period, and additional 110 intruders were asked to provide a text sample. [Obaidat,1997] build database information from 15 individuals providing genuine and forgery logging information everyday for an eight-week period. In [Araujo,2005] information was collected using two different keyboards and in different sessions. Thirty individuals provide both genuine and forgery logging information (knowing the string typed and also knowing the way the other users type). Other authors like [Lv,2006] or [Hu,2008] do not even provide information about time acquisition procedures.

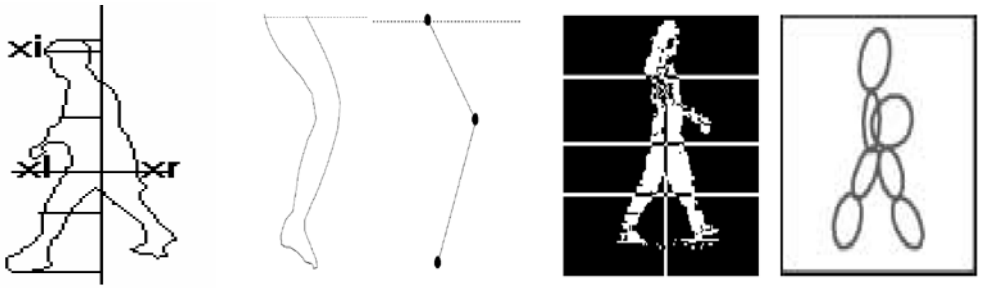
- **Gait**

Gait, defined as the way people walk is a quite novel characteristic used in biometric recognition. Most approaches use video information as the major source of information for gait recognition, therefore a great improvement has been made in this area as computational resources became sufficient to process video with reasonable performance [Nixon,2006]. Gait recognition is far away of achieving as good results as other biometric traits, mainly because of its small discriminatory power. However, it allows verification in some low-security or event oriented applications [Guo,2008], [Jain,2004-B], [Nixon,2006]. On the other side, gait is an attractive biometric feature for human identification because it can be acquired at distance, i.e., it can be measured at low resolution, and used in situations where face or iris information does not have sufficient resolution for recognition, it is difficult to disguise and it does not require a cooperative subject [Hong,2008], [Chai,2006].

In one of the earliest psychophysical studies of human perception of gait [Johansson,1975] people were recognised based on silhouette movement. Later on, gender identification attributed to anatomical differences, was performed in a similar manner, thus providing evidence that identity information is hidden in

gait and inspiring the development of vision-based algorithms for gait-based human recognition.

From a technological point of view, three main approaches have been considered for gait recognition purposes: vision-based recognition [Nixon,2006], floor sensor approaches [Middleton,2005], and wearable sensor based recognition [Gafurov,2007]. Technology aside, recent gait recognition methods can be roughly divided into two main classes: model-based methods and model-free methods. Both methodologies follow the general framework of image-processing, feature-extraction, feature correspondence and high-level processing. As reported in [Kale,2004-A], the main difference lies in the feature correspondence phase. Model-based approaches explicitly model the shape or motion trajectories of human body and the kinematics of joint angles, so feature correspondence is established by matching each frame of a walking sequence to the model. [Lee,2002] present a model based on specific information extracted from a set of ellipses in which the silhouette is divided and the mean and standard deviation of region features across time. Recognition is achieved using the K-nearest neighbour approach. In [Cunado,2003] a structural and temporal pendulum-based description of the thigh is developed and the k-nearest neighbour rule applied to the upper-leg motion Fourier components is used for classification. [Guo,2008] create an HMM gait-model for every known individual, based on anatomical information extracted from binarised silhouette. Classification is carried out using a Bayesian classifier. In a similar way, [Ye,2007] apply SVM's to model gait information extracted from individual binarised silhouettes after performing a data dimensionality reduction based on discrete wavelet transformation (DWT). The main advantages of model-based method are their ability to reduce the dimensionality to represent the data, and to deal with occlusion and noisy data. On the other hand, high computational cost is a drawback due to complex model search and matching.




a) b) c)

**Figure 1-11** Different approaches for human gait model. **a)** [Guo,2008] **b)** [Cunado,2003] **c)** [Lee,2002]

Model-free methods are based on the extraction of shape and/or movement information from the subject's silhouette, i.e., the motion pattern is characterised regardless the underlying structure of the body. [Kale,2004-A] propose two different set of features (the outer contour of the binarised silhouette and the binary silhouette itself) and two different approaches to gait recognition based on HMM representation of gait. In an *indirect approach*, frame to exemplar distance (FED) is used to extract structural and dynamic traits of a subject. The gait information reflected in the FED vector is modelled in a hidden Markov model (HMM). In the *direct approach*, a feature vector extracted from the image

is used to train the HMM model for each individual. In [Sarkar,2005] a four-part algorithm that relies on silhouette template matching was developed. After silhouette extraction, gait period is extracted to compute the similarity between two gait sequences based on spatio-temporal correlation. [Hong,2007] propose the use of sequences of mass vectors extracted from the binarised silhouette or its contour to represent human gait after applying PCA for dimensionality reduction purposes. DTW is chosen as the matching scheme. [Nandini,2008] use periodicity of legs and the maximal information compression index between silhouettes extracted from different stances of a person to perform gait recognition. The main advantage of model-free approaches is its lower computational complexity due to the use of dimensionality reduction techniques, especially PCA. Moreover, it is a holistic approach, so methods used for human gait recognition can be applied to non-human gait recognition with little modifications. A deeper and more detailed state of the art review can be found in [Nixon,2006].

A meta-analysis of gait identification rates as reported in state-of-the-art literature can be found in [Liu,2006-B], achieving over 90% of identification under situations where the training and test data are captured under similar conditions [Nixon,2006]. Performance of gait recognition systems is highly affected by the size of the database used in the research. Whereas it is comparatively effective regarding other biometric characteristics when tested with small databases, its performance is significantly reduced with larger ones (> 50 individuals) [Liu,2006-B]. Most of the more recent studies on gait recognition make use of the database collected during the DARPA HumanID program (USF/NIST HumanID gait challenge database). This dataset, see [Sarkar,2005], consists of 1870 sequences from 122 subjects spanning five covariates (surface type, shoe-wear type, weight carried, camera view-point, time acquisition). However, there are other databases also used for gait recognition research: CMU-MoBo database [Gross,2001], CASIA gait database (URL: CASIA), HID-UMD database [Kale,2004-A], [Kale,2004-B], the Southampton database [Shutler,2004], etc.

- **Odour**

  Despite being cited as a biometric characteristic for human identification purposes, it is still not clear whether body odour can be used in this way. Moreover, in a recent article posted in the Washington Post (URL: Washington Times), it is said that the United States Department of Homeland Security (DHS) through the department of Science and Technology, will conduct a research on this area to find evidences supporting the theory that an individual can be identify by its odour.

  A general approach to solve the problem of odour recognition, regardless human identification purposes, involves two main components: a sensing system that captures the odour and a pattern recognition platform that enables the recognition system to classify an odour based on the information supplied by the sensing system.

  As reported in [Korotkaya,2003] there are five categories of sensors available for building an electronic nose: Conductivity sensors, piezoelectric sensors, MOSFET odour-sensing devices, optical fibre sensors and spectrometry-based sensors. In the case of the pattern recognition subsystem for the electronic nose,

the most used approach is based on artificial neural networks [Ferreira,2005], although statistical methods and neuromorphic models have also been tested [Keller,1999].

The main problems affecting the used of odour as a biometric are

- Lack of an odour-sensing instrument (also known as electronic-nose) with enough capabilities to capture the volatile chemicals that the human body emits which make up the person's smell.
- The odour signature of the human body is affected by the use of deodorants or perfumes, diet and medication, odour transference between individuals or even environmental conditions such as pollution [Jain,2004-B], [Delac,2004].

- **DNA**

As far as DNA (Deoxyribonucleic acid) recognition is currently not performed by an automated method, there is no agreement on the research community on whether it can be considered a biometric trait. It provides the most accurate performance for positive person identification, as genotypic features are physical features that do not change under almost any circumstance and DNA model adaptation is not needed over time. Although it is extremely reliable, DNA recognition have some drawbacks when compared to other biometric characteristics

- It cannot distinguish between identical twins [Delac,2004].
- It is mainly used in the context of forensic applications, so DNA use can elicit some rejection among its users for its relation with crime investigation.
- Although DNA investigation is a hot topic and more research is needed to completely understand all the information embedded in it, it is clear that it carries information about susceptibilities of a person to certain diseases that can lead to some kind of discrimination, e.g., in hiring practices, enrolment in health insurance, etc.
- Real-time recognition is not currently possible, because DNA profile extraction involves the use of chemical methods that cannot be performed without expert involvement

A DNA profile can be generated using samples from blood, saliva, hair, bone or other body tissues. To extract this DNA profile different processes are involved related to biology, technology and genetics. Once the DNA profile from an individual is extracted, it is compared with other DNA profiles to establish the origin of the sample. The comparison is performed on the basis of samples, rather than on the basis of models as defined in other type of biometrics. More information on DNA forensics can be found in [Butler,2005], [Rudin,2001] and URL: DNA.

Although much research is done on DNA, due to its forensic application the development and maintenance of DNA databases for human identification purposes relies mainly in state governments and its access is usually restricted to law enforcement agencies.

- NDIS (National DNA Index System): It is the highest level of the CODIS (Combined DNA Index System) hierarchy, allowing the laboratories participating in the program to exchange and compare DNA profiles at the USA national level. The CODIS includes also a SDIS (State DNA Index System) and a LDIS (Local DNA Index System). By the end of 2008 the NDIS contains over 6,730,749 offender profiles and 248,943 forensic profiles. The profiles in the CODIS are categorised into one of the following categories: convicted offenders (individuals convicted of a crime), forensic (DNA profiles developed from crime scene evidence), arrestees (profiles of arrested people whenever the state law permits it), missing persons, biological relatives of missing persons and unidentified human remains. More information can be found in UTL: CODIS
- NDNAD (National DNA database in UK): The NDNAD includes DNA samples obtained both at crime scenes and taken from individuals in police custody. It is the largest DNA database (4,457,195 individuals) in proportionate terms in the world, with 7.39% of population in the database with no information of children younger than 10. More information about this database can be found in URL: DNA-database or URL: police.uk

Nevertheless, no single biometric characteristic will be able to fulfil all the requirements described by [Clarke,1994] and [Jain,1998] completely. In other words, no biometric trait is optimal taken alone; therefore the selection of one or another for a specific application will depend on the particular application. For example, a DNA-based technique is probably the most accurate technique however; it will be unsuitable for telephone base applications, as far as only voice-based biometric authentication seems to be applicable in that case. Moreover, error rate is not the only consideration to be taken into account. Other factors which have a strong influence when designing a biometric security system are: cost, integration in existing security infrastructure and user comfort, sense of hygiene (whether they require contact between the person and the sensor) and privacy (whether the information collected can be used for other purposes different from identification).

Depending on the security level that we want to achieve, and the specific requirements of the system, one or more of these biometric traits may be used, possibly under a multimodal fusion scheme

### 1.4.2 Multibiometrics

As we have already stated, some security systems combine different kinds of information for recognition purposes (something the person owns, something he/she knows and something he/she is). Another approach to increase the security of a system is biometric fusion. NIST defines biometric fusion as *"the use of multiple types of biometric data, or methods of processing, to improve the performance of a biometric system"*.

Multimodal biometric systems are expected to improve authentication rates if compared with unimodal systems due to two aspects related to the used of multiple traits: they ensure sufficient population coverage (not all people are suitable for all kind of biometrics) and provide anti-spoofing measures because an intruder needs to provide multiple biometric traits simultaneously. However, this kind of systems has also its

drawbacks. A multimodal biometric system requires the users to provide multiple traits, which may bother him/her. Additionally, the use of multiple sensors and the plausible increase of computational requirements will increase the system's cost.

Biometric fusion can be achieved through different techniques [Jain,2004-B], [Ross,2007] that include:

- Multiple sensor systems: Information from the same biometric is captured using different sensors. For example, voice can be captured using landline telephone, mobile telephone or different quality microphones, resulting in the acquisition of complementary information.
- Multiple biometrics: Different biometric traits (for instance, voice and face) are used to establish the identity of a user. These systems require the use of multiple sensors.
- Multi-sample systems: These systems used a single sensor to obtain multiple samples of the same biometric, in order to account for intra-class variations or to build an accurate model of the user. For instance, different images of the face in different positions or with different expression, or multiple voice utterances under different emotional conditions.
- Multi-instance systems: Refers to the acquisition of multiple instances of the same biometric trait. For example, palm prints of both hands or an image of each of the two irises can be acquired.
- Multi-algorithmic systems: This kind of systems applied different feature extraction techniques and/or matching algorithms on the same biometric characteristic.
- Hybrid systems: The term hybrid refers to the use of a combination of the different scenarios presented above. For example, we can combine a multi-biometric system and a multi-algorithmic system to improve the identification rates.

An additional classification can be performed regarding the stage at which the information fusion is done:

- Fusion before matching: Two different modalities can be classified into this category: sensor level fusion and feature level fusion. The first one refers to the used of information acquired using multiple sensors or multiple snapshots of the same biometric using a single sensor. In the case of feature fusion, the data obtained from different biometrics are combined into a single feature vector. [Samad,2007] proposed a multi-sample approach by fusing multiple spectrographic samples from different utterances at the score level, using the average operator. [Rattani,2007] proposed a feature fusion of fingerprint and face images. In order to make both set of points compatible for concatenation, a preprocessing step is applied consisting of feature normalisation, concatenation and reduction.
- Fusion at the matching score: Each biometric matcher provides a similarity score between the input feature vector and a model. These scores can be combined to improve the recognition rates. Additionally, if the biometric system operates in identification mode, the output of the biometric system is a ranking of identities instead of a match score, thus the aim of the rank fusion is to derive a consensus

rank based on the individual rankings. [Snelick,2005] experimented with five different fusion methods: simple-sum, min-score, max-scores, matcher weighting and user weighting. The matching scores were generated from three fingerprint and one face COTS (Commercial of-the-shell) biometric systems. [Monwar,2008] introduced a rank level fusion based on Borda count method and Logistic regression method. The unimodal ranks are obtained from face, ear and signature biometric systems.

- Fusion at the decision level: In this case, each biometric system provides its own recognition decision. Therefore, the decision fusion can be performed applying: AND/OR rules, majority voting schemes, etc. For instance, [Melin,2005] propose the use of a fuzzy system to implement the decision unit based on the results provided by the face, fingerprint and voice biometric systems.

The key when combining multiple biometric modalities is the selection of uncorrelated or negative correlated sources. A combination of this type will lead to higher improvement in performance.

There is a large amount of literature on fusion techniques in biometrics [Rodriguez,2008], [Ross,2007].

### 1.4.3 Multimodal biometric databases

Some researches in the area of multibiometrics used ad hoc databases compiled for the specific research carried on [Kim,2008], [Shahin,2008-B]. Others simply create a database combining unimodal biometric databases which provide information from different individuals, i.e. they associate an individual from a database with an individual from other database creating a virtual subject, as in the case of [Monwar,2008] or [Snelick,2005].

However, there are some multimodal databases available:

- NGI: NGI stands for Next Generation Identification (URL: NGI). It is not really a database but it's a program initiated by the FBI with the purpose of incrementally replace the IAFIS. The aim of the project is to incorporate to the fingerprint database other biometric modalities as they become cost effective, probably becoming one of the biggest multimodal biometric databases in the world, unfortunately only available for law enforcement purposes. Besides fingerprint information, it will include among other information, palm prints, face, iris and voice.

- BIOMET: BIOMET is a biometric database which includes five different characteristics: face, voice, fingerprint, hand, and signature data. The database was acquired in three different sessions with three and five months spacing between them. Among the 130 individuals who got involved in the database creation, only 91 completed the three sessions. The database is balance in gender and with ages ranging from 20 to 60 years. Audio and video (centred on face) were captured simultaneously using a digital camera. Additional face information is extracted with an infrared camera and a 3D acquisition system. 2D Hand information is captured with a scanner. Signature information includes both dynamic and static information. Finally fingerprint information was only collected for the index and middle finger [Garcia-Salicetti,2003].

- MyIdea: MyIDea includes talking face, audio, fingerprints and palmprints, signature, handwriting and hand geometry. 104 individuals were recorded over

three different sessions without any control on the interval between sessions. Like the BIOMET database, it offers two bimodal voice-signature and voice-handwriting simultaneous recordings. A digital video camera and a webcam were used for voice and face acquisition, additionally, high-quality audio is also available. For fingerprint acquisition two different sensors were used: optical and thermal sensors. Palm prints were acquired using a scanner, while the hand geometry information was extracted using a digital camera and a pegged-platform for standardisation purposes. Signature and handwriting was recorded in a similar way to BIOMET [Dumas,2005].

- MOBIO: MOBIO database contains both audio (both native and non-native English) and video (face) data recorded form 152 people (100 males and 52 females) from five different countries in 12 different sessions. The database was recorded under real noise conditions using two types of mobile devices: mobile phone (NOKIA N93i) and laptop computer (2008 MacBook). More technical details about MOBIO database can be found in [McCool,2012] and (URL: https://www.idiap.ch/dataset/mobio).
- BIOSEC: The BioSec corpus includes fingerprint, face, iris, and voice information. In the case of fingerprints, three different sensors were used to capture index and middle fingerprints of both hands. Four face images were acquired with a webcam, with different facial expressions in each shot. The database also includes four iris images of each eye. If the subject wears glasses, he/she is asked to take them off for iris acquisition but not for face registration. Voice is acquired in two different ways: using the webcam while face registration, and using a specific microphone. Four utterances of a user-specific 8-digits pin pronounced digit-by-digit is recorded. Additionally, three utterances of other users' pin to simulate informed forgeries are also recorded. The 8 digits were recorded both in Spanish and English. Two releases of the database are available. The BioSec-Baseline comprises information acquired from 200 subjects in two different sessions, while the BioSec-Extended comprises information acquired from 250 subjects in four different sessions. The age of the subjects range from 18 to 62 and it is not gender balanced [Fierrez,2007].
- XM2VTS: XM2VTS is a multi-modal face database captured onto high-quality digital video. It contains synchronised image and speech data from 295 individuals recorded in four different sessions at one month interval. Two recordings were made on each session. The first one includes a frontal-face video acquisition while the subject utters two sequences of digits and a phonetically balanced sentence. In the second part, video images were acquired while the subject rotated its head. When applicable, subjects were asked to take off their glasses [Messer,1999].
- MCYT: MCYT is a bimodal database which includes fingerprint and signature information collected from 330 individuals in a single session for each individual. The fingerprint sub-corpus consists of 12 samples from each finger of each individual acquired using an optical and a capacitive sensor. In the case of the signature sub-corpus, both on-line and off-line information are captured. For each subject, 25 true signatures and 25 skilled forgeries (produced by the five subsequent targets) were obtained using a WACOM pen tablet [Ortega-Garcia,2003].

- BiosecureID: This database registered in four different sessions distributed over 4 months, includes the following information from 400 subjects: voice, face, iris, fingerprints, hand, writing and signature and, keystroke biometrics. The database is gender balanced and covers the age range above 18. For each subject, ten short Spanish sentences and four realisations of a personal pin, uttered digit-by-digit were recorded in high-quality audio with noise cancellation. Iris information was acquired for both left and right eyes, twice. Although people were encouraged to remove their glasses during the acquisition process, the use of contact lenses was allowed. In the case of face, there are video and static images available. Four frontal static images are captured and a video sequence is recorded while the subject produced an 8-digits pin, thus voice and face biometric information is simultaneously recorded. The database also includes a total of 64 fingerprints for each individual, captured with a thermal and an optical sensor, from index and middle fingers of both hands. Complete 2D hand information is registered with a scanner. Behavioural biometrics where also included in the database, as both handwritten signature and text were registered, including on-line and off-line information. Finally, keystroke dynamics were included into the database, but registering only time information about keystroke for fixed text. The database contains information which simulates replay attacks for both speech and keystroke biometrics and skill forgeries for signature [Fierrez,2010].

A deep review of different multimodal databases can be found in [Faundez-Zanuy,2006].

| | # of subjects | # of sessions | # of characteristics | Voice | Fa | Ir. | Fp | Hp | Hw | Ks |
|---|---|---|---|---|---|---|---|---|---|---|
| **NGI** | n/a | n/a | n/a | x | x | x | x | x | | |
| **BIOMET** | 91 | 3 | 6 | x | x | | x | x | x | |
| **MyIdea** | 104 | 3 | 5 | x | x | | x | x | x | |
| **MOBIO** | 152 | 12 | 2 | x | x | | | | | |
| **BIOSEC** | 250 | 4 | 4 | x | x | x | x | | | |
| **XM2VTS** | 295 | 4 | 2 | x | x | | | | | |
| **MCYT** | 330 | 1 | 2 | | | | x | | x | |
| **BiosecureID** | 400 | 4 | 7 | x | x | x | x | x | x | x |

**Table 1-4** Summary of multimodal databases (Fa-face, Ir-iris, Fp-fingerprint, Hp-hand print, Hw-handwriting, Ks-Keystroke)

### 1.4.4 Biometric applications

Biometric-based authentication applications have experienced a great deployment in the last years, mainly due to cost reduction of biometric devices, the increase of computational power and the increase of authentication accuracy rates. According to [Jain,2004-B], biometric applications can be divided into different groups:

- Commercial applications.
    - Some notebooks (HP, Acer, Fujitsu, etc.) include a fingerprint sensor/reader which allows biometric login into the system instead of the classic user/password login. One of the latest companies in introducing it is Apple in the iPhone 5S.
    - Japanese financial (Mizuho Bank, Sumitomo Mitsui Banking, Bank of Kyoto) institutions have incorporated finger and palm vein

authentication technologies for their ATMs in order to increment security and reduce card fraud [Jones,2006].

- Biometric access control can be found in electronic games like the Password Journal® from Girl Tech® (URL: girltech), which incorporates a voice recognition application to provide access to the journal, or in building access control like the solutions propose by: IITS (UTL: iits), Sagem (URL: morpho), LG IrisAccess (URL: lgiris), or Mag-Gate physical access from Cogent Systems (URL: MagGate).
- Password reset via voice biometry is used by employees at Bankinter (URL: agnitio.es)

- Government applications.
  - The California Department of Social Services (CDSS) uses the Statewide Fingerprint Imaging Systems to eliminate duplicate aid in the State's public assistance program (URL: sfis).
  - The U.S. Department of Homeland Security introduced in 2004 the US-VISIT program to collect information about the travellers entering the USA. This program collects biometric information of individuals such as fingerprint or face image for authentication purposes. The United Kingdom (URL: ukba-homeoffice) or the United Arab emirates [Al-Raisi,2008] have also deployed a biometric system for border control, but based on iris information. At Ben-Gurion airport in Israel (URL: iaa-gov) the passport control is done by biometric reading of the hand.
  - Some nations use National ID systems that incorporate biometric information about the individual, such as the Spanish DNI (URL: DNI) based on fingerprint information or the British Identity Card (URL: ips-gov-uk), which registers iris, face and fingerprint information.
  - Access to the inner areas of the Nuclear power plants in USA, where vital equipment is located, is controlled among other systems, through the use hand geometry access control systems. (URL: nei-org)

- Forensic applications.
  - The FBI Laboratory's Combined DNA Index System (CODIS) is used for law enforcement purposes, matching DNA profiles with crime scenes and human remains. (URL: CODIS)
  - The IAFIS (Integrated Automated Fingerprint Identification System) provides automated fingerprint search capabilities for law enforcement purposes. Deployed in 1999, it holds the largest biometric database in the world (URL: IAFIS).
  - Suspect identification. There are many commercial applications which deals with the problem of suspect/terrorist detection based on facial recognition using video surveillance information, such as FRS Suspect Detection (URL: FRS-SD) or x-ident Video Identification (URL: x-ident VI).

Last trends in biometric technology and applications can be found in URL: Biometrics. Moreover, not only the Biometric Consortium maintains an on-line library of publications and a list of biometric vendors, but also sponsors the

Biometric Consortium Conference and Technology Expo (URL: Biometrics-events) which is focused on biometric technologies for homeland security, identity management, border crossing, electronic commerce, etc. This Expo is useful for surveying the commercial state of the art.

## 1.5 STATE OF THE ART IN SPEAKER RECOGNITION

Voice biometrics, also known as speaker recognition, can be defined as the application of speech processing techniques and theories to the identification of a specific person, in other words, it tries to establish the identity of a person analyzing its voice. Figure 1-12 shows how voice biometry relates to both areas previously cited (speech processing and biometrics).

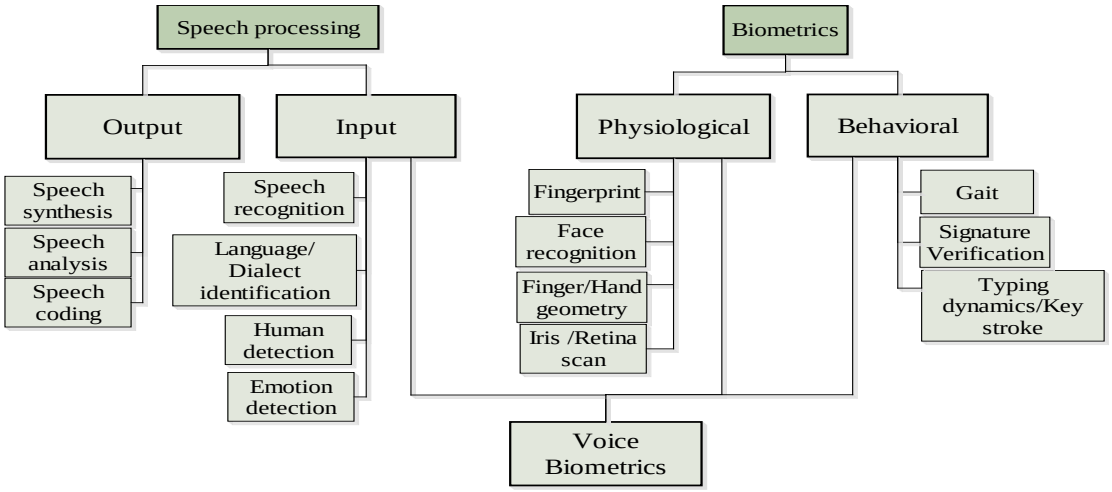



**Figure 1-12** Voice Biometry and its relation to Biometrics and Speech Processing

Speaker recognition is related to biometrics in the sense that the aim of biometric systems is to identify or recognise a person by using intrinsic characteristics of him/her. The use of voice, unlike other biometric characteristics, such as fingerprints, retinal pattern, genetic information, gait, etc., represents an additional challenge because it is influenced by both physiological and behavioural characteristics. However, the use of voice in a biometric system has also its advantages:

- Although our world is mostly visual, language is probably the most important feature that distinguishes humans from non-human beings. In this sense, the exchange of information among people is usually done by means of speech and this makes the use of voice for security purposes more acceptable than other characteristics such as retinal or iris scan (considered invasive), fingerprints (for its relation with criminality), etc.
- The use of voice as a biometric characteristic for identification does not require (generally speaking) additional expensive hardware, just a microphone to capture the voice. For example, nowadays, almost everybody have a mobile phone or a laptop, and in these devices there is always a microphone, which can be used to capture voice from a person and verify its identity before allowing him/her to use the device. However, it is more difficult to find a fingerprint reader, iris scanner, not to mention the hardware needed to perform a DNA test.

On the other hand, voice biometrics is related to speech processing, because to achieve its goal, it extracts the information from an utterance.

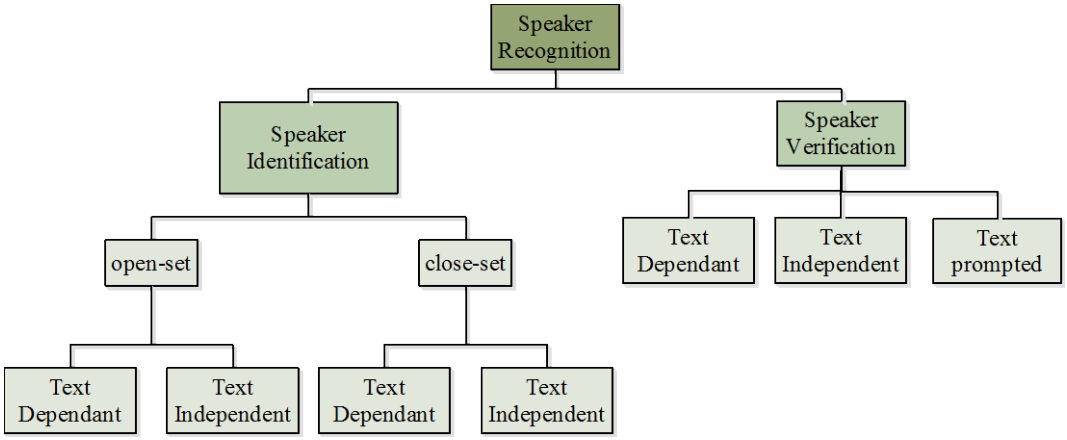



**Figure 1-13** Speaker Recognition

Voice Biometrics, from now on Speaker Recognition (SR), is a general task that can also be divided into two specific tasks that have been extensively studied over the last decades: Speaker Identification (SI) and Speaker Verification (SV) (see Figure 1-13).

Even though the main goal of both tasks is to establish the identity of the person responsible for a specific utterance, we can distinguish both tasks for the aim of its application. In the case of identification (Figure 1-14), the objective is to establish the identity of the person producing the utterance among a group of, N, previously modelled speakers, therefore known to the system, being the target speaker part of this set. This task is also known as closed-set speaker identification [Reynolds,1995-B]. If we assume, that the unlabelled voiced can belong to an unknown speaker, that is to say, not previously modelled, then the problem can be formulated as open-set speaker identification [Park,2006]. In this type of systems, it could be convenient to allow a dynamical update of speaker models.

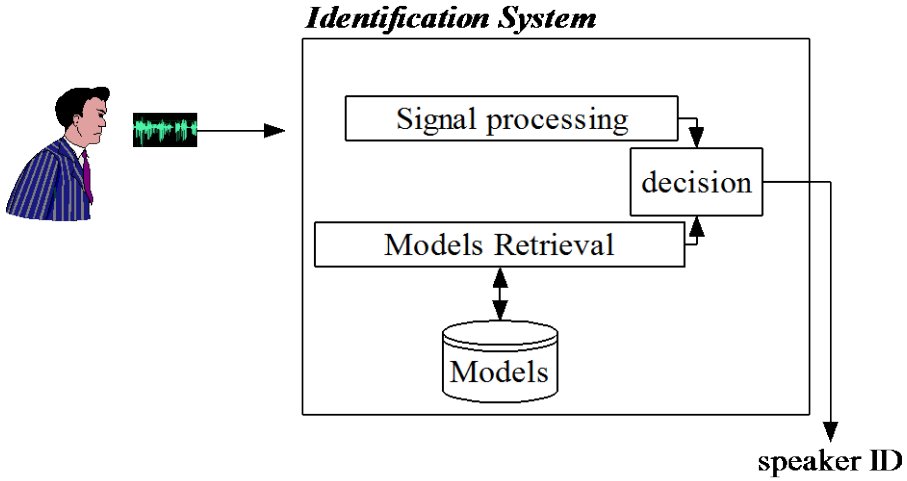



**Figure 1-14** General Speaker Identification System

In speaker verification systems (Figure 1-15), the goal is to determine if a person is who he or she claims to be. For this reason, it is also known as open-set speaker verification system, because it has to distinguish the speaker's voice, already modelled, among a huge set of possible impostors. Clearly, the decision to be made involves accept the utterance as belonging to the claimed speaker or reject the speaker, as it is supposed to be an impostor. Verification is a more complex task than identification, but it is also more challenging for its commercial applicability.

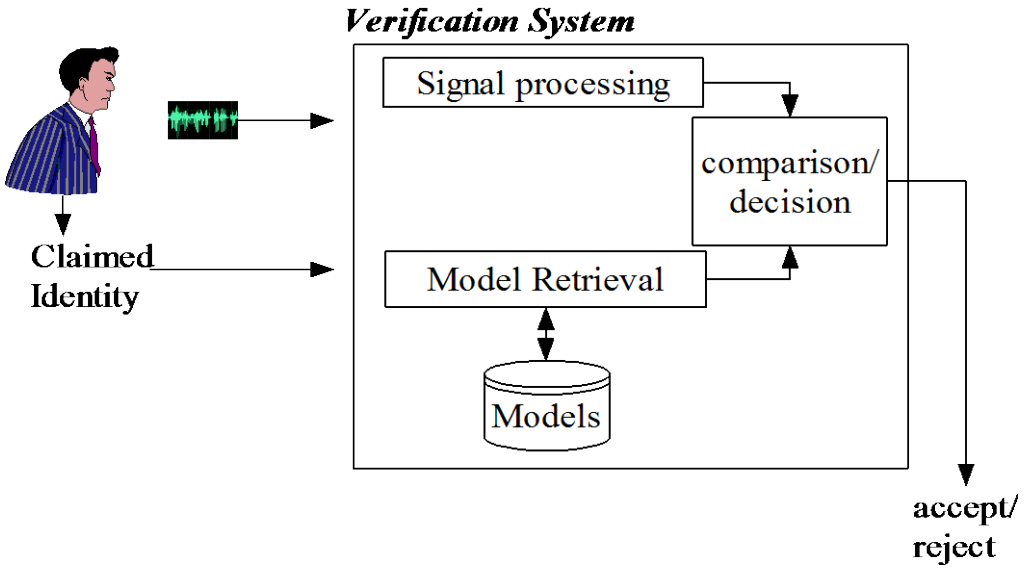



**Figure 1-15** General Speaker Verification System

Another characteristic of Speaker Recognition is its dependence or independence from the text used both to model the set of speakers who are going to be known to the system and to get access to the system. From this point of view, we can classify the system as:

- Text-independent: In this mode, the message emitted by the speaker is not constrained to any previous restriction.
- Text- dependent: In this kind of systems, the message produced for the verification process is already known by the system and by the speaker. This provides additional security, as far as the speaker should have not only the correct voice, but also has to know the correct password.
- Text- prompted: This is a special case of text-dependent system, because now it is the systems who ask the user to repeat a specific message that can be a sequence of digits, some specific words or some phrase.

Each of these modes has its pros and cons. For example, text-independent systems are more user-friendly as far as identification can be performed while the speaker is producing any message so he/she is not expected to say any specific text. However, it requires more data to train specific models and usually performs worse than text-dependent or text-prompted systems. Text-dependent systems perform better than text-independent systems and need less information to build speaker models, but require the user to be more collaborative. Finally, text-prompted systems provide additional security measures which make them suitable for security access control; because the user does not know in advance the specific message he/she will have to produce, thus making the system more robust to impostor access through the use of recordings.

So far we have only reviewed the text phase, but clearly, both in SI and in SV, it is necessary to perform a previous train phase in which the voice of each of the speakers supposed to use the system will be processed to get a model of each one. This model will be used later for comparison purposes.

Speaker recognition has been an active area of research since late 70s and early 80s [Atal,1976], [Doddington,1985]. There are even previous works on the area such as [Kersta,1962] which is one of the earliest works in using spectrograms in speaker

recognition. However it keeps on being an active area of research which means that speaker recognition has not achieved optimum performance and more research needs to be done.

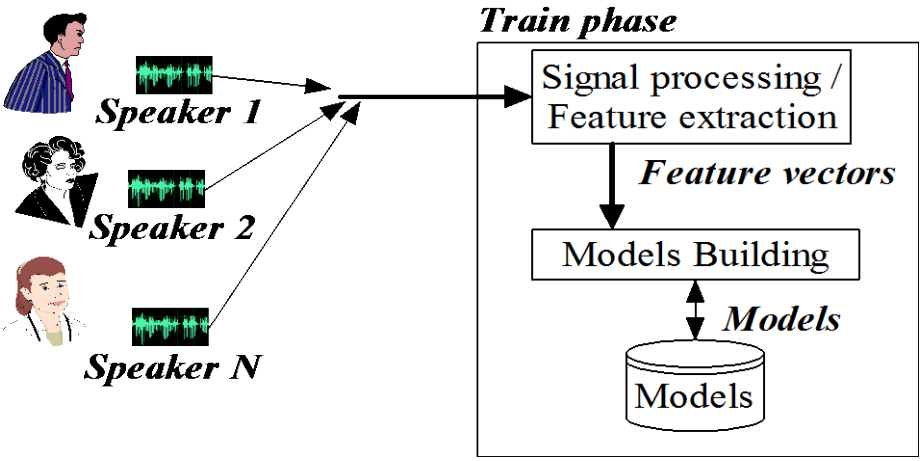



**Figure 1-16** Train/Enrolment phase

To effectively perform a revision of the state-of-the-art in SR systems, it is necessary to split it in the 4 different modules that are involved in such a system [Jain,2004-B]:

- Data acquisition: Responsible for capturing the voice signal. The study of this module is out of the scope of this thesis, and an extensive review does not make any sense, but just as a hint, it can be as simple as a PC microphone or more complicated as the whole telephone system or a specifically designed room with different types of microphones, to deal with microphone and channel variability during speech collection.
- Feature extraction: Responsible for extracting the most relevant features for the speaker recognition task from the data collected by the data acquisition module.
- Class Modelling/Class Decision: As it has been previously said, a model of each of the speakers involved in the system will be created. The class decision module is in charge of confirming a claimed identity (speaker verification system) or to establish the identity associated to a specific utterance (speaker identification). This decision will be taken comparing the input utterance with the models already known.
- Database: Used by the system to store the models or relevant information for each of the speakers. Depending on the specific application in which the system will be used, this database can be centralised or distributed, in this last case, each user will be in possession of its own model in a smart card for verification purposes.

Let's review how the feature extraction and the class problem have been classically addressed in the literature.

### 1.5.1 Feature extraction

A large amount of information can be extracted from the voice signal that can be useful for the speaker recognition task. The most important aspect in choosing a good set of parameters for speaker recognition is, generally speaking, that they provide high speaker discrimination power, that is to say high inter-speaker variability while keeping

low intra-speaker variability. More precisely we can enumerate some specific characteristics [Atal,1976], [Jain,2004-A] that are desirable when choosing the features:

- Universality: Each person should have the characteristic, although specific features, such as some kind of pathologies, not present in all people, can be also suitable for speaker recognition.
- Distinctiveness: Different speakers should be sufficiently different in terms of the characteristics. This also implies that the metrics should be not easily subject to mimicry or impersonation.
- Permanence: The characteristics should be stable over time and change little over different productions of speech.
- Collectability: The set of characteristics should be easy to measure and occur naturally and frequently in speech.

Classical approaches in speaker recognition focus in using, almost exclusively, a statistical spectrographic representation of voice. That is to say, devoting all the efforts to spectral information rather than temporal, although as it is already known, some important features for speaker characterisation are hidden in temporal patterns. Some examples of these temporal features can be the use of specific words or phrases, and some specific patterns related to them, such as intonation, temporisation within certain words, or other acoustic events [Scheffer,2005]. Taking into account the different nature of the features use for speaker recognition, we can classify feature extraction modules in three categories:

- Low level features: Biometric and spectral levels can be considered as low-level features. Biometric level refers to the use of specific characteristics in the speaker's production of voice difficult to impost as they are related to physiological or/and behavioural aspects. Among these features we can cite: short-term perturbation in the fundamental frequency (jitter) [Jankowski,1995], perturbations on the cycle-to-cycle phonation amplitude (shimmer) [Farrús,2007]. On the other hand, spectral level has been extensively used in speaker recognition systems for feature extraction. Typical methods in this spectral level are: *Short-time spectrum* (no matter if we use the exact representation or its approximation by filter banks) [Burget,2007], [Seddik,2004], [Xiang,2003], *predictor coefficients* (based on a linear model of speech production: [Park,2006]), *formant frequencies and bandwidth* (defined as the resonance frequencies of the vocal tract: [Fatima,2004]) or even the *formant trajectories* [Tanabian,2005].

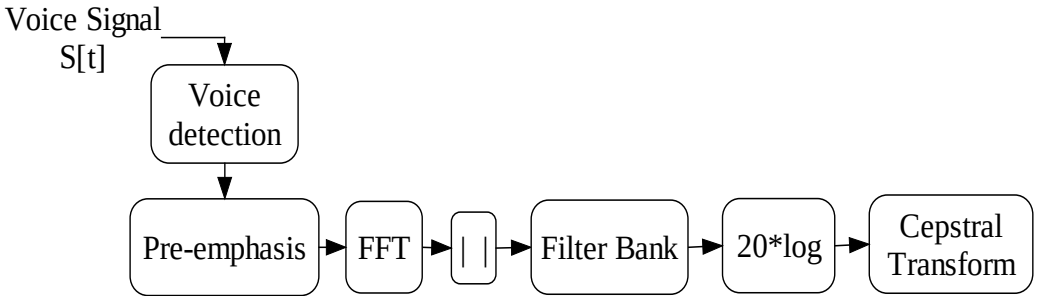



**Figure 1-17** General filter bank-based cepstral parameterisation

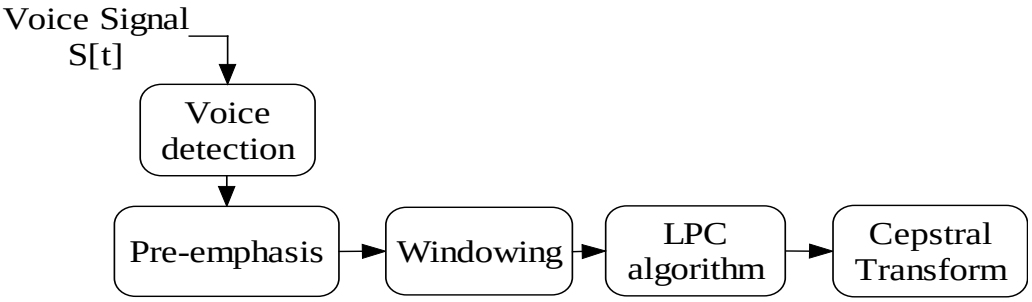



**Figure 1-18** General LPC-based cepstral parameterisation

However, the short-time spectrum has formed the basis for most works in characterizing speech in parametric form. The following procedure is usually followed: first of all, the voice signal is filtered to detect speech activity, erasing segments in which message is not present. Then, the power spectral density of the remaining segments is estimated by a specific procedure (FFT – see Figure 1-17, LPC – see Figure 1-18) over a sliding overlapped window whose duration in time is less than the whole signal (typically 20 ms). The use of sliding overlapped windows allows a good estimation of the temporal variability of the signal while granting the stationary assumption of the signal for processing. Finally, this PSD is parameterised to obtain the MFCC (Mel-Frequency Cepstral Coefficients) templates and optionally we can incorporate some dynamic information, using the $\Delta$ and $\Delta\Delta$ (polynomial approximations of the first and second derivatives), that give and estimation of how these MFCC features vary over the time. These feature vectors are aligned in streams to build templates for classification purposes.

One of the drawbacks of this low level feature approach is that they fail to capture long-term habitudes in a speaker's style, such as duration and pausing patterns, intonation and the use of specific words or phrases. Moreover, the performance of these systems is also affected by acoustic environment variations, noisy channels and microphone degradation.

Another important fact is that the MFCC obtained from the power spectrum hides some harmonic structure related to the fundamental frequency, which is considered a high level feature. Some attempts have been made to remove the influence of pitch in the speaker recognition systems, such as [Zilea,2003].

- High level features: This set of information reflects behavioural characteristics of speakers, such as prosody (pitch, duration and energy), phonetic information, pronunciation, emotion, stress, idiolect word usage, conversational patterns, etc. These differences in the speaking habits result from the manner in which people have learned to use their speech mechanism; but at the same time, the sociolinguistic context, the education and the socio-economic environment playing an important role in these differences. In high level feature recognition systems, a sequence of symbols is extracted from the acoustic signal, and speaker recognition can be carried out analyzing the frequency and occurrence of these symbols. The most investigated high level features for speaker recognition are:
    - Prosody: Prosodic level uses the fundamental frequency (F0) and energy variations over long-term segments (words or phrases) of voice to capture long-term information. The rhythmic level deals with speaking rates, pausing, balance between voiced sounds (presence of vocal chords

vibration) and unvoiced sounds (articulation without vocal fold vibration). [Sönmez,1998] use a stylised F0 contour to produce some linear model parameters in conjunction with intonation features (duration of continuously voiced regions and pauses). These parameters are then used as statistical features for speaker verification. [Kajarekar,2003] extract four types of prosodic features: pitch, duration, pause and energy, over different regions. These features are modelled using Gaussian Mixture Models (GMM). [Adami,2003] use bigrams to model the slope's sign of the F0, energy trajectories and duration of specific segments. Moreover, these features have also proven their value in emotion identification tasks as reported by [Barra,2006], and this is also a drawback, because pitch is susceptible of changing over time due to emotional state or even effort level.

- Phonetic information: Phonetic level deals with the detection of spectral structures known as phones, diphones or triphones in which specific articulation features are present. Usually Hidden Markov Models (HMM) are used for phone detection and modelling [Campbell,2003-B], [Kohler,2001]. This phone detection can be performed using phonetic transcriptions or using data-driven phone-like units derived directly from un-transcribed speech as in [El Hannani,2007]. The next step is to compute relative frequency of specific tokens through the relative frequency of each phone n-gram, and to use them as features for training the speaker's model [Hatch,2005]. Relative frequencies computed, typically, over bigrams or trigrams are then used as inputs to recognition algorithms based on likelihood ratios between target speaker's model and background speaker's model, resulting in an n-gram model for each speaker [El Hannani,2007]. Other approach uses Support Vector Machines (SVM). In these systems, a phone sequence is vectorised by computing frequencies of N-grams. This vector becomes the input of a SVM to produce a score, which is compared to a threshold to make a decision of acceptance or rejection [Campbell,2003-B].
- Idiolectal word and phone usage: The use of certain words or phrases by speakers, have proved to be useful in characterizing speakers for identification purposes. Moreover, intonation, stress and timing related to this words or phrases, are also useful features. This idea was implemented in [Doddington,2001], where the lexical content of the speech (bigrams, such as "you bet", "it were", "so forth", etc.) was modelled through the use of n-grams. In this work he applied conventional log likelihood ratios for test. [Sturim,2002] use cepstral GMM to build text-constrained GMM-UBM systems, modelling acoustic units or groups of units (specific words or groups of words). In recent approaches such as [Boakye,2004], speaker models consist of whole-word (word bigrams) models represented by adapted world-level HMMs. However, these techniques actually change a text-independent system to a text-dependent system or; as it is called, text-constrained speaker recognition systems.

The main drawback, as reported in different studies of this kind of systems [Doddington,1985], [Campbell,2007], [Campbell,2004-A], is its necessity for more information for both training and testing phases if compared to low level

feature systems. Another problem is that speaker recognition results tend to be corrupted by the errors in detecting phonetic events. On the other hand, [Petrovska-Delacrétaz,2007], pointed out that these high level features are less sensitive to noise and channel mismatch than the low level ones.

- Multilevel features: Simultaneously to the development of high-level feature systems, most researches tried also to fuse this information with that provided by low-level systems, yielding further improvements in speaker recognition. These improvements are due to the complementary characteristics of acoustic and high level features [Gonzalez-Rodriguez,2007]. Different methods have been used to fuse low-level and high-level feature system results: SVMs, GMMS, and MLP (MultiLayer Perceptron). In [Stolcke,2007-B], subsystem scores are used as inputs for an SVM-based combiner, trained to minimise the NIST decision cost function, reporting some improvements in speaker recognition performance. The low-level features subsystems used are classical cepstral SVM and GMM systems. For high-level feature subsystems, they use among others an SVM-based system that model syllables and word NERFs (Nonuniform extraction region features). [Garcia-Romero,2003] test two different fusion methods, SVM and Sum combination, yielding to an improvement in performance only for the SVM fusion method. In this case only two systems were fused: GMM acoustic system and word bigrams as reported in [Doddington,2001]. Again, GMM-UBM and SVM spectral based verification systems were used as low-level feature systems. As high-level feature systems they used the one described in [Doddington,2001], and also incorporated information from pitch, energy, prosody (intonation, rhythm and stress), and phone stream by both N-gram phone modelling and phone-SVM classifier. However, no improvements were reported due to cross-lingual degradation. [Campbell,2003-A] also used a single layer perceptron for fusion purposes, but also reported no significant differences between GMM and perceptron fusion systems. Different tests were carried out to determine which set of subsystems provide better results, leading to a 0.22% *EER* for the fusion of cepstral GMM-UBM system, pitch and energy track dynamics, duration and pitch related features, phone binary trees, and word n-grams. It is also relevant, that they achieved a 0.9% *EER* fusing only high-level feature systems.

From this review, we can extract some conclusions:

- Multilevel feature systems provide better performance than low-level and high-level alone, but a specific fusion strategy is necessary.
- The performance of the systems depends not only on the set of selected features, but on the classification method as well.
- Mel-frequency cepstral coefficients and their dynamic coefficients have been the dominant feature parameters, and used as baseline for systems comparison.
- Although there are some sets of features that provide a good performance for speaker recognition purposes, no best combination have been found yet.
- There are some works that used genetic algorithms for feature selection such as [Zamalloa,2006].

### 1.5.2 Class modelling and classification methods

As we have already pointed out, performance of speaker recognition systems highly depend on the class modelling and classification methods. That is to say, the same feature set may provide different performance depending on the classification used.

For speaker recognition, class modelling refers to the process of building a speaker model based on the specific features extracted from the voice signal. Classification refers to the process of evaluating the membership of a specific input feature vector, during test phase, to a specific class or speaker model. However, a perfect classification performance is most of times impossible, due to different problems in finding the appropriate features to build the models. For that reason, a more general task consists in determining the membership probability of a specific input vector for each class or category, and then deciding on the class for which the likelihood measure is largest.

Different class modelling techniques have been successfully tested in speaker recognition systems, which can be classified into three categories, depending on how they model the data (generative models, discriminative models and fusion). A review of all these techniques is beyond the scope of this document. More information on these methods can be found in [Friedman,1999], [Duda,2000], [Quatieri,2001], [Rabiner,1993], [Bengio,1995], [Burges,1998]. However, a more strict review of the specific techniques used will be presented latter.

- Generative models: These models are designed to capture the joint probability distribution of the training data. In other words, they compute the mean and variance of the training data available for a specific speaker. These methods have the advantage of using only data for the specific class to be modelled for training purposes, therefore adding a new class (speaker) to the speaker recognition system is done straight forward by learning its joint probability distribution without taking into account or modifying all previous models.

  The most popular generative method used for speaker recognition is without any doubt Gaussian mixture models (GMM). Presented in [Reynolds,1992], it has been accepted as a baseline for comparison with other methods, due to the high accurate results in text-independent speaker recognition applications. This method have also been used in text-dependent [Kim,2004-B], text-constrain [Sturim,2002] and text-independent [Reynolds,2000] speaker verification, in high-level feature speaker verification systems [Kajarekar,2003], and especially in low-level feature systems [Yang,2005], [Gómez,2008].

  Not only GMM's have been used as a generative method for SR. Hidden Markov Models, have been also used. Although they have been extensively used for speech recognition [Rabiner,1989], it has been successfully applied to SR as reported in [Kohler,2001], [Ang,1997], [Dumitru,2006].

  - Template models: Considered to be the simplest of the classifiers, they can be regarded as a special type of generative models because they only model the mean of the available data but not the variations. These methods were used in the past, and have been replaced mainly by GMMs or SVM, although they are still used in combination with other methods as in [Ilyas,2007]. As template models we cite Dynamic Time wrapping (DTW) and Vector Quantisation (VQ). Both methods have been extensively used in speaker recognition, although VQ mostly in text-independent systems and DTW in text-dependent systems [Yu,1995].

- Discriminative models: In this case, we are not interested in modelling the joint probability distribution of the training data, but just the discriminative regions of the distribution, that is to say, the border between the different classes (speakers). For that reason, for training a specific class or speaker it is necessary to provide training data for both the target speaker and all the other speakers supposedly using the system. So actually they model differences between speakers rather than a specific speaker. The main problem of this kind of systems is that if a new speaker is to be incorporated in the system, all the previous models in the system need to be recomputed. Moreover, these methods are less robust against impostors, since each speaker is modelled in relation with all the other known speakers. So a new impostor in the system will not be easily classified as an impostor.

  The discriminative methods most used in speaker recognition are: Support Vector Machines (SVM's) as in [Campbell,2002-A], [Campbell,2003-B], [Campbell,2004-A], [Campbell,2007], [Hatch,2005], [Lin,2006], [Stolcke,2008]; Neural Networks (NN) as in [Badran,2000], [Timms,1992], [Seddik,2004], especially Multilayer Perceptrons (MLP's) as in [Konig,1998], and polynomial classifiers as in [Campbell,2002-B].

- Fusion: Some works focus on fussing generative and discriminative methods to improve speaker recognition system's performance. There are different ways of combining these methods. We can use a generative method to train discriminative methods or the other way around, or more simply, we can fuse results from different methods in a collaborative way to get a global score from the individual scores. In this last case, each method can use different sets of features, allowing us to fuse results from low-level feature systems and high-level features.

  As an example of the first way of combining different methods in [Liu,2006-A] speech data is used to train a GMM adapted from a UBM. Then means from Gaussians are used to train an SVM system.

  [Hou,2003] used GMM's output to adjust the probabilistic output of SVM. Both SVMs and GMMs are trained independently using low-level features, and during the test phase the probability outputs are used to adjust the posterior probability output of SVMs. Another example in which results from low-level feature GMM and SVM systems are combined is [Campbell,2004-B], but instead of using a probabilistic approach they used both a linear combination and a perceptron approach to fuse the results. [Xiang,2003] used a multilayer GMM to model each of the speakers in the system and a multilayer UBM from which they adapted each of the speakers. The scores obtained from each layer are then fused using a multilayer perceptron to get a final decision.

  There are also other works in which, they fuse results from different systems which deal with different types of features, i.e., low-level feature systems and high-level feature systems. [Reynolds,2005] and [Campbell,2003-A] fused the scores from systems dealing with spectral, prosodic, phonetic and idiolectal sources using a perceptron classifier. Other approaches used SVMs as a fusion method [Garcia-Romero,2003].

### 1.5.3 Class decision

The last step in a speaker recognition system is the decision making, i.e., whether to accept or reject an utterance as belonging to a specific user. In the speaker identification task, there is no threshold, as the most likely model for the input utterance is chosen as the target speaker. So this problem mainly affects speaker verification tasks. Assuming that we can compute the likelihood that an input utterance belongs to a specific speaker (model) then, if this likelihood (or matching score) is higher than a decision threshold the speaker will be accepted as the origin of the utterance. Although computing the likelihood is a straight forward task once we have chosen the feature extraction and the modelling technique, finding an optimum threshold is not easy and is usually fixed empirically.

Why is it so difficult to find a threshold? In [Li,1988], it is shown that there is a strong variability of matching scores even for a single speaker. And conclude *"no absolute threshold on raw scores can be chosen which will give reliable decisions even if the threshold is speaker dependent"*. This matching score variability is due to different aspects:

- Unless specifically design, training data is going to be very different form each speaker in the system. These differences are related with sound acquisition device variations, transmission variability, environment noise, phonetic content and finally the amount of available training data between speakers, which can also be a problem.
- As for the training data, test data also have a deep impact on scores. Even in the case in which differences in training data are negligible, test utterances can be captured in different conditions as the ones in training data. So the problems are the same: channel and acquisition devices mismatch, environment noise, duration and phonetic content of test segments.
- Finally, the matching scores are also affected by intra-speaker variability, i.e., variations in speaker's voice due to different issues: emotional state, health problems or even changes over age.

To make speaker recognition tasks more robust to these problems, several score normalisation techniques have been investigated. Although a deep review on normalisation techniques can be found in [Bimbot,2004] and [Auckenthaler,2000] where a performance comparison between different normalisation techniques is carried out, the most popular normalisation techniques are:

- Cohort-base and world model normalisation: These score normalisation techniques work under the assumption that an input utterance from a known speaker (or target) will match its reference model better than they match other speaker's model, as the reference model is built from utterances from the same speaker as the input utterance.

  Speaker verification systems that use these normalisation techniques are also known as likelihood-ratio-based speaker verification systems, because to decide whether to accept or reject an utterance the following test is performed:

$$\frac{P(X|\lambda_{tar})}{P(X|\lambda_{\overline{tar}})} \begin{cases} > \theta, accept \\ < \theta, reject \end{cases} \qquad \textbf{Eq. (1-4)}$$

The difference between cohort-base normalisation and world base normalisation is based on how they build $\lambda_{\overline{tar}}$, as it must represent, at least in theory, the whole space of possible impostors to the target speaker. In the case of using the cohort approach, we can compute $P(X|\lambda_{\overline{tar}})$ as:

$$P(X|\lambda_{\overline{tar}}) = f\{P(X|\lambda_1), \ldots, P(X|\lambda_b)\} \quad \textbf{Eq. (1-5)}$$

where $B=\{\lambda_1,\ldots,\lambda_b\}$, is a set of background speakers, and $f$ is a function over all the probabilities that the input utterance belongs to a background speaker. Examples of cohort normalisation can be found for example in [Reynolds,1995-A], where $f$ is defined as $P(X|\lambda_{\overline{tar}}) = \frac{1}{B}\sum_{b=1}^{B} P(X|\lambda_b)$, or in [Ilyas,2007].

The main problem in cohort normalisation is how to select the background speaker models (same or different sex as the target speaker, same or different age, etc.) and the number of background speakers that represent the population of expected impostors (ideally the larger the better, but there are still computational and storage constraints). Some works have also been carried out on cohort selection [Colombi,1996]. To overcome these problems, the set of background models was replaced with a single model (also known as world model or universal background model). This idea, introduced by [Carey,1991], consists in building a general or world model using utterances from a large number of speakers of the population in general, and use this model as the alternative (non-target) hypothesis. This method has been extensively used with successful results [Heck,1997], [Reynolds,2000], [Sturim,2002], [Zheng,2006].

- Zero Normalisation: Also known as ZNorm, it was first introduced by [Li,1988]. In this case, the normalised score is given by:

$$S_{ZNorm} = \frac{\log\big(P(X|\lambda_{tar})\big) - \mu_\lambda}{\sigma_\lambda} \quad \textbf{Eq. (1-6)}$$

where $\mu_\lambda$ and $\sigma_\lambda$ are the mean and standard deviation for the impostor distribution. In other words, the aim of ZNorm is to scale and shift the distribution of scores between a target speaker and a set of impostor utterances to the standard normal distribution. To estimate these values, $\mu_\lambda$ and $\sigma_\lambda$, the target speaker model is tested against utterances from impostors, this result in a set of likelihood scores from which the impostor distribution is estimated. This normalisation technique has also been applied in [Barras,2003] for speaker verification over cellular data.

- Handset Normalisation: Also known as HNorm. This normalisation technique arises from the problem presented in [Reynolds,1996], in relation with handset variability between utterances collected for training and testing. In this work several channel compensation techniques and a new cepstral mean subtraction was proposed without a great improvement. So in [Reynolds,2000] a variation of the ZNorm, called HNorm was proposed to improve performance. During the test phase, the handset type for each utterance is established and the handset specific normalisation parameters are applied:

$$S_{HNorm} = \frac{\log\big(P(X|\lambda_{tar})\big) - \mu\big(HS(X)\big)}{\sigma\big(HS(X)\big)} \quad \textbf{Eq. (1-7)}$$

To compute handset normalisation parameters, log-likelihood ratio scores are computed for each speaker model using handset-dependent impostor test segments. From these results, and assuming they have a Gaussian distribution, mean and standard deviation for each handset type are computed.

- Test Normalisation: Also known as TNorm, proposed by [Auckenthaler,2000], it can be considered as a mix-up between ZNorm and cohort normalisation, as it uses mean and standard deviation as normalisation parameters, and during the test a set of impostors is used to estimate log-likelihood for test input utterance. So a mean and variance is computed for these impostor scores and the normalised score is computed like in previous normalisation techniques. This type of normalisation is also used in [Barras,2003] and [Hebert,2005].
- ZT Normalisation: Also known as ZTNorm, it cannot be considered as a new normalisation technique, but it is appropriate to cite it, as it is widely used in speaker recognition systems. Actually ZTNorm is the combination of ZNorm and TNorm.
- Symmetric Normalisation: Also known as SNorm has been proposed in [Shum,2010] and applied to the cosine similarity decision (CSD) metric. In this case, the normalisation parameters are applied to both the speaker model and the test utterance as follows. First of all a set, $I_{imp}$, of impostor utterances and impostor models must be defined. Then for each speaker model $\lambda_i/i=1...n$, and for each test utterance $X_j/j=1...k$, the SNorm parameters can be defined as the mean and standard deviation of the scores between the given utterance or speaker model (i.e. $X_j$ or $\lambda_i$) and all the elements of $I_{imp}$, obtaining for each model, $\lambda_i$, and test utterance, $X_j$, the following values: $SNorm(\lambda_i)=(\mu_{\lambda i},\sigma_{\lambda i})$ and $SNorm(X_j)=(\mu_{Xj},\sigma_{Xj})$. Finally for a specific trial, the score can be computed as follows:

$$score_{SNorm}(\lambda_i, X_j) = \frac{score_{CSD}(\lambda_i, X_j) - \mu_{\lambda i}}{\sigma_{\lambda i}} + \frac{score_{CSD}(\lambda_i, X_j) - \mu_{Xj}}{\sigma_{Xj}} \quad \textbf{Eq. (1-8)}$$

- Distance Normalisation: Originally proposed by [Ben,2002], distance normalisation, also kwon as DNorm, is based on the use of the Kullback-Leibler (*KL*) distance between the speaker model and the world model. The normalised score is given by:

$$score_{DNorm} = \frac{log(P(X|\lambda_{tar}))}{KL2(\lambda_{tar})} + \frac{score_{CSD}(\lambda_i, X_j) - \mu_{Xj}}{\sigma_{Xj}} \quad \textbf{Eq. (1-9)}$$

This means that the score for a speaker $X$, with claimed identity $\lambda_{tar}$, is normalised with the symmetrised *KL* distance corresponding to the speaker $\lambda_{tar}$. A Monte-Carlo method is used to estimate GMM synthetic data that client and world models follow, which are used to compute the *KL* distance of the client and world model. According to [Ben,2002], DNorm performs comparably to ZNorm.

### 1.5.4 Experimental recognition results

Table 1-5, presents recognition results achieved in some representative works on Speaker recognition. For each system, the database used on the experiments, the amount of information used for training each speaker model, the type of speech used (Telephone

channel/microphone channel), normalisation methods, features and classifiers employed, and best *EER*s obtained are presented. The *EER* quality measure is not directly comparable since experiments are conducted under very different scenarios: different databases, different number of train and test speakers and different amounts of training information per speaker. In order to compare two different systems, both should be tested under the same conditions, which is not the case in the following Table. Therefore, this table simply allows us to verify which are the different parameters and classifiers mostly used as under different scenarios.

| *Reference* | *Database* | *Information /speaker (train)* | *Channel Type* | *Normalisation* | *Features* | *Classifier* | *EER* |
|---|---|---|---|---|---|---|---|
| **[Adami,2003]** | NIST SRE 2001 (Switch board -I) | 2/4/8/16/32 minutes | Telephone | Cohort | 3-systems:<br>• Log F0, log-Energy, first order derivatives of them<br>• Sign of the F0 and energy slopes. Duration of segment. Phone information<br>• F0 contours + DTW (on selected words) | GMM/UBM<br>DTW<br>Single-layer perceptron fusion. | 3.7% |
| **[Ang,1997]** | TIMIT | 8 sentences | Microphone | | LPC-based Cepstral coefficients | VQ<br>Phoneme-based HMM | 4.5% |
| **[Auckenthaler,2000]** | NIST SRE-1998 (SwitchBoard-II phase 2) | 2 minutes | Telephone | Cohort<br>ZNorm<br>TNorm<br>HTNorm | 12 Cespstral +Energy + Δ + ΔΔ | GMM | ≈3.5% |
| **[Barras, 2003]** | NIST SRE-1998 (SwitchBoard-II phase 4) | 2 minutes | Telephone | CMS<br>Variance norm.<br>Feature warping<br>TNorm | 15 MEL-PLP cepstrum + Δ + ΔEnergy | GMM | 8.5% |
| **[Xiang,2003]** | NIST SRE 1999 (Switchboard-II phase 3) | 2 minutes | Telephone | Feature warping<br>SUBM(structural UBM) | 19 MFCCs + Δ. | SGMM(structural GMM)<br>MLP | 12.1% |

| *Reference* | *Database* | *Information /speaker (train)* | *Channel Type* | *Normalisation* | *Features* | *Classifier* | *EER* |
|---|---|---|---|---|---|---|---|
| **[Boakye, 2004]** | NIST SRE 2001 (Switchboard-I) | 2/4/8/16/32 minutes | Telephone | CMS /UBM | 19 MFCCs + C0 + Δ | Word-based HMM | 1.25% |
| **[Burget,2007]** | NIST SRE 2006 | 5 minutes | Telephone | Feature Warping<br>CMS<br>HDLA<br>Eigenchannel Adaptation<br>UBM<br>TNorm | 13 MFCC + C0 + Δ + ΔΔ + ΔΔΔ. | GMM | 4.7% |
| **[Campbell, 2007]** | NIST SRE 2005 (Mixer) | 40 minutes | Telephone | RASTA<br>CMS | 3 systems:<br>• 19 MFCC + Δ (SVM)<br>• Word sequences (SVM)<br>• Phone sequences (SVM) | SVM<br>Fusion based on linear equal weighting of scores. | 3.43% |
| **[Campbell,2002-B]** | Yoho database | 288 two-digit numbers (6 minutes) | Microphone | UBM | 12 LPCC + Δ | Polynomial classifier | 0.07% |
| **[Campbell,2004-B]** | NIST SRE 2003 (Cellular switchboard) | 2 minutes | Telephone | TNorm<br>UBM | 18 LPCC + Δ (SVM)<br>19 MFCC + Δ (GMM) | SVM<br>GMM | 5.73% |

| *Reference* | *Database* | *Information /speaker (train)* | *Channel Type* | *Normalisation* | *Features* | *Classifier* | *EER* |
|---|---|---|---|---|---|---|---|
| **[Campbell,2003-A]** | Switchboard-I | 2/4/8/16/32 minutes | Telephone | UBM | 5-systems:<br>• Cepstral features (GMM/UBM).<br>• Pitch and energy slopes, dynamics and phoneme context<br>• Prosodic statistics<br>• Phone binary trees<br>• Word n-gram idiolect | Single layer perceptron fusion | 0.22% |
| **[El Hannani,2007]** | NIST SRE 2006 (English trials) | 40 minutes | Telephone | UBM | 3 systems:<br>• 16 LFCC + Δ + ΔEnergy (GMM)<br>• 15 MFCC + energy + Δ (ALISP symbols – 3-grams HMM)<br>• Symbol duration (GMM) | GMM<br>HMM<br>SVM fusion | 5% |
| **[Farrús,2007]** | Switchboard-I | 40 minutes | Telephone | UBM<br>Cohort<br>ZNorm | 2 systems:<br>• Prosody (word and segmental duration, fundamental frequency, jitter and shimmer)<br>• Spectrum (20 FF + Δ + acceleration) | GMM<br>k-nearest neighbour<br>weighed fusion. | 6.8% |

| *Reference* | *Database* | *Information /speaker (train)* | *Channel Type* | *Normalisation* | *Features* | *Classifier* | *EER* |
|---|---|---|---|---|---|---|---|
| **[Hou,2003]** | NIST SRE 2003 | 30 seconds | Telephone | UBM | 12 MFCC + Δ + ΔEnergy | SVM with GMM adjustment | 6.1% |
| **[Gómez, 2008]** | ALBAYZIN | 30 seconds | Microphone | | 14 MFCC (raw voice) + 14 MFCC (Vocal tract) + 8 MFCC (Glottal Source) + pitch + energy | GMM | 0.35% |
| **[Gonzalez-Rodriguez,2007]** | NIST SRE 2005 | 40 minutes | Telephone | Cohort KL-TNorm | 3 systems:<br>• MFCC (GMM + SVM)<br>• phone tokens (HMM)<br>• prosodic tokens (HMM) | GMM<br>SVM<br>HMM<br>SVM fusion | 7.13% |
| **[Ilyas,2007]** | Malay Spoken digit database | 0 to 9 digits | Microphone | | 14 LPCC | VQ + HMM | 11.72% |
| **[Kim,2004-B]** | Yoho database | 1,5 minutes | Microphone | | 12 PSMFCC (pitch synchronous MFCC) | GMM | 2.83% |
| **[Kohler, 2001]** | NIST SRE 2001 (Switchboard-I) | 20 minutes | Telephone | UBM | (12 Cepstral + Δ) Phone n-grams | HMM / LLR | 7.2% |

| *Reference* | *Database* | *Information /speaker (train)* | *Channel Type* | *Normalisation* | *Features* | *Classifier* | *EER* |
|---|---|---|---|---|---|---|---|
| **[Konig,1998]** | NIST SRE 1997 (SwitchBoard-II phase 1) | 2 minutes | Telephone | | 2 systems:<br>• NLDA features<br>• 18 Cepstral + pitch | MLP | 7.1% |
| **[Liu,2006-A]** | NIST SRE 2004 | 2 minutes | Telephone | UBM | 16 MFCC + Δ | GMM<br>SVM | 11.92% |
| **[Stolcke,2007-B]** | NIST SRE 2006 (English trials) | 5 minutes | Telephone | Feature-level intersession variability with nuisance attribute projection<br>Factor analysis<br>TNorm<br>UBM | 8 systems:<br>• Cepstral (GMM)<br>• Cesptral (SVM)<br>• Gaussian supervector (SVM)<br>• N-gram frequencies that record duration of frequent words (SVM)<br>• SNERFS + GNERFS (SVM) → prosodic features over syllables and words.<br>• Word duration (GMM)<br>• Phone duration (GMM)<br>• MLLR (SVM) | GMM<br>SVM | 2.59% |

**Table 1-5** Different recognition results in speaker recognition

## 1.6 SPEAKER RECOGNITION COMMERCIAL APPLICATIONS

Commercial applications based on speaker recognition goes beyond its incorporation in electronic games, as previously cited. There are several companies that offer voice biometry solutions for identification purposes. Table 1-6 provides a list of different commercial applications with a brief description of their capabilities or applications. Obviously, none of these commercial applications provide information related to the technology used in their systems. They just provide software development kits, API's, etc., i.e., they provide black box solutions for speaker recognition.

| *Company* | *Website* | *Product* | *Brief description* |
|---|---|---|---|
| **NUANCE** | http://www.nuance.com/verifier/ | Nuance Verifier 4.0 | Nuance Verifier provides secure access to sensitive information over the telephone. The system is language independent and performs an ongoing adaptation of voiceprint characteristics as voices change or age, improving the quality of voiceprints for faster verification. Nuance Verifier 4.0 provides enrolment and verification using rotating questions, or even verifies callers in the background while the callers are completing other tasks. |
| **Speech Sentinel Ltd.** | http://www.securivox.co.uk/ | SecuriVox | Reading between the lines, due to the lack of much information, we can infer that this system can work in text-independent mode, for internet, telephony or electronic device based verification applications. |
| **MEK Soft. Technologies** | http://www.meksofttech.com/ | SecurPBX | Provides a text-dependent verification system to secure the access to organisation's PBX via telephone. |
| **IBM** | http://www-01.ibm.com/software/pervasive/voice_server/fpsiv/ | WebSphere | According to the manufacturer, WebSphere provides a speaker verification technology which is grammar, language, and text independent, i.e., you can enrol saying anything, in any language, and have it verify you, saying anything, in any language. Additionally, the system provides Speaker Tracking, and Speaker Change Detection mechanism. |

| Company | Website | Product | Brief description |
|---|---|---|---|
| **AGNITIO** | http://www.agnitio.es/ | • KIVOX | The application developed by Agnitio, is channel-independent and language-independent, providing both text-dependent and text-independent solutions. Additionally it provides automatic conversation indexing per speaker. |
| **AUTHENTIFY** | http://www.authentify.com/ | • Authentify Solutions | The Authentify solution employs a model of voiceprint comparison known as *text independent directed* speech. In this model, verification is performed against a phrase that is randomly generated. |
| **ZEHU Technologies** | http://www.zehu.com/ | • Zehu authenticator | Zehu's products are designed as software development kits (SDKs) that provide a package of APIs, libraries and tools to OEM applications. Zehu's application includes mechanisms to adapt to the changes in the user voice over time and is secure against playback attacks. It also provides both text-dependent and text-independent speaker verification. |
| **DIAPHONICS** | http://www.diaphonics.com/ | • Spike Server | Spike Server is an integrated hardware and software platform that verifies the identity of callers with biometric voice verification. In addition to verifying identity, Spike Server also records voice transactions, and creates a secure audit trail of all interactions with the system |
| **PORTICUS** | http://www.porticusinc.com/ | • Porticus Versona | Versona delivers speaker authentication solutions for enterprise systems, wireless carriers, IVR systems/call centres and device manufacturers. Versona relies on the physiological aspects of the human vocal tract, and as a result, is less susceptible to background noise interference, recorded playback and intra-speaker variability. |

| *Company* | *Website* | *Product* | *Brief description* |
|---|---|---|---|
| **VoiceVerified** | http://www.voiceverified.com/ | • Voice Authentication Suite 2.0<br>• Vocal Reset 2.0<br>• Vocal PIN 2.0<br>• Vocal Time Tracker 2.0 | VoiceVerified offers 4 different products that can be used with four different voice biometric engines: **Static Passphrase** (An engine that allows users to enrol by repeating simple phrases or numeric strings → text-dependent), **Random Fusion** (An engine that allows users to enrol using random numeric strings → text-dependent ), **Mixed Case** – (A combination of the static and random engine), N**atural Speech** (An engine that allows users to speak naturally in order to enrol or verify → text-independent) |
| **ANOVEA** | http://www.anovea.com/ | • ANOVEA Authentication Technology | The Anovea Speaker Authentication System provides high noise immunity, minimal bandwidth requirements, and high-channel invariance. Additionally it can automatically adapt to the natural variability of each speaker's voice. Both enrolment and verification steps are text-dependent. |
| **PerSay** | http://www.persay.com/ | • VocalPassword<br>• FreeSpeech<br>• S.P.I.D | PerSay provides language- and accent-independent speaker verification solutions through 3 different products: VocalPassword™ 6.5 (A biometric speaker verification system that verifies a speaker during an interaction with a voice application. It supports text-dependent, text-independent and text-prompted technology), FreeSpeech™ 6.5 (A unique text-independent biometric speaker verification system that transparently verifies the identity of a speaker during the course of a natural conversation) and S.P.I.D™ 6.5 – (An advanced voice mining and speaker identification system for law enforcement and intelligence agencies). |

| *Company* | *Website* | *Product* | *Brief description* |
|---|---|---|---|
| **Voicevault** | http://www.voicevault.com/ | • Password Reset<br>• Caller Authentication<br>• Payment Verification<br>• Voice Sign<br>• Web authentication<br>• Voice Track | VoiceVault uses spoken words to calculate measurements of the speaker's vocal tract, and used them to verify a person's identity. The company provides different products which include both text-dependent (i.e. Caller Authentication) and text-independent (i.e. Voice Track) solutions. |
| **MISTRAL** | http://mistral.univ-avignon.fr/ | ALIZE Library/Mistral | Open Source toolkit which provides an efficient and modular platform capable of managing different biometrics. Including Speaker ID modelling. |

**Table 1-6** Industry vendors for Speaker Recognition Systems

# 2 SPEECH PRODUCTION AND BIOMETRIC CHARACTERISATION OF VOICE

From the usage point of view, speech can be regarded as the vocalised form of human communication. From the biological point of view, like in any other important motor activity, speech communication requires the interaction of neurophysiologic systems, the motor system and the sensory system. A simplified description of speech communication can be as follows:

> "A speaker uses his/her brain to conceptualise the idea that he/she wants to communicate to a listener. Following this process, the brain translates these concepts into neurological processes and motor nerve commands to produce an acoustic sound pressure wave. This acoustic sound pressure wave, also called speech signal, results from the appropriate action combination of different muscles of the vocal organs. After its propagation through a transmission channel, the speech signal is perceived by the listener's auditory system. At this point, the auditory system translates the speech signal into nerve impulses that are transmitted to the listener's brain through the auditory nerve system. Finally the brain is supposed to satisfactorily reconstruct and hopefully understand the original idea."

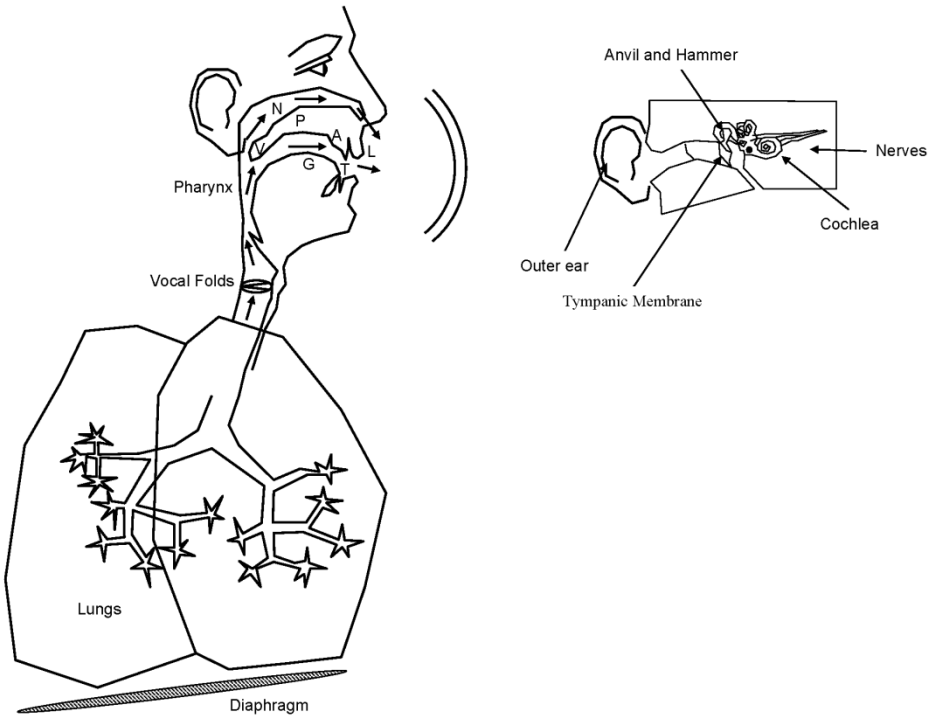



**Figure 2-1** A simplified diagram of speech communication (without taking into account brain interaction). V: Velum. G: Tongue. L: Lips, P: Hard Palate. A: Alveolus. N: Nasal Cavity. T: Teeth

Figure 2-1, provides a graphical representation of the above description. As far as speech production exists due to our ability to hear, or the other way around, this intrinsic connection between speech production and hearing is referred to as the *speech chain* by some authors [Denes,2012].

Although *speech chain* research encompasses a broad set of topics, in this chapter we will only focus on the motor system of speech production. Section 2.1, reviews the biological process of speech production at motor system level. Section 2.2 introduces the acoustic theory of speech production. Section 2.3 will describe the algorithm developed to separate glottal source and vocal tract components from speech signal, and finally section 2.4 will define the new biometric speaker characterisation from voice.

## 2.1 SPEECH PRODUCTION

As we have already highlighted, the motor system plays an important role in speech production, particularly, the speech tract which can be roughly divided in three different areas (infraglottic, glottal and supraglottic) depending on the location or functionality of the different organs and muscles. A full description of the motor system is beyond the scope of the present document. However, the interested reader can find a complete description for example in [Raphael,2006].

- Infraglottic cavities: Under this concept we refer to all the organs that are related somehow with the respiratory activity in the lower respiratory system: lungs, bronchi and trachea. When we speak, the lungs are responsible for providing the air flow needed to generate a sound, thanks to the contraction of the diaphragm.
- Glottal cavity: Again under this concept we refer to a set of muscles and cartilages that conform what is also known as the voice box or larynx. During expiration the air flow coming from the lungs passes by the vocal folds. The vocal folds, which are involved in the opening and closing of the glottis and in the production of sound, are a pair of muscles, epithelium, mucosal and connective tissues that stretch from the back of the larynx to the front (see Figure 2-1). At rest, in respiratory mode, the vocal folds remain separated (creating a V-shaped glottal space) to allow the air flow to pass directly from the lungs into the vocal apparatus, and the other way around (see Figure 2-2- right).

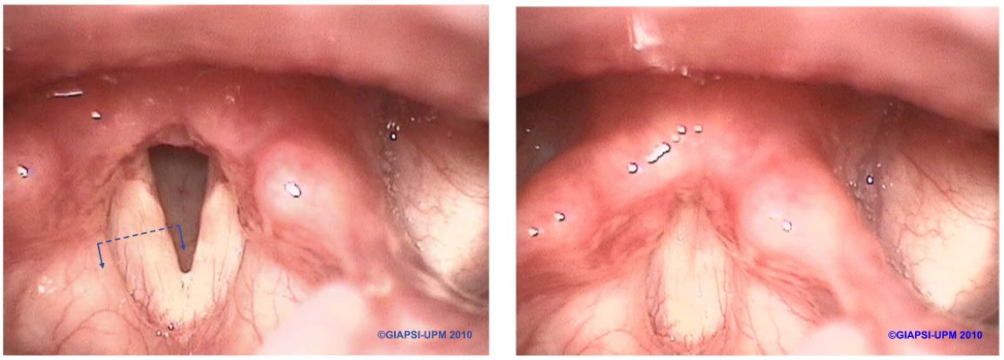



**Figure 2-2** Vocal Folds: Endoscopic view of vocal folds at inspiration (left-hand side) and in pre-phonation position (right-hand side)

In tense mode (or phonation mode), the two folds remain closed retaining the airflow from the lungs (see Figure 2-2- right). This situation causes subglottal pressure increase. When pressure exceeds the elastic force of the vocal folds, the air pressure pushes them apart creating an opening known as the glottis and allowing a small air flow to pass. As the air flow passes through the glottis the subglottal pressure decreases again leading to the closure of the glottis under the restoring forces of the muscles. These small and periodic releases of air flow are kwon as glottal pulses. The repetition of this process – opening and closure of

the vocal folds - leads to a periodic vibration of the vocal folds which produce sound waves that are transmitted to the supraglottic cavities.

The rate of the vocal fold vibration (phonation) is called the fundamental frequency, also known as F0 or pitch. Although different classifications of phonation can be performed, for the sake of simplicity and utility we are going to distinguish between two broad classes: voiced and unvoiced speech. Voiced speech has been already described as the one that is produced thanks to the vocal fold vibrations. In the case of unvoiced speech, there is no vibration, as vocal folds are open and air flow passes through the glottis without any constriction resulting in a turbulent noise-like behaviour. Instead, sound may be produced by the constrictions in the supraglottic cavities.

Moreover, for voiced speech, we can also establish and additional classification, distinguishing between vocalic and consonant sounds. In the production of vocalic sounds, vocal folds come together tightly, therefore vocal fold vibrations are stronger and frequency is high. Glottal aperture, under this constraint, is minimal and so is the air flow. In the case of voiced consonant sounds, vocal folds are less tightened, the glottis aperture is wider, and the air flow from infraglottic cavities to supraglottic cavities is larger.

- Supraglottic cavities: Under this concept we group the set of cavities, membranes and muscles that are situated above the glottal cavity. The supraglottic cavities are made up mainly by the pharynx cavity, until we reach the tongue and the uvula, at which point it is divided into the nasal cavity and the oral cavity (see Figure 2-1- left). While the vocal folds, roughly speaking, classify speech into voiced or unvoiced, supraglottic cavities modulate the sound wave to produce different categories of sounds/phonetic classes. In other words, these supraglottic cavities influence the manner of articulation.

  The different categories of speech in which voiced and unvoiced sounds are classified depend on the manner and places/points of articulation. These points of articulation correspond with the points of narrowest vocal tract constriction, i.e., the occlusions and narrowing on the vocal tract due to the action of mobile (velum, lower row teeth, lips, tongue, etc.) and fixed (pharynx wall, hard palate, upper row teeth, alveolar arch) organs/muscles. The manner of articulation is related with the airflow path, i.e., whether it flows through the nostrils (nasal sound) or through the lips (oral sound).

The characterisation of the vibration is heavily dependent on the mass and tension of the vocal folds, so a deeper review on the vocal fold characteristics and its behaviour is needed.

In voiced speech production, i.e. when vocal fold vibration exists, the joint effects of the subglottal and supraglottal air pressure difference, the laryngeal muscle tension and the elasticity of the vocal folds, causes an opening and closure of the vocal folds which produces a glottal flow. The glottal flow can be defined as the sound pressure pattern that is produced in the supraglottic cavities immediately after the vocal folds, which is related to the Liljencrants-Fant (L-F) excitation, in its ideal form [Fant,1985].

The physiological structure of vocal fold involved in the opening and closure of the glottis can be seen in Figure 2-4 (a), in which the massive part of the fold, composed mainly of muscles is referred to as “the body”, while the epithelial (lamina propria) and tissular envelope is referred to as “ the cover”. Thus the resultant vibration of the vocal

folds is composed of the coupled oscillations of the body and the cover. However, from a detailed analysis of video-endoscopic images it can be concluded that in modal phonation less than the third part of the vocal fold vibrates at full elongation in its transversal extension (i.e. following the dot-line direction in Figure 2-2- right) and only in its ventral area of its longitudinal extension. If we assume that the longitudinal extension size is 1.5 cm (average female case) and that the transversal extension size is 0.3 cm (Figure 2-3) then the static mass will be of about 0.106g. Given the above vibratory conditions, the involved dynamic mass will be of 0,011g.

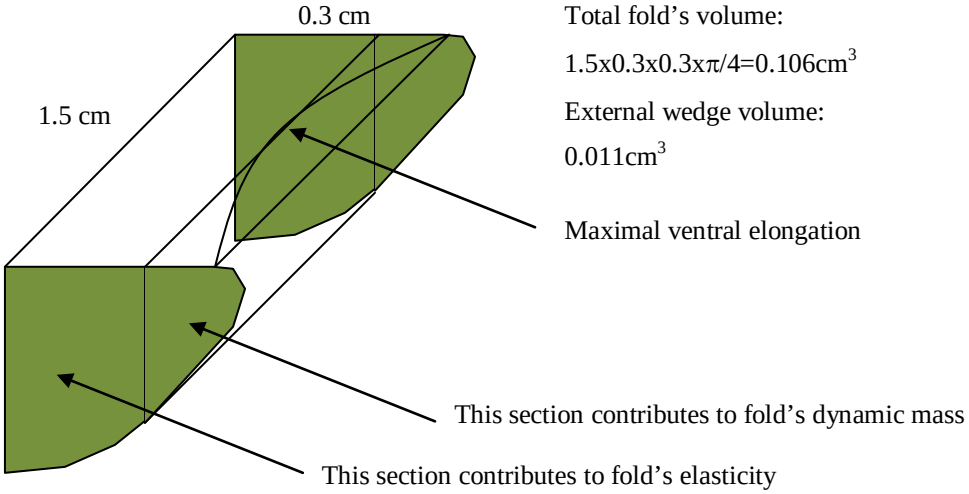



**Figure 2-3** Equivalent dynamic mass of the vocal fold. The section that mainly contributes to the vibration of the fold corresponds to the wedge on the right, corresponding to an inertial mass of approximately 11mg. The left-hand side section which corresponds mainly to the body of the fold, exhibits an elastic rather than inertial profile

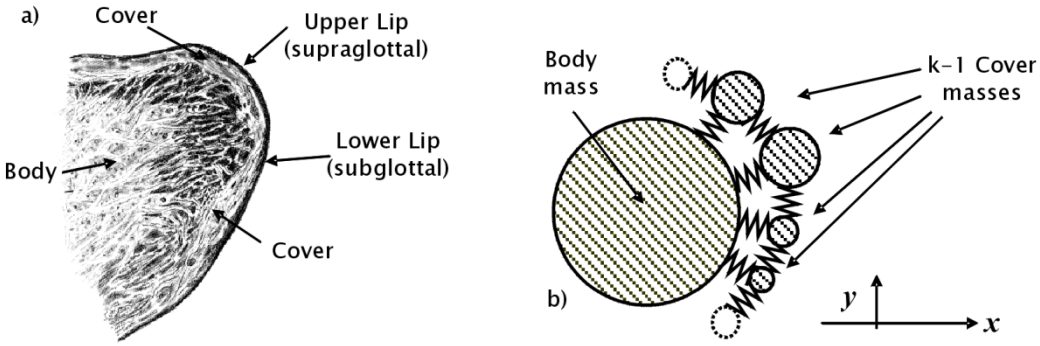



**Figure 2-4** Cross section of the left vocal fold: (a) body-cover structure, following the blue dash-line of the endoscopic view in Figure 2-2 right, (b) k-mass (body mass + k-1 cover masses) equivalent mechanical model.

The core of the body is slightly involved in the vibration, thus showing an elastic behaviour instead of providing dynamic mass. However most of the cover is heavily involved in the vibration, thus contributing to the vibration with a dynamic mass comparable to that provided by the body of the fold. This situation is depicted in Figure

2-4. Under this assumption, most of the structure of the vocal fold behaves like a spring, and the estimates of dynamic mass are considerably lower than the inertial mass.

Based on the body-cover model, Figure 2-4 (b) depicts the equivalent mechanical model of the vocal fold, as a set of lumped masses joined by elastic springs which represent the properties of the different tissue layers involved in the vocal fold vibration. Each mass will be linked to the body mass and to two neighbour ones, except the end masses. In general, the largest masses would be associated with the supraglottal and subglottal lips controlling the glottal closure.

Going back to video-endoscopic images we can distinguish three different phases in vocal fold vibration: a closing phase in which vocal folds start to get closer (adduction), a closed phase in which vocal folds block air flow, and an opening phase in which they start to move away (abduction). However, a deeper analysis of this vibration pattern reveals that the supraglottal and the subglottal lips do neither move nor close in phase, thus showing a phase lag between supraglottal and subglottal lips. This vibratory pattern, known as the "*mucosal wave*", as well as a complete phonation cycle sequence from glottal closure to glottal closure are depicted in Figure 2-5.

Due to the physiological characteristics of the vocal folds, we can characterise the vocal fold movement as a rhythmic approach-separation movement of the body plus an independent movement, thought subject to certain bond of the cover. This second movement is more complex as the different cover parts, and especially the subglottal and supraglottal lips move independently but also subject to certain bonds, as depicted in Figure 2-4 (b). The dynamic differential expression of the vibration pattern of the vocal folds is known as *mucosal wave*, as already mentioned. The *mucosal wave* is in essence a kind of a *travelling wave* propagating along the cover, and inducing nonlinear effects when both folds are close enough due to fold collision.

Assuming that the vocal folds were formed only by the body component, then the dynamic movement of the vocal folds could be approximated by a sinusoid as depicted in Figure 2-6. If we also assume the non-existence of the vocal tract, the glottal source would be a semi-sinusoidal pressure wave, which will reach a null value at the closing instant and an overpressure at the middle of the open phase. The inertial nature of the air column in the vocal tract would introduce a depression at the closing instants (known as the MFDR: maximum flow declination rate) and the glottal flow would present some phase shift with respect to the opening interval. Finally, if we assume movement independence in the cover components, then the glottal source pattern will follow the classical L-F model, whose ideal form is depicted in Figure 2-7.

Using this model (L-F excitation model), we can define the cover dynamic component, also known as mucosal wave correlate, as the observable glottal source component which is produced by a differentiated dynamic pattern of the different cover fold sections, especially the supraglottal and subglottal lips, during the glottal cycle [Story,1995], [Story,2002], [Titze,1988].

Although the k-mass equivalent mechanical model (see Figure 2-4) can be used to explain the effect of the mucosal wave ([Berry,2002], [Story,2002]), the least complex model which incorporates both cover and body dynamics of the vocal folds is the 3-mass model, also known as the Story-Titze model, described in [Story,1995] and depicted in Figure 2-8. For the interested reader a brief review of other models is also presented in [Story,1995], [Story,2002]

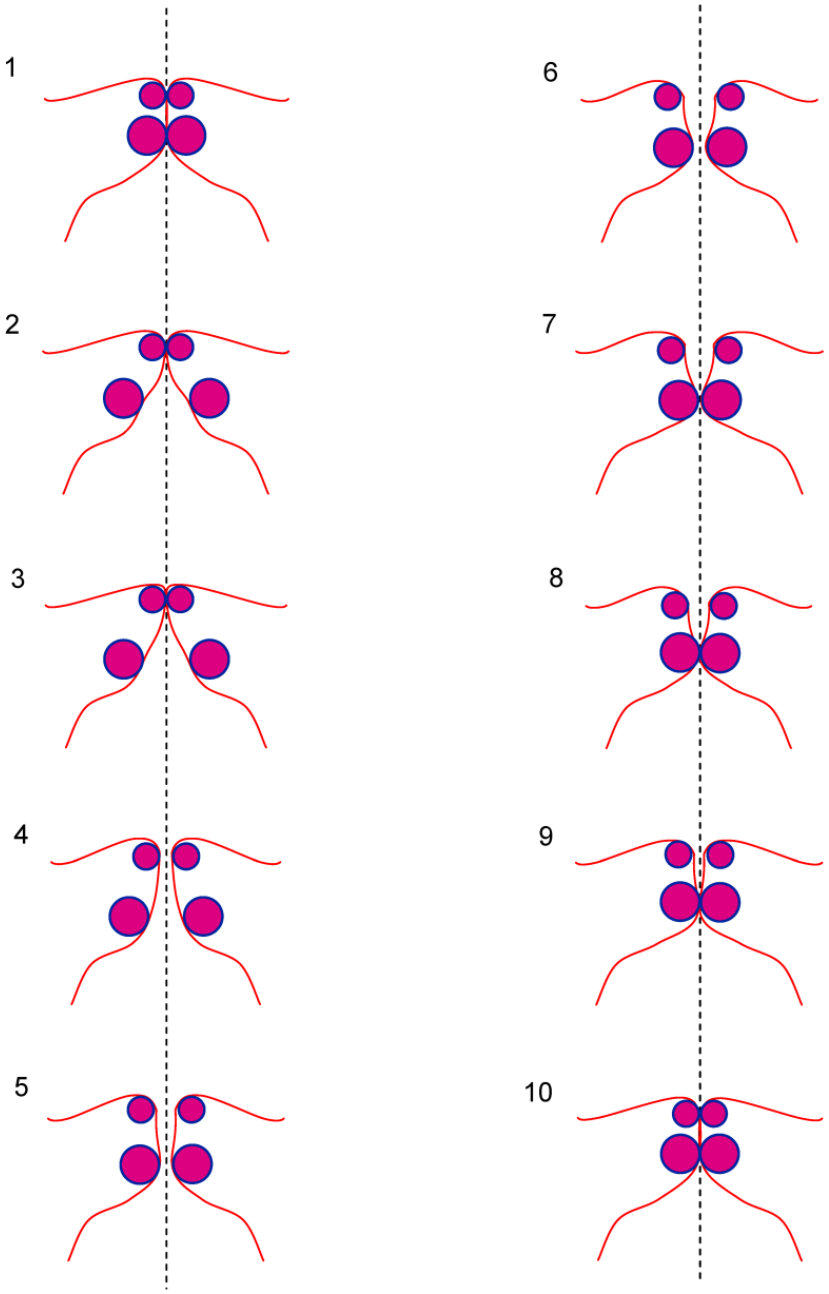



**Figure 2-5** Phonation cycle sequence from glottal closure to glottal closure (1-10), describing the Mucosal Wave. 1: Both subglottal and supraglottal lips are in contact, thus the glottal tract is closed. 2-3: Subglottal lips start to move away from each other, while supraglottal lips keep in contact, thus the glottal tract is still closed. 4-6: Subglottal lips get closer while supraglottal lips move away from each other, thus the glottal tract is opened from 4. 7-10: Subglottal lips are in contact, and supraglottal lips follow this closing tendency with certain phase lag. The glottis is closed from (7-10). As it can be seen, the mucosal wave dynamics also requires that a slight movement of subglottal and supraglottal lips in the vertical axis must be taken into account.

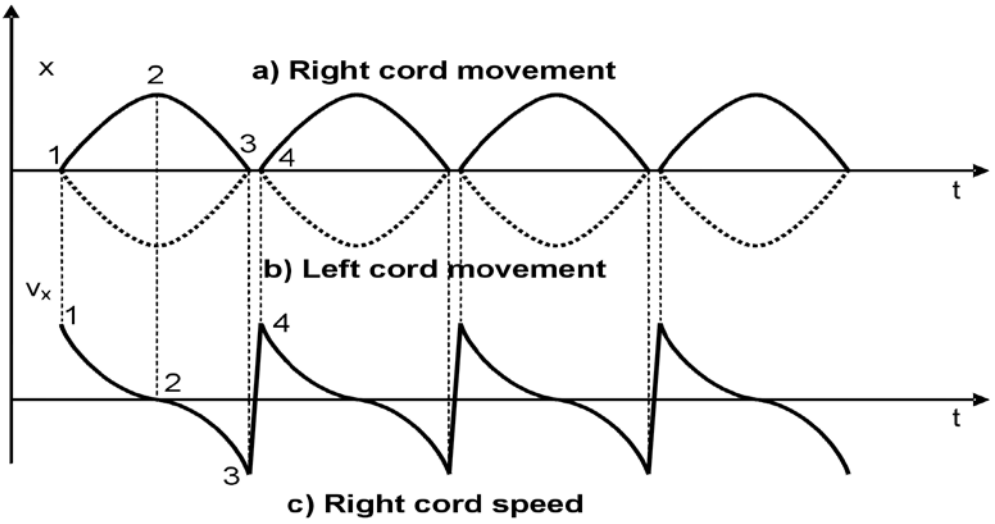



**Figure 2-6** Simplest vibration mode of the vocal folds. Top: Solid line represents the position of the right body mass, while dash line represents the position of the left body mass. Bottom: Right cord mass velocity. The vibration cycle can be divided in four stages: (1) the body masses start to move apart. (2) Maximum separation is reached, the relative velocity becomes null, and the fold tension inverts the movement (the velocity becomes the opposite – negative in the right cord and positive in the left cord). (3) Both masses come in close contact during a small fraction of time to move away again and start a new cycle (4).

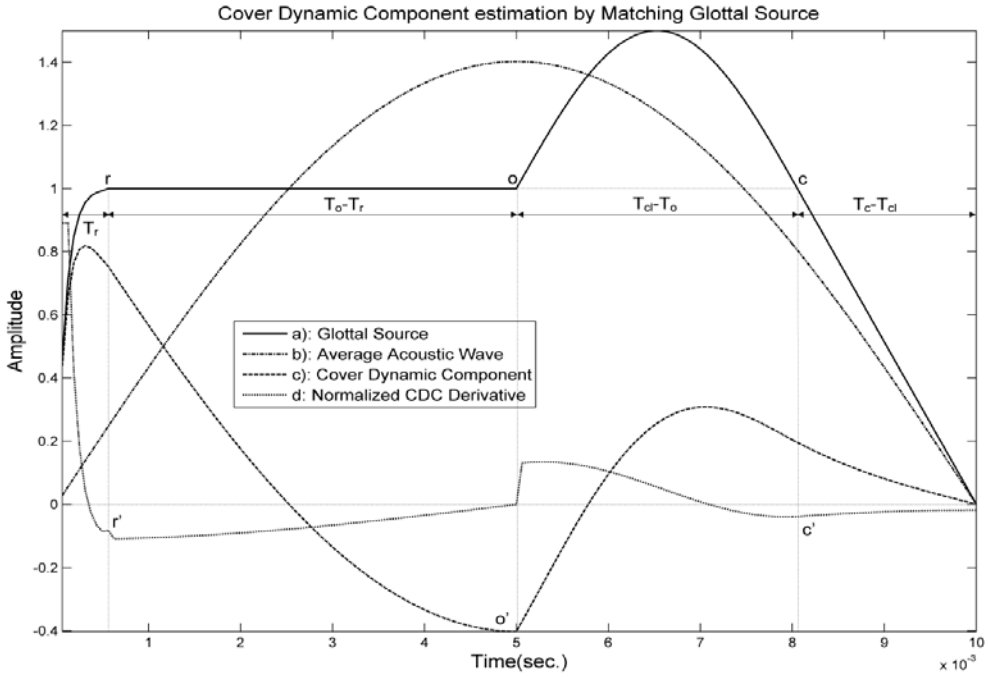



**Figure 2-7** Glottal source ideal model. a) Glottal source, in which we can highlight the following instants: t=0 → closure instant in which the minimum relative pressure is reached (allegedly p=0) ; $t=T_r$ → recovery from the minimum pressure achieved at closure time to the moment it reaches static or equilibrium atmospheric pressure (p=1, r-

point); t=To → opening starting point, when vocal folds start to move apart from each other (o-point); t=$T_{cl}$ → closure starting point, when vocal folds start its closing approximation (c-point). t=$T_c$ → closure instant, when pressure reaches the minimum normalised pressure (MFDR), starting a new phonation cycle. b) Average acoustic wave, according to [Titze,1994-B], corresponding to the resulting flow wave produced by the opening and closure of the vocal folds, characterised as two masses attached by springs to the walls of the larynx, and with no vocal tract present. c) Cover Dynamic Component (CDC), or mucosal wave correlate, which results from removing the average acoustic wave from the glottal source. Its minimum value at o-o' is used to establish $T_o$. d) Normalised CDC derivative. Its minimum values at r' and c' are used to establish $T_r$ and $T_{cl}$.

### 2.1.1 3-mass model of the vocal folds (body-cover structure)

The mucosal wave behaviour has been explained as a travelling wave [Berry,2001-B] propagating along the cover, having its most observable evidence in the phase lag between supraglottal and subglottal lips. In order to estimate the biomechanical parameters of the vocal folds, it is necessary to define a framework in which a relation between these parameters and the observable parameters (average acoustic wave and mucosal wave correlate - obtained from the speech signal) can be established.

As we have already said, the least complex model which incorporates the dynamics of both cover (whose correlate is the mucosal wave) and body (whose correlate is the average acoustic wave) is the 3-mass model (see Figure 2-8). From this model and following a *divide and conquer* approach, both correlates will be associated with a 1-mass model (for the body biomechanical parameter estimation) and a 2-mass model (for the cover biomechanical parameter estimation, see Figure 2-9)

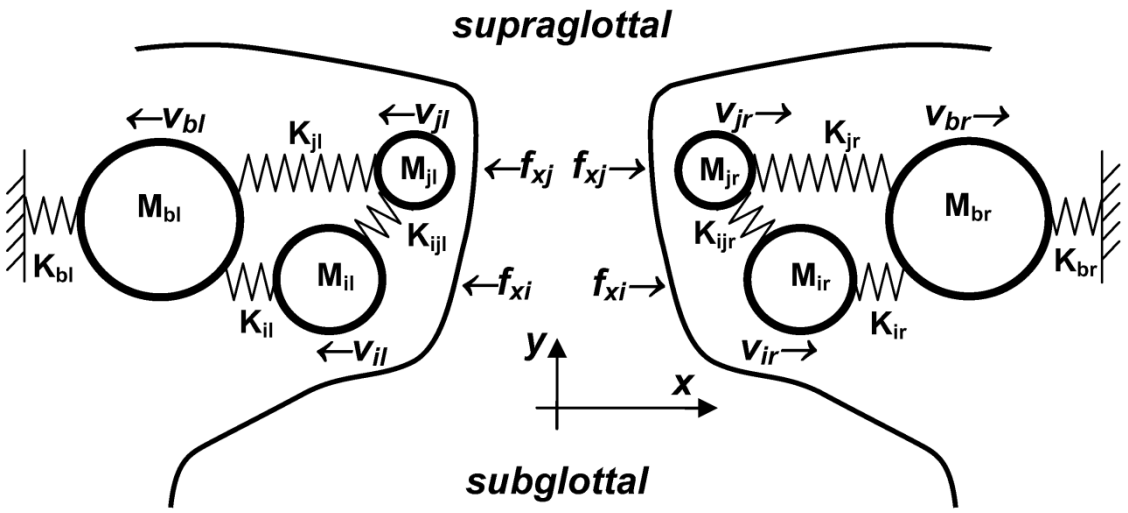



**Figure 2-8** Simplified version of the well-known Story-Titze model, which is agreed to give a more accurate explanation of the *mucosal wave phenomenon* than Ishizaka-Flanagan's 2-mass model ([Ishizaka,1972]). Schematic structure of one section of the body-cover modelled as a 3-mass system. The vocal fold body is represented by a bulky mass $M_{bl,r}$ (left and right respectively) which is concentrated in the vocal fold body's centre of mass. The vocal fold cover is represented by a pair of lumped masses $M_{il,r}$ and $M_{jl,r}$, linked together and to the body by elastic springs. Cover masses, $M_{il,r}$ and $M_{jl,r}$, allow to represent the independent movement of subglottal and supraglottal lips of the

cover. Under the adequate assumptions, this model may be reduced to the 2-mass equivalent model of the cover dynamics referenced to the body masses.

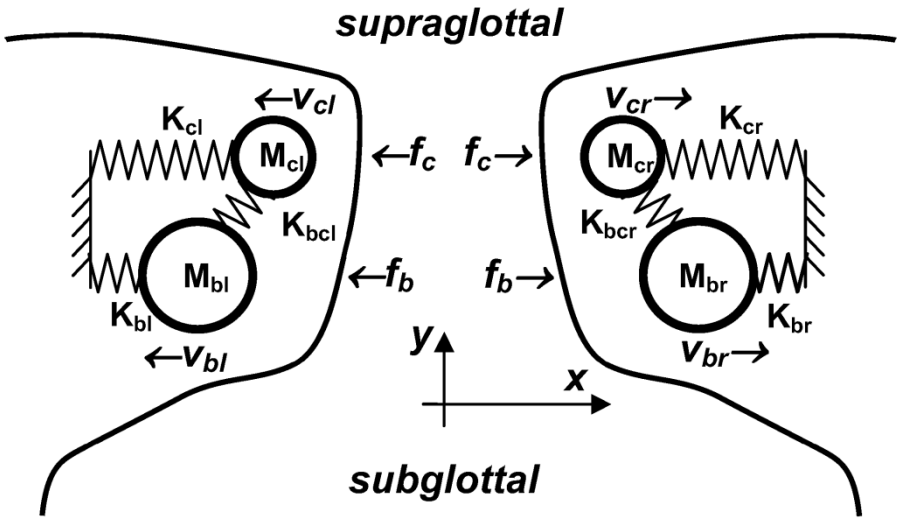



**Figure 2-9** 2-mass and 3 spring by vocal fold model. Schematic structure of one-section of the right (r) and left (l) vocal folds. Supraglottal and subglottal lips refer to the segment of the larynx that is closer to the lips and to the lungs respectively. The left supraglottal lip is represented by a mass, $M_{cl}$, tied to larynx walls by the spring $K_{cl}$, whereas the subglottal lip is represented by another mass $M_{bl}$ tied to larynx walls by a spring, $K_{bl}$. Both masses are tied to each other by another inter-spring $K_{bcl}$. The right vocal fold structure is completely analogous to the left one. $f_c$ and $f_b$ forces influence symmetrically on both vocal folds, resulting in a mass movement against the springs, so that the dynamic variables of each mass correspond to the linear velocity in the x-axis ($v_{bl}$, $v_{cl}$, $v_{cr}$, $v_{br}$).

In the attempt to develop a biomechanical model of the vocal folds that simulate their dynamics, Flanagan's 2-mass model [Ishizaka,1972] was capable of reproducing many features of phonation and has been successfully used for speech synthesis. Since its introduction, different refinements have been applied adding complexity through the use of finite methods [Alipour,2000], [Berry,2002], [Titze,1973], [Titze,1974]. All these methods have been successfully applied to voice production, speech processing, clinical studies, etc. Moreover, as we have already said and presented, the k-mass model also can be used to describe mucosal wave effects [Story,2002], [Berry,2002]. However, for the sake of simplicity only the 3-mass, 2-mass and 1-mass models will be presented in the present study.

In Figure 2-8 the vocal fold body is represented by a bulky mass $M_{bl,r}$ (left and right respectively) which is concentrated in the vocal fold body's mass centre. The vocal fold cover is represented by a pair of lumped masses $M_{il,r}$ and $M_{jl,r}$, linked together ($K_{ijl,r}$), to the body ($K_{il,r}$ and $K_{jl,r}$), and to the reference ($K_{bl,r}$) by elastic springs. Cover masses, $M_{il,r}$ and $M_{jl,r}$, allow to represent the independent movement of subglottal and supraglottal lips of the cover. The influence of viscous and inelastic losses are taken into account in the model by parameters ($R_{bl,r}$, $R_{jl,r}$, $R_{il,r}$, $R_{ijl,r}$). However, in order to reduce the complexity of this model, some assumptions have been done:

- Each vocal fold (of complex 3-dimensional structure) is represented by the vibration of the specific masses only in the x-axis direction.

- Each specific model mass represents the dynamic interaction between forces and acceleration that only affect a fraction of equivalent 3D structure (as depicted in Figure 2-3).
- As the result of the action of the forces in the x-axis, the masses move only in the direction of this axis.
- Biomechanical parameters are time-invariant in contrast with the studied interval.

Under these assumptions, the resulting model (Figure 2-8) is the well-known Story-Titze model, described in [Titze,1974]. This model is the less complex model which incorporates a good description of the cover and body dynamics, representing the mucosal wave while retaining the simplicity of a low order model.

Cover masses, $M_{il,r}$ and $M_{jl,r}$, allow to represent the independent movement of subglottal and supraglottal lips of the cover. The dynamic variables, represented by $f_{xi,j}$ are the forces acting on each cover masses as the result of the pressure difference between both regions, and the associated velocities to the specific masses: $v_{bl.r}$, $v_{il,r}$ and $v_{jl,r}$. The fact of maintaining a difference between right and left folds relies on the possibility of finding left/right vocal fold differences even in normophonic speakers.

The behaviour of the vocal fold can be characterised by the estimate resulting from removing the effect of the vocal tract [Alku,1992] on the speech wave that is also known as glottal residual correlate. The first and second integrals of this correlate represent the glottal source and the glottal flow, as reported in [Gómez,2005], [Gómez,2004]. The influence of the cover and body dynamics on the glottal source is reflected in the average acoustic wave and in the mucosal wave correlate respectively.

Following a *divide and conquer* approach, the mucosal wave dynamics can be characterised by the movement of the two cover masses ($Mj_{l,r,}$ $Mi_{l,r}$) with respect to the associated body mass ($M_{bl,r}$).

### 2.1.2 2-mass model of the cover structure of the vocal folds

As we are able to distinguish between the body and the cover dynamics, we can use the Ishizaka-Flanagan's 2-mass model [Ishizaka,1972] to characterise the later one. Thus, the dynamics of this system can be characterised by the following four integro-differential equations:

$$f_{il,r} = M_{il,r}\frac{\partial v_{il,r}}{\partial t} + R_{il,r}v_{il,r} + K_{il,r}\int_{-\infty}^{t} v_{il,r}d\zeta + K_{ijl,r}\int_{-\infty}^{t}\left(v_{il,r} - v_{jl,r}\right)d\zeta$$

$$f_{jl,r} = M_{jl,r}\frac{\partial v_{jl,r}}{\partial t} + R_{jl,r}v_{jl,r} + K_{jl,r}\int_{-\infty}^{t} v_{jl,r}d\zeta + K_{ijl,r}\int_{-\infty}^{t}\left(v_{jl,r} - v_{il,r}\right)d\zeta \qquad \textbf{Eq. (2-1)}$$

If we associate forces with electric potentials and velocities with currents, then the equivalent electromechanical model will be the one depicted in Figure 2-10.

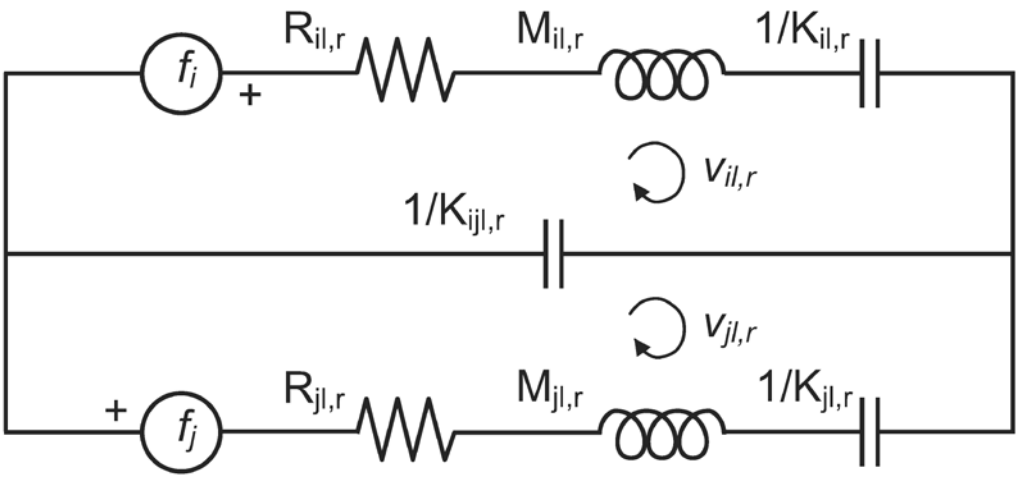



**Figure 2-10** Equivalent electromechanical model of the 2-mass and 3-spring model of the vocal folds depicted in Figure 2-9. Additionally to the mass and elasticity springs represented by $M_{il,r}$, $M_{jl,r}$, $K_{il,r}$, $K_{jl,r}$, two resistor elements ($R_{il,r}$, $R_{jl,r}$) have been added which represent the elastic and viscous losses in the behaviour of the system.

The validity of the differential analysis which results from the independent characterisation of the body and cover dynamics, can be checked by the relation that can be established between the PSD profile of the mucosal wave correlate and the 2-mass model transfer function of the cover as it is described in [Gómez,2005], [Gómez,2004].

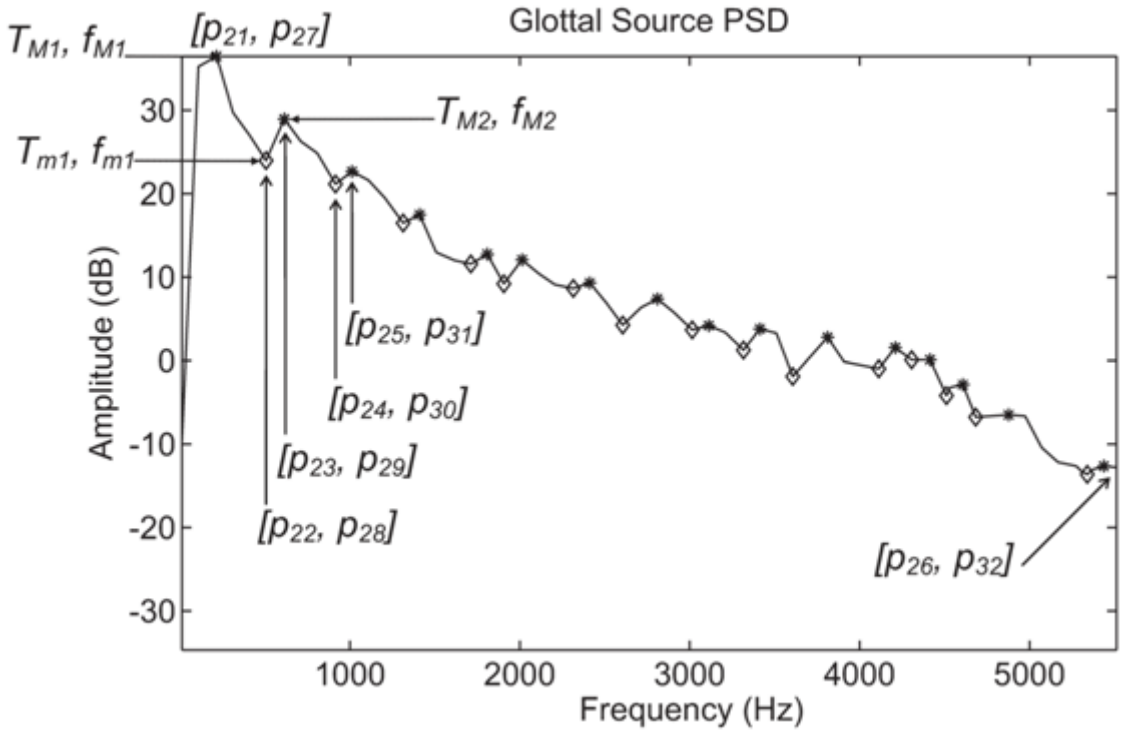



**Figure 2-11** Power spectral density (PSD) of male voice segment synchronously evaluated in a phonation cycle, which match the harmonic envelope or the PSD profile. The following singular points have been highlighted: $p_{21} \rightarrow$ maximum PSD value in dB scale; $p_{22} \rightarrow$ first minimum value related to the first maximum in dB scale; $p_{23} \rightarrow$ second PSD maximum value related to the first maximum in dB scale; $p_{24} \rightarrow$ second

PSD minimum value related to the first maximum in dB scale; $p_{25} \rightarrow$ third PSD maximum value related to the first maximum in dB scale; $p_{26} \rightarrow$ PSD value at the maximum Nyquist value relative to the first maximum in dB scale; $p_{27} \rightarrow$ relative position in frequency of the first maximum; $p_{28} \rightarrow$ relative position in frequency of the first minimum; $p_{29} \rightarrow$ relative position in frequency of the second maximum; $p_{30} \rightarrow$ relative position in frequency of the second minimum; $p_{31} \rightarrow$ relative position in frequency at the third maximum; $p_{32} \rightarrow$ relative position at the end for Nyquist frequency related to the first maximum.

Figure 2-11 provides a representation of a male voice PSD profile whereas Figure 2-12 depicts the supraglottal-subglottal trans-admittance transfer function response (i.e. the relation between the supraglottal mass velocity and the subglottal force in frequency domain). If we contrast Figure 2-11 with Figure 2-12, the depth of the first V-profile of the PSD profile is also present in the trans-admittance transfer function response and can be explained through the valley that appears in the trans-admittance modulus.

By tuning the values of the biomechanical parameters $\{M_i, M_j, R_i, R_j, K_i, K_j$ y $K_{ij}\}$ of the symmetric 2-mass model, the trans-admittance clearly matches the main characteristics of the PSD profile of the mucosal wave correlate: a rapid rise from low frequencies to a first amplitude maximum given by $T_{M1}$ centred at a frequency $f_{M1}$, followed by a minimum $T_{m1}$ at $f_{m1}$ and a new rise to a second amplitude maximum $T_{M2}$ at $f_{M2}$. It can be shown [Berry,2001-A] that this V-profile is the result of the interaction between the 2 cover masses lumped by the spring $K_{ij}$. This "V" profile may appear several times more, while the envelope of the curve shows a decay of *1/f*. The conclusions reached so far were empirically shown by Švec et al., [Svec,2000], in the experimental measures taken over real vocal folds, where the spectral density of the movement in a supraglottal edge point of both vocal folds is measured, showing the same V-profiles.

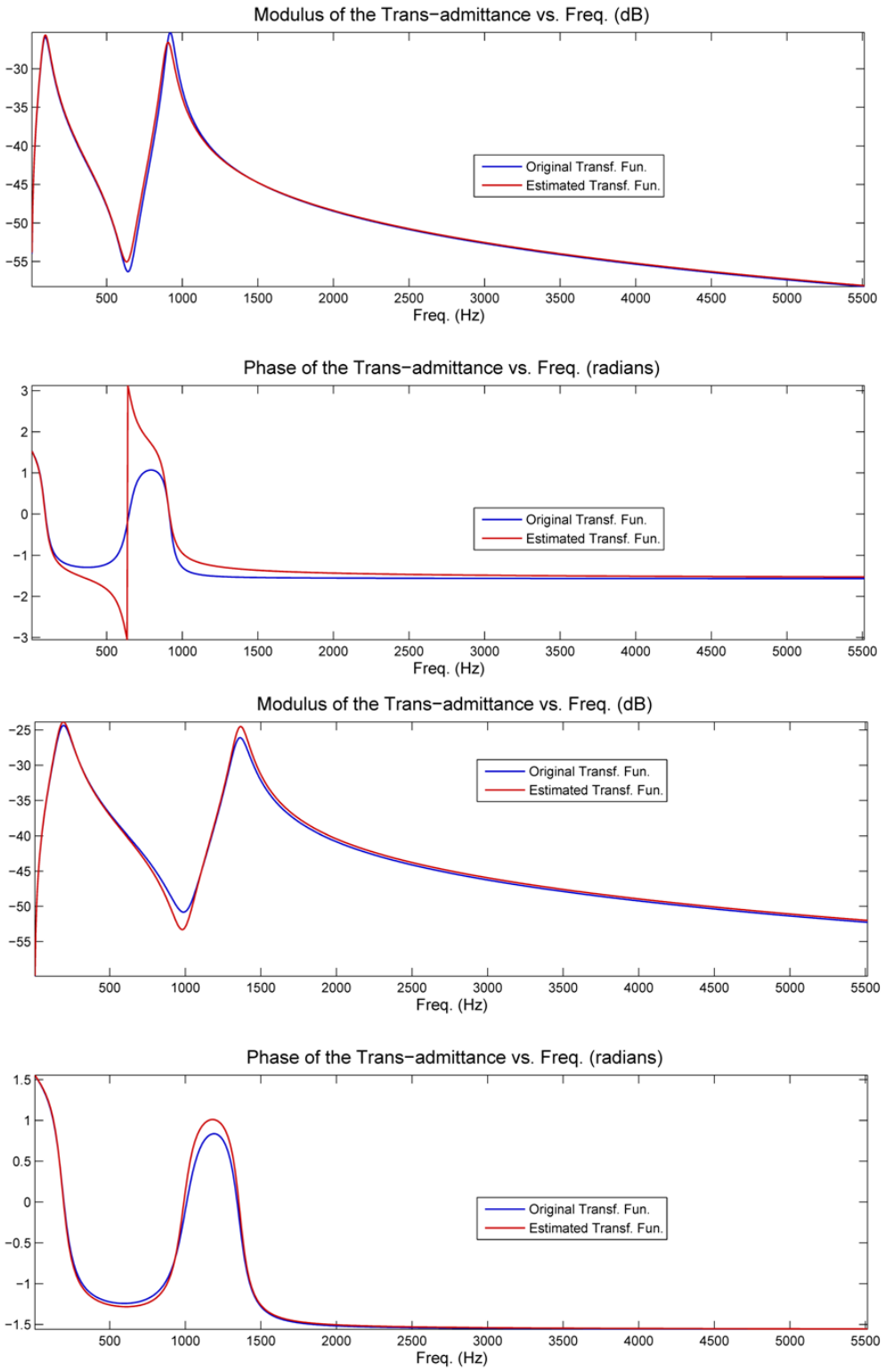



**Figure 2-12** Modulus and phase frequency response of the trans-admittance between the force that affects one mass ant the velocity that appears in the opposite mass. The first 2 plots represent a male voice prototype, while the last 2 plots represent a female voice prototype. Blue lines corresponds to the behaviour of the trans-admittance function from a prefixed circuit elements ($M_{il,r}$, $Ml,r_j$, $R_{il,r}$, $R_{jl,r}$, $K_{il,r}$, $K_{jl,r}$ y $K_{ijl,r}$), while red lines represents the behaviour with the elements that results from inverting the system and following the biomechanical parameterisation method.

### 2.1.3 1-mass model of the body structure of the vocal folds

As previously said, the 1-mass model of the body structure will determine the behaviour of the average acoustic wave. This model consists of two masses, $M_{bl,r}$, which represent the left and right fold bodies. Each of these masses is joined to the ideal walls of the system by a spring, $K_{bl,r}$, while $R_{bl,r}$ represents the influence of the viscous losses. The dynamic behaviour of this system can be characterised by the following integro-diferential equations:

$$f_{bl,r} = M_{bl,r}\frac{\partial v_{bl,r}}{\partial t} + R_{bl,r}v_{bl,r} + K_{bl,r}\int_{-\infty}^{t} v_{bl,r}d\zeta \qquad \textbf{Eq. (2-2)}$$

The equivalent masses of each vocal fold experiment a movement in response to the forces, $f_{bl,r}$ that result from the pressure difference between subglottal and supraglottal regions. Vibratory movement along x-axis will be described by the position of each mass $M_{bl,r}$, considered as a point mass. In absence of the vocal tract, the trajectory (or vibratory movement) described by the masses follows the description depicted in Figure 2-6, which consists of semi-sinusoidal arches that reproduce separation, elastic retention, approach and collision events (the degree of inelasticity depends on losses incurred). Assuming the linearity of the system between collisions, the frequency of the arches can be express as:

$$\omega_{bl} = \sqrt{\frac{K_{bl}}{M_{bl}}}; \omega_{br} = \sqrt{\frac{K_{br}}{M_{br}}} \qquad \textbf{Eq. (2-3)}$$

Moreover, Eq. (2-3) must be taken as a rough estimation in order to validate the body-cover decomposition approach, as we are considering that the body of the vocal folds vibrates without cover influence.

## 2.2 ACCOUSTIC THEORY OF SPEECH PRODUCTION

As described in previous sections and following the Gunnar Fant's [Fant,1985] production model (see Figure 2-13), the voiced speech production process can be described as the result of the sound pressure wave generated by the vibration of the vocal folds flowing through the laryngeal and pharyngeal cavities and finally radiated through the lips and/or nostrils (vocal tract). However, as we have already established, humans can also generate unvoiced speech (for instance whispered speech), in which there is no vibration on the vocal folds. In order to deal with this situation, Fant's production model assumes the existence of two different excitation sources *e(n)*: Glottal Excitation and Turbulent Excitation. These excitation sources can be applied, either alternately or in combination, to the phonation model (laryngeal, pharyngeal, nasal and oral cavities) and then radiated, resulting in a signal, $s_r(n)$, that can be heard by the ear or captured with a microphone.

The sound pressure pattern that occurs in the supraglottic cavities immediately after the vocal folds (see Figure 2-7) is what we have called glottal source, voice glottal correlate or Liljencrants-Fant excitation in its ideal form. Meanwhile, the vocal tract, which includes the speech production organs above the vocal folds, can be regarded as a filter that alters the frequency content of the glottal source due to its resonances (also known as formants: energy amplification) and antiresonances (energy attenuation). This circumstance allows the estimation of the vocal tract shape from the spectral shape of the voice signal ([Campbell,1997]).

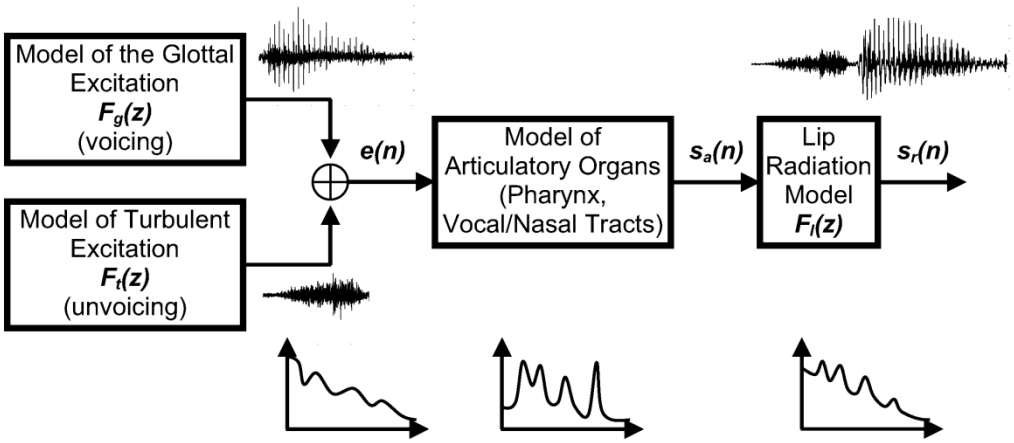



**Figure 2-13** Gunnar Fant's production model. An excitation source *e(n)*, either a glottal excitation for voiced voice or a turbulent excitation for unvoiced voice, feeds the phonetic and articulatory organs, represented by an all-pole transfer function. The resulting active signal, $s_a(n)$, is finally radiated to become a signal, $s_r$(n) that can be captured either by the ear or by a microphone.

According to [Fant,1971], the voice production process can be acoustically simulated by a source-filter model (source excitation and filter modulator). In other words, voiced speech may be seen as the output of a generation model, $F_g(z)$ (where $f_g$ is the glottal impulse response), excited by a train of delta pulses, its output spectrally conditioned by the vocal tract with transfer function given by $F_{vt}(z)$ (where $f_{vt}$ is the vocal tract impulse response) to produce the speech signal before radiation $s_l(n)$ and after radiation $s(n)$, where $F_l(z)$ represents the lip radiation model, and $r=R^{-1}(z)$ represents the inverse lip radiation model.

$$s = \left\{\left\{\delta * f_g\right\} * f_{vt}\right\} * r = \left\{f_g * f_{vt}\right\} * r = s_l * r \qquad \textbf{Eq. (2-4)}$$

In this simplified approach we assume that the vocal tract and the glottal excitation signal have separated acoustic effects.

If we now focus the attention on the vocal tract, it can be roughly modelled with a series of acoustic tubes with specific cross-section areas [Fant,1971]. As it can be seen in (see Figure 2-14), the first acoustic tube is supposed to start at the glottis while the last one represents either the lips or nostrils. The vocal tract length of adult men is typically of 17cm. while that of adult women is 14cm. and in children is of 10cm. Regarding the diameter and length of each acoustic tube, this two characteristics are related with the sound produced (i.e. the manner and place of articulation). Obviously, the acoustic configuration of the vocal tract also determines the resonances and antiresonances of the tubes. Moreover, these resonances and antiresonances can be characterised by the poles and zeros of a digital filter, although for the sake of simplicity an all-pole model (with transfer function described in Eq. (2-5)) can also be used. In this last case, the poles of the digital filter define the formants of the speech wave.

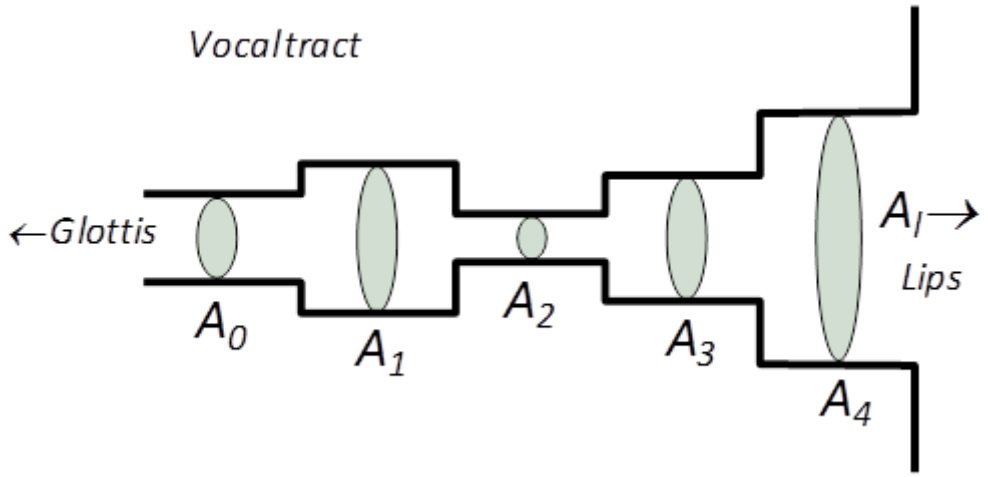



**Figure 2-14** Acoustic tube model of speech production ([Campbell,1997])

$$F_{vt}(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \qquad \textbf{Eq. (2-5)}$$

Since its practical introduction at the late sixties by F. Itakura and S. Saito [Itakura,1970], for modelling the vocal tract by inverse filtering of speech signal, Linear Prediction has become the *de facto* method for this purpose. Although a deep review on Linear Prediction can be found in [Makhoul,1975], a general review of the relevant aspects involved in the inverse filtering performed for the estimation of the glottal source form the voice signal, will be done later.

Finally, the radiation effect due to lips/open air interaction can be acoustically modelled by a first-order differentiation of the volume velocity at the lips. This means that in order to compensate this radiation effect on the voice signal, a first order integrating filter can be used.

## 2.3 SOURCE-TRACT SEPARATION OF THE SPEECH SIGNAL

As we have already established, both glottal source and vocal tract systems are involved in speech production processes. It may be expected that glottal information will be more influenced by the speaker's phonation habits, while the description of the vocal tract will be more conditioned by the phonetic structure of the message. On its turn the power spectral density of the glottal source is strongly conditioned by the biomechanics of the vocal folds. Thus, both vocal tract and glottal information seem to be relevant when characterizing a speaker. However, in most real-life scenarios (regardless medical scenarios), we can only observe the speech signal, which as we have already established is the convolution reflected in Eq. (2-4).

In order to study the influence of vocal tract and glottal information in speaker recognition applications, it would be useful to break down the speech signal in a glottal source estimate and a vocal tract estimate. In what follows a description of the algorithm applied to source-tract separation will be exposed, preceded by a review on linear prediction.

### 2.3.1 Linear Prediction review

Linear Prediction (LP) of a temporal signal is a well-known and extensively used signal processing technique in speech signal analysis, but having wide application in other areas. Although most of signal processing books devote a section to linear prediction

theory, in [Kailath,1974] the interested reader may find a remarkable review of the history of linear estimation traced back to its very roots.

Since a deep review of LP is beyond the scope of this thesis, in what follows, only the general aspects of LP closely related to inverse filtering processes required to estimate the glottal source from the voice signal will be highlighted.

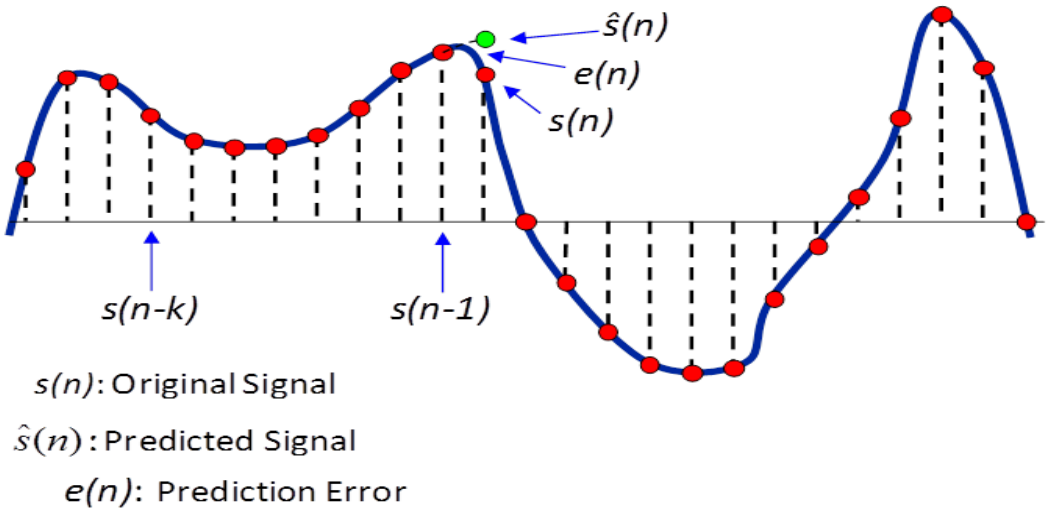



**Figure 2-15** Linear Prediction over a discrete signal which results from sampling a continuous signal at regular intervals

The linear prediction problem (as depicted in Figure 2-15) can be defined as the estimation of future values of a discrete-time signal from a linear combination of previous samples. Or more specifically:

- Let $s(n) \in \mathbb{R}$ be a set of real samples ordered in time domain according to an integer variable $n \in \mathbb{Z}$, obtained as a result of sampling a continuous signal, *s(t)*, at intervals $t=n\tau$, where $t, \tau \in \mathbb{R}$ and $\tau$ is the sampling period. Let $S_K = \{s(n-i); 1 \leq \mathrm{i} \leq K\}$ a set of K consecutive samples defined over the temporal window $\{n-i; 1 \leq \mathrm{i} \leq K\}$. The aim of linear prediction is to estimate the next sample, s(n), outside the temporal window $\{n-i; 1 \leq \mathrm{i} \leq K\}$, as a linear combination of $S_K$ samples and the predictor coefficients $A_K = \{a_i; 1 \leq \mathrm{i} \leq K\}$:

$$\hat{s}(n) = a_1 s(n-1) + a_2 s(n-2) + \cdots + a_K s(n-K) = \sum_{i=1}^{K} a_i s(n-i) \qquad \textbf{Eq. (2-6)}$$

  where $a_i \in \mathbb{R}, 1 \leq i \leq K$ are the optimum set of coefficients (optimum predictor) evaluated as the ones that minimise the mean square error between the prediction and the actual value of *s(n)*:

$$e_K(n) = s(n) - \hat{s}(n) = \sum_{i=0}^{K} h_i s(n-i)\,; \qquad \textbf{Eq. (2-7)}$$

$$h_0 = 1;\ h_i = -a_i;\ 1 \leq i \leq K$$

  evaluated over the temporal window W=$\{n, N_1 \leq n \leq N_2\}$:

$$L_K = \sum_w e_K^2 \qquad \textbf{Eq. (2-8)}$$

NOTE: The prediction error can be regarded as the output of a FIR (Finite Impulse Response) filter, $A_K(z) = \sum_{i=0}^{K} h_i z^{-i}; h_0 = 1;\ h_i = -a_i; 1 \le i \le K$ in response to the input, *s*(*n*). Therefore, the IIR (Infinite Impulse Response) filter 1/ $A_K(z)$ can be used to reconstruct the original signal from the error signal.

Consequently, LP in essence transforms the signal *s(n)* into a set of *K* numbers $\{a_i;\ 1 \le i \le K\}$ and an error signal $e_K(n)$. For large *K*, the spectral information of *s(n)* is mostly contained in the predictor coefficients.

It can be shown that the function $L_K$ has a minimum in the complementary vector space defined by the weights of the linear combination $\{a_i;\ 1 \le i \le K\}$; therefore in the minimisation process the following condition must be fulfilled:

$$\frac{\partial L_K}{\partial a_j} = 0; 1 \le j \le K \qquad \textbf{Eq. (2-9)}$$

which is equivalent to:

$$\frac{\partial L_K}{\partial a_j} = \frac{\partial}{\partial a_j}\left[\sum_W e_K^2(n)\right] = \sum_W \frac{\partial}{\partial a_j}\left[s(n) - \sum_{i=1}^{K} a_i s(n-i)\right]^2 = -2\sum_W s(n-j)\left[s(n) - \sum_{i=1}^{K} a_i s(n-i)\right] = 0; 1 \le j \le K \qquad \textbf{Eq. (2-10)}$$

and can be reformulated as:

$$\sum_W s(n-j)s(n) = \sum_{i=1}^{K} a_i \sum_W s(n-j)s(n-i) \qquad \textbf{Eq. (2-11)}$$

Eq. (2-11) are also known as the Normal Equations (*Yule-Walker equations or Wiener-Hopf equations*). If prediction coefficients, $\{a_i;\ 1 \le i \le K\}$, have been estimated through the Normal Equations, then the estimation error, *e(n)*, is orthogonal to the extended sample vector $S_{k+1}=\{S_k, s(n)\}$ on the temporal interval considered:

$$\sum_W s(n-j)e(n) = 0; 1 \le j \le K \qquad \textbf{Eq. (2-12)}$$

Moreover, if we define $c_{ij}$, as

$$c_{ij} = \sum_W s(n-j)s(n-i) \qquad \textbf{Eq. (2-13)}$$

then, we can reformulate the Normal Equations as:

$$\sum_{i=1}^{K} a_i c_{ij} = c_{j0}; 1 \le i \le K; 1 \le j \le K \qquad \textbf{Eq. (2-14)}$$

$$Ca = q \qquad \textbf{Eq. (2-15)}$$

This reformulation allows for an intuitive solution though the use of matrix inversion (covariance method): $a=C^{-1}q$

Since its original formulation, different algorithms have been proposed for solving the Normal Equations for $\{a_i;\ 1 \le i \le K\}$. An alternative method to the covariance one, is the autocorrelation method, in which in order to simplify $c_{ij}$, the assumption $s(n)=0$ outside the temporal window under analysis is made.

$$c_{ij} = \sum_{W} s(n-j)s(n-i) = \sum_{W} s(n)s(n+|i-j|) = r_{ij} \qquad \textbf{Eq. (2-16)}$$

In this way, $c_{ij} = c_{ji} = r_{ij}$, and $r_{ij}$ receive the name of autocorrelation coefficients of the autocorrelation matrix. The autocorrelation matrix has the form of a Toeplitz matrix, which is symmetrical and has the same values along the lines parallel to the diagonal. According to this approach, Eq. (2-14) can be expressed by matrix representation:

$$\begin{bmatrix} r_0 & r_1 & r_2 & \dots & r_{K-1} \\ r_1 & r_0 & r_1 & \dots & r_{K-2} \\ r_2 & r_1 & r_0 & \cdots & r_{K-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{K-1} & r_{K-2} & r_{K-3} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_K \end{bmatrix} \qquad \textbf{Eq. (2-17)}$$

where

$$r_\tau = \sum_{W} s(n)s(n+\tau) = \sum_{W} s(n)s(n+|i-j|) = c_{ij} = c_{ji};\ \forall \tau = |i-j| \qquad \textbf{Eq. (2-18)}$$

NOTE: Although the covariance and correlation methods give almost the same results when the original signal is large enough and stationary, their results differ when $s(n)$ is short and weakly stationary or quasi-stationary.

The equations expressed in Eq. (2-16) can be efficiently solved using the Levinson-Durbin recursion (originally proposed by Norman Levinson and improved later by J. Durbin [Durbin,1960]). This method states that a solution to the Normal Equations, Eq. (2-11), expressed by matrix representation in Eq. (2-17) can be achieved using the following recursion:

- Initialisation of the mean quadratic value of the fragment of the signal under study:

$$L_0 = r_0 \qquad \textbf{Eq. (2-19)}$$

- Loop $1 \le j \le K$:

  Where $a_i^j$, represents the estimation of $i^{th}$ coefficient at $j^{th}$ iteration, and $L_j$, represents the mean squared error produced at iteration $j$.

  - Evaluation of the $j^{th}$ reflection coefficient (also known as PARtial CORrelation – PARCOR – Coefficient).

$$c_j = \frac{1}{L_{j-1}}\left[r_j - \sum_{i=1}^{j-1} a_i^{j-1} r_{j-i}\right] \qquad \textbf{Eq. (2-20)}$$

- Evaluation of the prediction coefficients of $j^{th}$ order.

$$a_j^j = c_j;$$
$$a_i^j = a_i^{j-1} - c_j a_{j-i}^{j-1} \qquad \textbf{Eq. (2-21)}$$

- Estimation of mean square error obtained at $j^{th}$ iteration.

$$L_j = L_{j-1}\left(1 - c_j^2\right) \qquad \textbf{Eq. (2-22)}$$

- If $j \neq K \rightarrow$ Go to 1

- Finally:

$$a_i = a_i^K; a \leq i \leq K \qquad \textbf{Eq. (2-23)}$$

Levinson-Durbin recursion can be reformulated in order to evaluate the reflection coefficients from the statistical correlation between the forward prediction error, $f_k(n)$, and the backward prediction error, $b_k(n)$. This last one can be viewed as the error incurred in predicting $s(n\text{-}K\text{-}1)$, the first sample out of the prediction window set $S_K$, looking backwards with a coefficient set $BK = \{b_i;\ 1 \leq i \leq K\}$. In order to guarantee the stability of the solution, [Itakura,1970] proposed the used of the geometric mean between the forward prediction coefficients and the backward prediction coefficients, leading to the following expression to evaluate the reflection coefficients from which the prediction coefficients can be obtained:

$$c_j = \frac{\sum_n f_{j-1}(n)\, b_{j-1}(n-1)}{\sqrt{\sum_n f_{j-1}^2(n) \sum_n b_{j-1}^2(n-1)}} \qquad \textbf{Eq. (2-24)}$$

This coefficient, also known as PARCOR coefficient, can be regarded as the cosine of the angle formed by the forward and backward prediction errors, considered as vectors ($f_j$ and $b_j$, respectively).

The complete Itakura-Saito algorithm is as follows:

- Initialisation of the backward and forward prediction errors from the original signal:

$$f_0(n) = b_0(n) = s(n) \qquad \textbf{Eq. (2-25)}$$

- Loop $1 \leq j \leq K$:
  - Evaluation of the $j^{th}$ reflection coefficient (also known as PARtial CORrelation – PARCOR – Coefficient) using Eq. (2-24).

- Evaluation of the prediction coefficients of $j^{th}$ order.

$$a_j^j = c_j;$$
$$a_i^j = a_i^{j-1} - c_j a_{j-i}^{j-1} \qquad \textbf{Eq. (2-26)}$$

- Evaluation of the new forward and backward prediction errors.

$$f_j(n) = f_{j-1}(n) - c_j b_{j-1}(n-1)$$
$$b_j(n) = -c_j f_{j-1}(n) + b_{j-1}(n-1) \qquad \textbf{Eq. (2-27)}$$

- If $j \neq K \rightarrow$ Go to 1

- Finally:

$$a_i = a_i^K; a \leq i \leq K \qquad \textbf{Eq. (2-28)}$$

Although Itakura-Saito and Levinson-Durbin algorithms almost arrive to the same solution, the former is more compact, modular, extensible and reusable.

Forward and backward prediction error filters can be implemented by transversal filters. However, filters derived from Eq. (2-24) and Eq. (2-27), known as PARCOR filters or Lattice Filters (see Figure 2-16), have the advantage of allowing the extension of the filter from order K to order K+1 with the addition of a new lattice section - Eq. (2-27). This means that there is no need of re-evaluating neither PARCOR coefficients nor forward or backward prediction errors previous to the new added lattice.

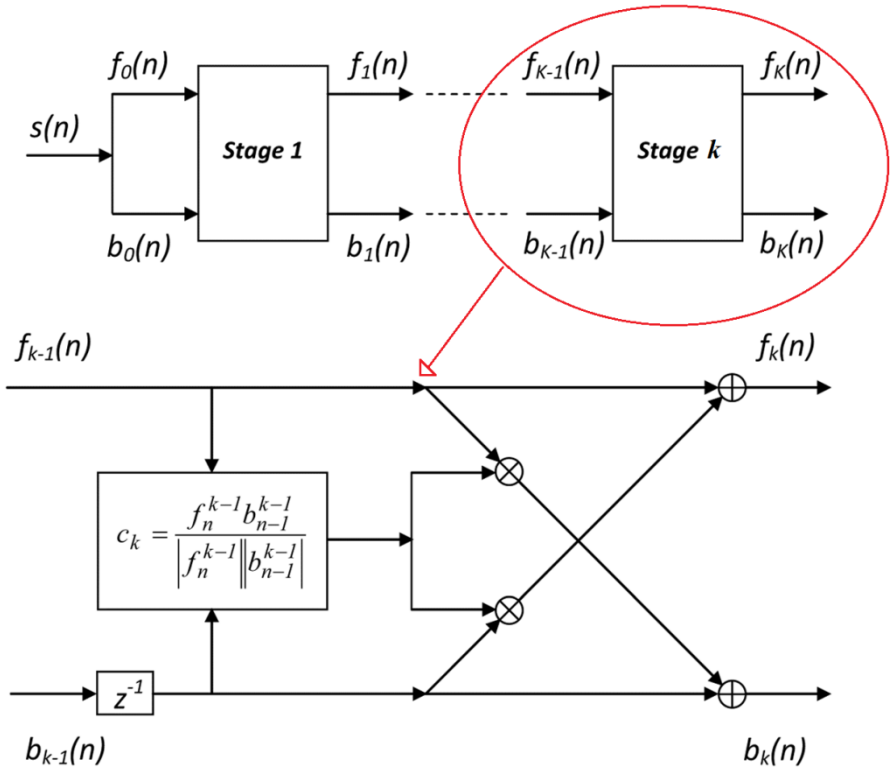



**Figure 2-16** Lattice structure that generates forward and backward predictor errors for all optimal predictors with order $1 \leq k \leq K$, and details of the $k^{th}$ lattice

#### *2.3.1.1 Properties of the prediction error filters*

- Minimum-phase property of the prediction error filters.

  For the optimal predictor, i.e. the optimum set of prediction coefficients, it can be shown that it is minimum-phase. In other words, it has all of its singularities inside the unit circle, unless the original signal, *s(n)*, represents a line spectral process. The important consequence of this property is that the prediction error filter has a causal stable inverse $1/A_K(z)$.

- Constrained log-spectrum.

  Another consequence derived from the minimum phase property is the fact that the log spectrum is constrained:

$$\int_{-\pi}^{\pi} \log(|A(w)|^2)\, dw = 0 \qquad \textbf{Eq. (2-29)}$$

  In fact, the mean of the log spectrum for a causal minimum-phase prediction error filter, is:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|A(w)|^2)\, dw = 2\log(|a_0|) = 2\log(1) = 0 \qquad \textbf{Eq. (2-30)}$$

- Orthogonality Principle.

  One of the most important properties of the prediction error filters is the uncorrelation of forward and backward prediction errors respect to the input signal, as it is shown by the set of *K* Eq. (2-9). In fact this set of equations reflects the *Orthogonality Principle*, which states that for the optimal predictor coefficients, the forward error prediction and the delayed data of the reference signal over the temporal window *W* are orthogonal. This principle is geometrically depicted in Figure 2-17 for a 2-dimension vector space.

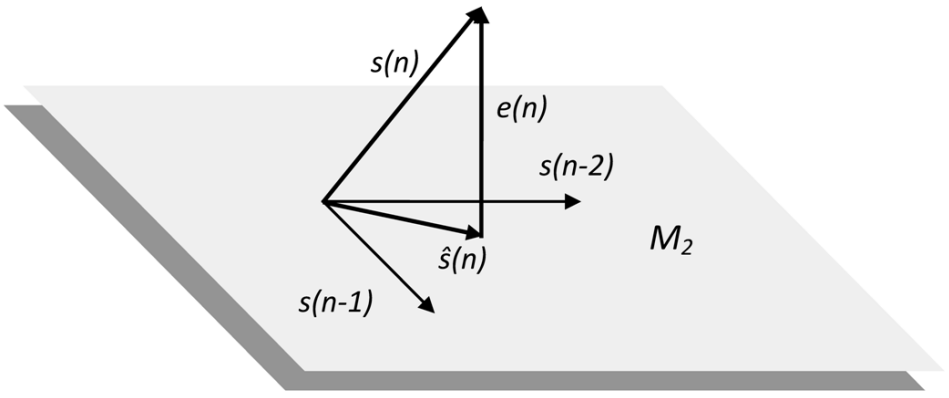



**Figure 2-17** Geometrical interpretation of the *Orthogonality Principle* in a prediction error filter of order 2

In Figure 2-17, the image shows that the norm of the forward error, *e(n)*, will be minimum when the vector is perpendicular to the basis B={$S_K$}, where $S_K$={*s(n-1),…s(n-K)*} are the set of delayed data of the reference signal that build a vector subspace of $M_K$-dimension (*K*=2 in Figure 2-17). The most important consequence of the Orthogonality Principle is the spectral properties of the

prediction process. As far as the forward prediction error, *e(n)*, is orthogonal to all delay samples of the reference signal (i.e. *e(n)* and the delayed samples of the reference signal are order-2 statistically independent), we can conclude that *e(n)* is also orthogonal to all the delayed estimations of it, i.e. is orthogonal to {*e(n-l) ; l > 0*}. Therefore, the statistical nature of *e(n)* is that of a white stochastic process if *K* is large enough. If $K \to \infty$, the resulting filter is known as *Wiener Filter* [Wiener,1949].

Given a stochastic process *s(n)* zero mean and $\sigma^2$ variance; if all the delayed samples over an interval *j* are mutually orthogonal, then it is considered a white stochastic process:

$$\sum_{W} s(n)s(n-j) = \begin{cases} \sigma^2; j = 0 \\ 0; j \neq 0 \end{cases} \qquad \textbf{Eq. (2-31)}$$

If we apply this test to forward error prediction, *e(n)*:

- By the *Orhtogonality Principle* we already know that *e(n)* is orthogonal to all delayed samples of *s(n)*.
- We also know that *e(n)* is a linear combination of present and past samples of *s(n)*:

$$e(n) = s(n) - \sum_{i=1}^{K \to \infty} a_i s(n-i) \qquad \textbf{Eq. (2-32)}$$

  This equation can be expressed in terms of the convolution product as:

$$e = s * h_W \qquad \textbf{Eq. (2-33)}$$

  Where $h_w$ = *1-*$a_w$ is the impulse response of the equivalent Wiener prediction error filter, assuming that the $a_w$ is the impulse response of the equivalent Wiener prediction filter.

- Then:

$$\begin{aligned} \sum_{W} e(n)e(n-j) & \\ &= \sum_{W} e(n)s(n-j) \\ &- \sum_{i=1}^{\infty} a_i \sum_{W} e(n)s(n-j-i) \\ &= \begin{cases} \sum_{W} e^2(n); j = 0 \\ 0; j \neq 0 \end{cases} \end{aligned} \qquad \textbf{Eq. (2-34)}$$

  Therefore *e(n)* is white.

### 2.3.2 Source-Tract separation proposed algorithm

The theory of inverse filtering via linear prediction, applied to Fant's production model (see Figure 2-13), has been used for the reconstruction of the glottal source. Through

this method, the prediction error can be regarded as a glottal residual, which is related with the first derivative of the glottal source, which at the same time is related with the first derivative of the glottal flow. Glottal flow, glottal source and glottal residual are also known as Glottal Correlates of voice.

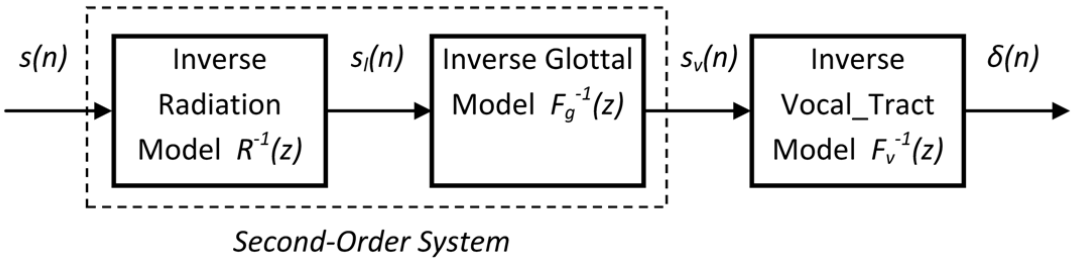



**Figure 2-18** General filtering model for the inversion of Fant's voice production model (see Figure 2-13)

Figure 2-18 provides a general block diagram in which the inverse functions needed to generate an estimation of the voice signal before lip radiation and an estimation of the glottal source are included in the corresponding block. The inversion of the model is performed by means of linear prediction methods already presented and defined over the convolution product in Eq. (2-33). More specifically, predictive structures based on the Itakura-Saito PARCOR algorithm have been modified by the research group in order to model and invert the system, providing a highly efficient algorithmic structure, known as paired lattice.

- **Lip radiation compensation model**:

  In order to compensate the lip radiation effects, a first order transversal filter, like the one depicted in Figure 2-19.b, can be used. Additionally this process can be implemented using a first order prediction error lattice like the one in Figure 2-19.a, which operates like an FIR filter according to the recursion:

$$f_k(n) = f_{k-1}(n) + c_{k-1}b_{k-1}(n-1) \qquad \textbf{Eq. (2-35)}$$

  So when $k$=1 and $c_0$=-$r_f$ (which is the first reflection coefficient) and given:

$$f_0(n) = b_0(n) = s(n) \qquad \textbf{Eq. (2-36)}$$

  the lattice behaves like a first order differentiator:

$$s_l(n) = f_l(n) = s(n) - r_f s(n-1) \qquad \textbf{Eq. (2-37)}$$

  with transfer function given by:

$$H_l(z) = R^{-1}(z) = 1 - r_f z^{-1} \qquad \textbf{Eq. (2-38)}$$

  This cancels the first order pole introduced by lip radiation effects.

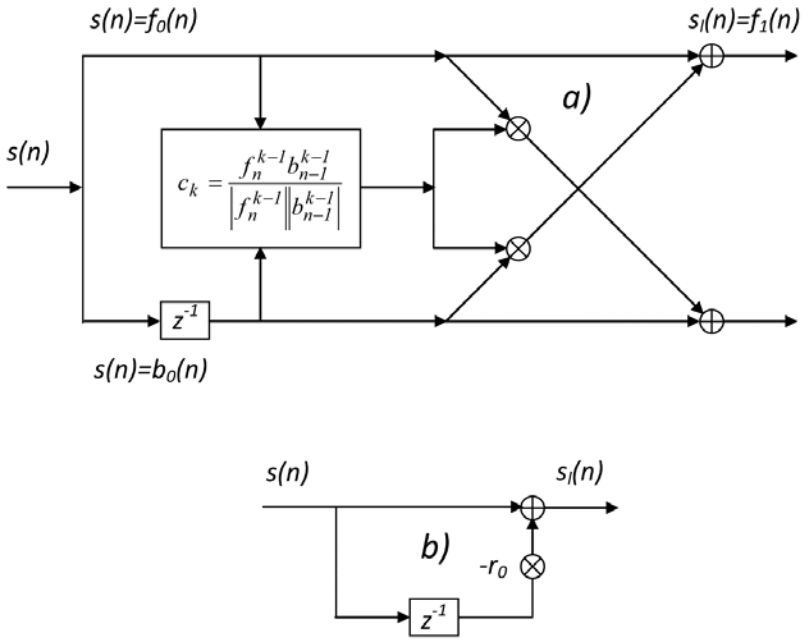



**Figure 2-19** Lip radiation cancel filter implemented as a first order lattice. a) PARCOR lattice implementation. b) First order transversal filter

- **Source compensation model:**

  Source model inversion can be carried out according to the following recursion:

$$e(n) = u(n) + \gamma u(n-1) \qquad \textbf{Eq. (2-39)}$$

  which corresponds to a first difference filter given by:

$$H_g(z) = 1 - \gamma z^{-1} \qquad \textbf{Eq. (2-40)}$$

  It can be implemented using a first order prediction filter with coefficient $\gamma$. This procedure is just an approximation, as glottal source is not a minimum-phase signal [D'Alessandro,2007]. However, this does not prevent the problem to be solved by the iterative method proposed, as it can reach a satisfactory solution in a few iterations.

- **Vocal tract transfer function estimation:**

  As depicted in Figure 2-18, a combination of the glottal and radiation models is possible whenever we consider the operation of the different blocks as commutative operators. The association of the first two blocks may be inverted using a second order prediction error filter. Thus the speech trace is processed by a second order inverse filter with transfer function given by:

$$\begin{aligned} H_{lg}(z) = R^{-1}(z)F_g^{-1}(z) &= \left(1 - r_f z^{-1}\right)\left(1 - \gamma z^{-1}\right) \\ &= 1 - \left(r_f + \gamma\right)z^{-1} + r_f \gamma z^{-2} \end{aligned} \qquad \textbf{Eq. (2-41)}$$

  where $R^{-1}(z)$ and $F_g^{-1}(z)$ respectively represent the inverse transfer functions of the lip radiation model and the glottal pulse generation model. This second

order inverse filtering may be implemented by a second order prediction error lattice. Obviously, the output of this second-order filter, $s_v(n)$, is the signal contributed by the vocal tract model:

$$s_v(n) = s(n) - (r_f + \gamma)s(n-1) + r_f\gamma s(n-2) \quad \textbf{Eq. (2-42)}$$

This equation can be expressed in terms of convolution product as:

$$s_v = s * h_{gl} \quad \textbf{Eq. (2-43)}$$

Thereby, the vocal tract inverse model will then be the Wiener Filter reducing $s_v(n)$ to a signal with white power spectral distribution:

$$s_v * h_{gl} = \delta(n) \quad \textbf{Eq. (2-44)}$$

The Wiener filter may be implemented by a prediction error lattice with a dimension *K* large enough to reduce the output power spectrum to a flat behaviour in the frequency domain.

With what has been presented, the actual signal processing procedures used to estimate the glottal source, based on Alku et al. IAIF method [Alku,1992], [Alku,1994], [Cheng,1989], are explained in detail below. Figure 2-20.a provides a block diagram of a complete system to estimate the glottal residual $u(n)$ from the speech signal, expressed in terms of convolution in Eq. (2-4). The first stage in the process consists in estimating the speech trace at lips $s_l(n)$, i.e. remove the radiation effects to get the radiation compensated voice. The construction process of $s_l(n)$ can be express in terms of convolution products as:

$$s * h_{gl} = \{s_l * r\} * h_l = s_l * \{r * h_l\} \cong s_l \quad \textbf{Eq. (2-45)}$$

where it has been assumed that the operators $r$ and $h_l$ are inverse to each other with respect to the convolution.

In addition, a first estimation of the glottal impulse response $h_g$ is also obtained, which permits to remove the influence of the glottal residual from the radiation compensated voice:

$$s_v = s_l * h_g \quad \textbf{Eq. (2-46)}$$

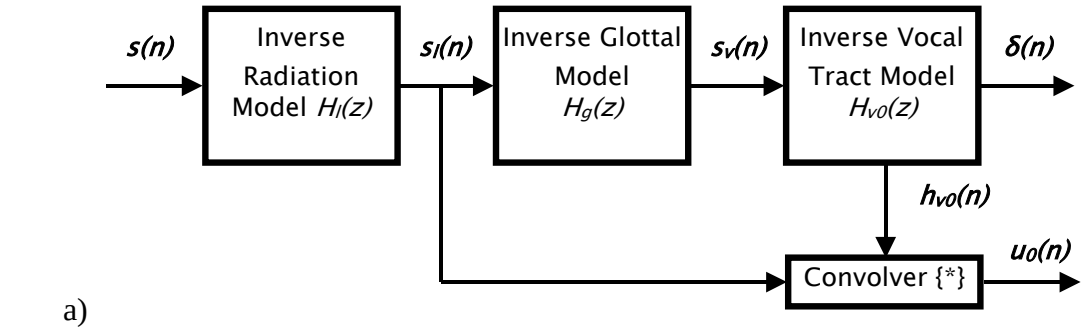
The resulting first estimation of the de-glottalised voice, $s_v$, may be inverted finding its equivalent inverse impulse response $h_{v0}$ (equivalent Wiener filter), such that:

$$s_v * h_{v0} = \delta \quad \textbf{Eq. (2-47)}$$

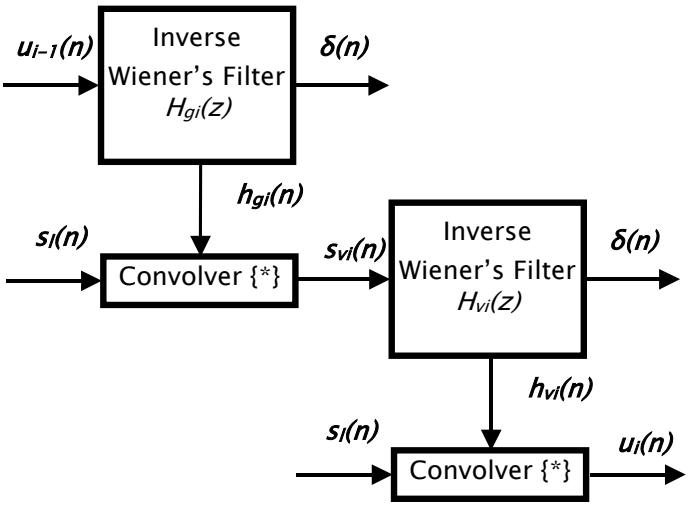
The inverse impulse response of the vocal tract may be used to remove the influence of the vocal tract from the radiation compensated voice, $s_l$, by direct convolution producing a more accurate estimation of the glottal residual $u(n)$:

$$s_l * h_{v0} = \{f_g * f_{vt}\} * h_{v0} = f_g * \{f_{vt} * h_{v0}\} \cong f_g = u_0 \qquad \textbf{Eq. (2-48)}$$

In order to obtain a more accurate estimation of the glottal residual, Figure 2-20.b describes an iterative system, in which the first estimation of the glottal residual is used to improve the estimation of the vocal tract transfer function from the radiation compensated speech. According to the block diagram, the (i-1)$^{th}$ estimation of the glottal residual, $u_{i-1}(n)$, is Wiener-inverse filtered, reducing it to a white process, $\delta(n)$. The prediction coefficients of the equivalent Wiener filter, $h_{gi}(k)$, constitute the impulse response of such filter, which when convolved with $s_l(n)$ produced the i$^{th}$ estimation of the de-glottalised speech, $s_{vi}(n)$. Again, $s_{vi}(n)$ is Wiener-inverse filtered, reducing it to a white process, $\delta(n)$.The prediction coefficients of the equivalent *Wiener filter* $h_{vi}(k)$ when convolved with $s_l(n)$ remove the influence of the vocal tract, reducing it to the $i^{th}$ estimation of the glottal residual $u_i(n)$.




a)




b)

**Figure 2-20** Block diagram method for the reconstruction of the glottal residual correlate by complemented inverse filtering: a) Initial estimate of the glottal residual correlate. b) i$^{th}$ iteration of the tuning process**.**

Our approach (Figure 2-21) is based on Alku's iterative process, but taking into account that the structure of the Wiener filter (implemented by a PARCOR lattice) and its associate convolver (see Figure 2-20.b) used to remove the influence of the transfer function estimated by inverse filtering are integrated together in a single structure (paired lattice) as put forth in Figure 2-22.

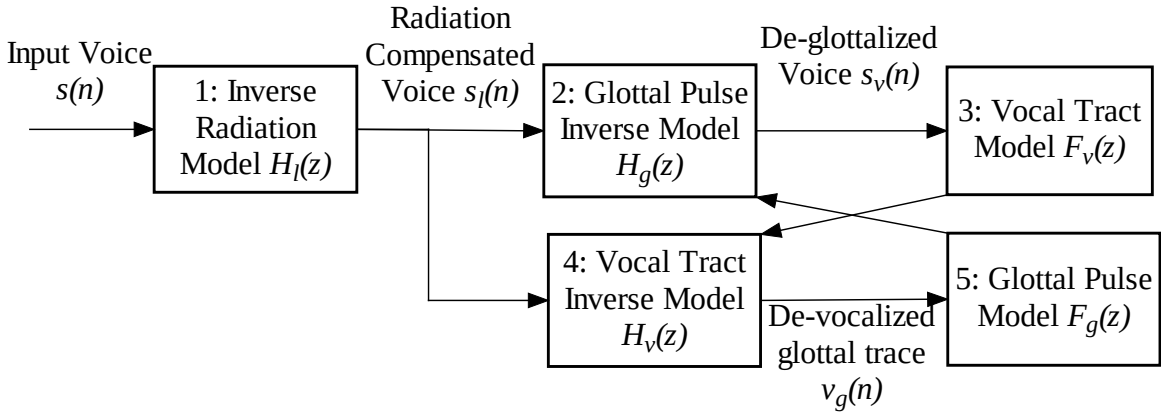



**Figure 2-21** Iterative estimation of the vocal tract transfer function $F_v(z)$and the glottal residual $v_g(n)$. Blocks $F_v(z)$ and $H_v(z)$ or $F_g(z)$ and $H_g(z)$ are implemented by successive chains of adaptive lattice filters

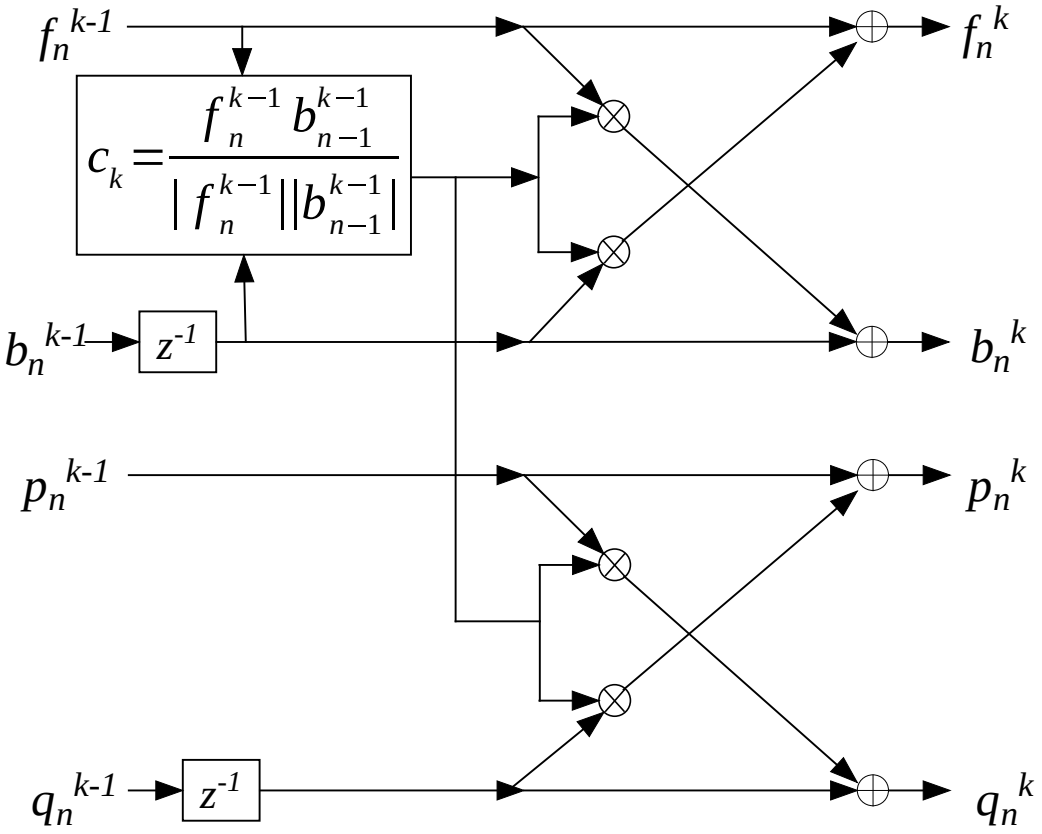



**Figure 2-22** Paired lattice joint estimator that combines a modelling filter section with an inverse filtering section, joining in a single structure the two blocks of each diagonal in Figure 2-21

As depicted in Figure 2-21, the first step consists in removing the radiation effects from voice signal, $s(n)$, by filtering with $H_l(z)$. This is performed with a first order prediction error lattice, like the one in Figure 2-19.a, which operates like an FIR filter. In the next step, an inverse filter, $H_g(z)$, of the glottal pulse generating model, $F_g(z)$, is used to remove the influence of the glottal source from the radiation compensated voice, generating a first estimation of de-glottalised voiced, $s_v(n)$. This trace is somehow

equivalent to the vocal tract response to a delta pulse train. In this first iteration, $H_g(z)$, do not need to be a very precise estimation, as it will be refined in successive iterations. In step 3, the vocal tract model, $F_v(z)$, is estimated by inverse filtering of this last trace using another adaptive lattice (in practice the order of this filter will depend on the quality of speech signal and gender of speaker). In the next step, the radiation compensated voice will be filtered with the vocal tract inverse model, $H_v(z)$, removing the vocal tract influence. Therefore, the residual $v_g(n)$ contains only glottal source information (in fact, depending on the model order of $H_l(z)$ and $H_v(z)$, $v_g(n)$ contains direct estimations of the glottal source or its derivatives). Finally, step 5 produces a more precise glottal source generating model, $F_g(z)$, by inverse filtering of the residual trace obtained in the previous step, allowing an improvement of $H_g(z)$. Repeating steps 2-5 allows the successive improvement of estimations $s_v(n)$ and $u_g(n)$ (obtained by integrating $v_g(n)$). The iteration loop is repeated as many times as necessary, according to a stabilisation criterion. In practice, usually two or three iterations are enough to concurrently produce a reliable estimation of the glottal source ($u_g(n)$) and vocal tract ($s_v(n)$) in a real case. This last trace, almost free from glottal influence and associated to the unit impulse response of the vocal tract, is particularly relevant in the accurate estimation of formants (resonances) of the vocal tract as it is almost free from the influence of the glottal dynamics in the vocal tract pattern, as opposed to whether this estimate was made on the original speech signal.

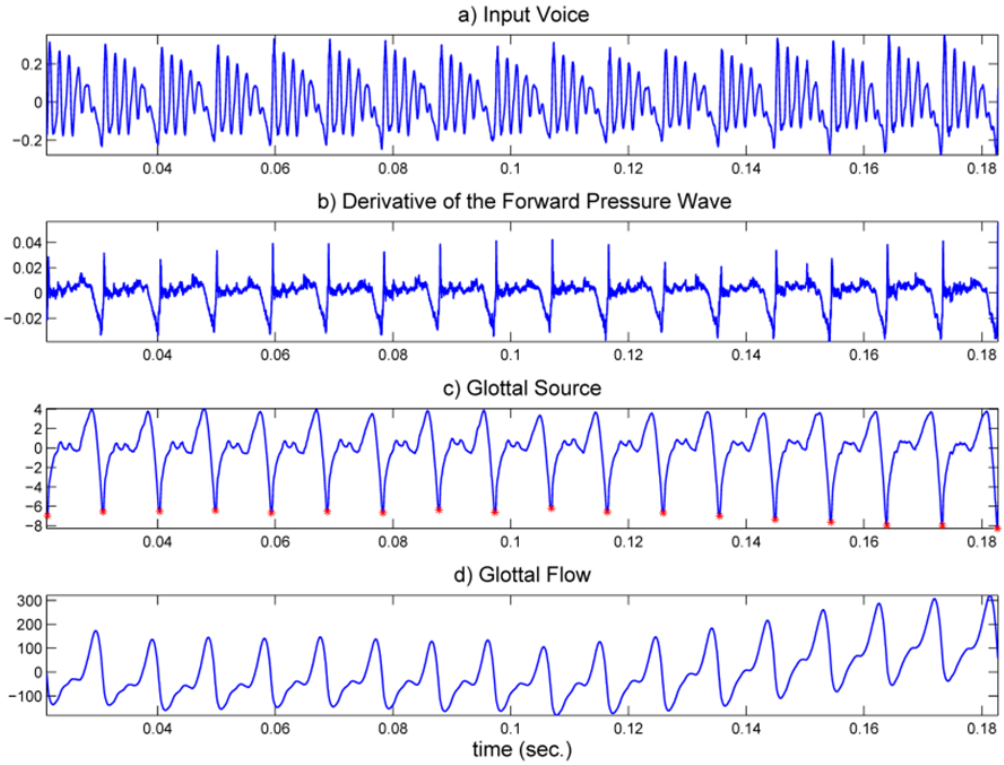



**Figure 2-23** Speech glottal traces that result from the described separation algorithm: a) Input voice (original speech wave); b) Glottal residual that results from linear prediction estimation; c) Glottal source that results from integrating the glottal residual; d) Glottal flow, resulting from integrating the glottal source.

Figure 2-23 depicts the set of signals that can be reconstructed from the speech wave through the application of the separation algorithm described above. In Figure 2-23.c it

is shown how the reconstructed glottal source clearly follows the L-F model, although not exactly, as the L-F model rarely appears in real speech.

Once the glottal source have been estimated, the next stage involves the differentiate estimation of the body and cover correlates of the glottal source. This have been achieved, as described in [Gómez,2004-A], [Gómez,2004-B], using an adaptive method that evaluates the amplitude of the semi-sinusoidal arch corresponding to the average acoustic wave defined in [Titze,1994-A]. In fact, it is an adaptive method, synchronous with the glottal cycle, which evaluates the amplitude of a semi-sinusoidal arch, $s_g(n)$ that minimises the error energy between the glottal source, $u_g(n)$, and the arch in the glottal cycle, defined as having $N_k$ samples:

$$L = \sum_{n \in W_k} \varepsilon_k^2 = \sum_{n \in W_k} \left(u_{gk}(n) - s_{gk}(n)\right)^2 \qquad \textbf{Eq. (2-49)}$$

Where $W_k$ is the associated window to the $k^{th}$ phonation cycle, and:

$$s_{gk}(n) = s_{0k} \sin(\omega_k n\tau); n \in W_k \qquad \textbf{Eq. (2-50)}$$

The optimal amplitude for each arch can be obtained through the minimisation of the cost function L, with respect to the amplitude $s_{0k}$:

$$\frac{\partial L}{\partial s_{0k}} = 0 \rightarrow s_{0k} = \frac{\sum_{n \in W_k} s_{gh}(n) \sin(w_k n\tau)}{\sum_{n \in W_k} \sin^2(w_k n\tau)} \qquad \textbf{Eq. (2-51)}$$

The cover dynamic component or mucosal wave correlate is defined as the difference between the glottal source and the average acoustic wave, evaluated over each glottal cycle:

$$s_{ck}(n) = u_{gk} - s_{0k} \sin(w_k n\tau)\,; n \in W_k \qquad \textbf{Eq. (2-52)}$$

From Eq. (2-52) we can define the cover dynamic component (or mucosal wave correlate) as the minimum energy signal that is obtained after subtracting the body dynamic component of the vocal folds from the glottal source. Figure 2-7 shows an estimation of these components on the ideal L-F model of the glottal source, while Figure 2-24 and Figure 2-25 respectively show these components on male and female real estimations of the glottal source.

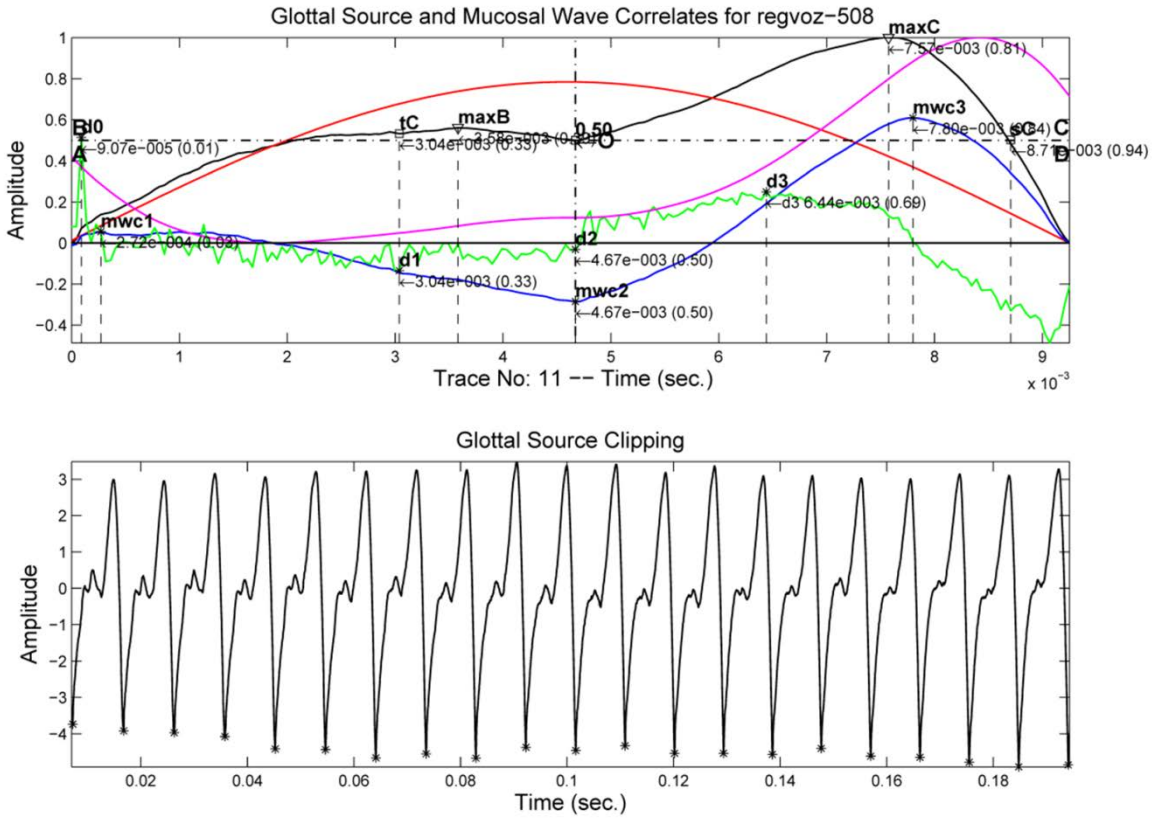



**Figure 2-24** Real glottal source models reconstructed according to the presented method for a male speaker (black: Glottal Source; red: Average Acoustic Wave; magenta: Glottal Flow; blue: Mucosal Wave Correlate; green: derivative of the Mucosal Wave Correlate). The upper image provides a graphical representation of the mucosal wave correlate estimation by subtracting the body dynamics to the glottal source. The lower image provides a representation of the detected glottal cycles.

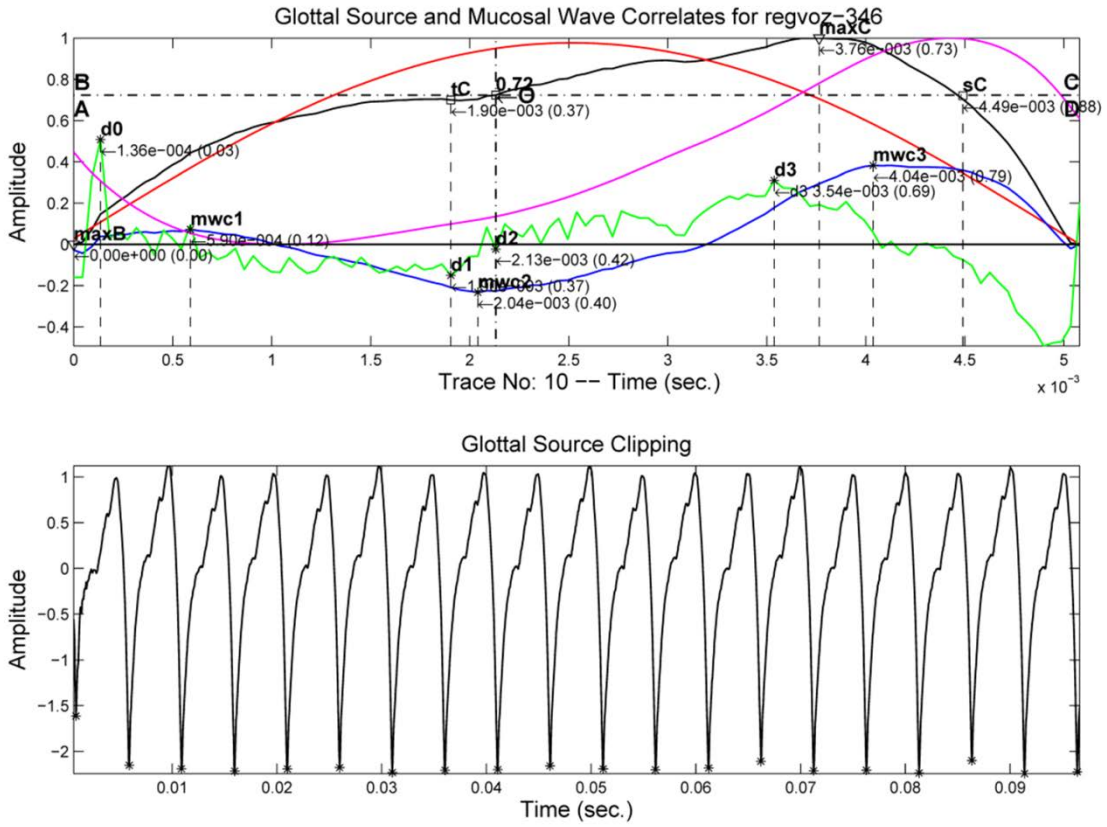



**Figure 2-25** Real glottal source models reconstructed according to the presented method for a female speaker (black: Glottal Source; red: Average Acoustic Wave; magenta: Glottal Flow; blue: Mucosal Wave Correlate; green: derivative of the Mucosal Wave Correlate). The upper image provides a graphical representation of the mucosal wave correlate estimation by subtracting the body dynamics to the glottal source. The lower image provides a representation of the detected glottal cycles.

### 2.3.3 Comments on the source-tract separation algorithm

The most relevant features of the source-tract reconstruction process employed, which constitutes a novel approach respect to previous works in the area are:

- The structure of the prediction error filters, used in the model inversion steps, follows the lattice structure depicted in Figure 2-16, instead of the transversal filters typically used in the state-of-the-art. This methodology allows the reconstruction of the transversal section of the vocal tract from the PARCOR coefficients of the vocal tract model, $F_v(z)$. As it has already been said, this model is almost free from the influence of the glottal source dynamics, therefore the all-pole model is more precise and the reflection coefficients are of better quality; consequently the reconstruction of the transversal area of the vocal tract is better than in the state-of-the-art systems. The reconstruction can be expressed as:

$$s_i = s_i \frac{1 - r_i}{1 + r_i} \qquad \textbf{Eq. (2-53)}$$

where $S_i$ represents the equivalent section of the $i^{th}$ stage of the acoustic tube (see Figure 2-14), and $r_i$ is the $i^{th}$ reflection (PARCOR) coefficient. Moreover, if the sampling frequency of the voice signal is adequately established, then more

accurate estimates of the reflection waves inside the acoustic tube that models the vocal tract (related with pressure waves inside it) can be produced. Following this approach a non-invasive and more precise reconstruction of the glottal source can be performed.

- The lattice filter implementation follows an adaptive approach, so it better models the non-stationarity of the signal
- As it has already been said, and reflected in Figure 2-21 and Figure 2-22, a paired lattice can integrate the Wiener Filter and its associated convolver, which allows removing the influence of a transfer function estimated by inverse filtering. In fact, lattice filter properties have been applied to carry out the described inversion process. More specifically, the residual error from a lattice filter, as shown for example in [Deller,1999], may be seen as the output of an all-pole filter inverse to the lattice input trace. This result allows to jointly build the inverse impulse response to $s_v$ by a lattice reducing this signal to a white series, and at the same time convolve the associated signal $s_l$ with the same inverse impulse response by means of a paired lattice (lower) which uses the same reflection coefficients derived in the driving lattice (upper) as shown in Figure 2-22.

As we have pointed out, the present method follows the IAIF-method presented by Alku in [Alku,1992], which has found considerable applications in the clinical field and in speech quality analysis in general [Gómez,2009], [Gómez,2007-A]. However this is not the only existing method to accurately estimate the source and vocal tract component of voice. [Backstrom,2002] also present a modification to Alku's method which consists in estimating the vocal tract transfer function using a discrete all-pole (DAP) modelling technique instead of LPC. According to their results, this modification provides better estimation of the first formants, especially the first one, thus decreasing the amount of formant ripple in the estimated glottal flow. However, this improvement is more relevant when applied to high pitch frequency voices and when the vocal tract can be well modelled using an all-pole envelope, which is not always possible.

Previously, [Plumpe,1999] introduced a robust approach to identify the glottal closed-phase based, in which the absence of source-filter interaction will result in no or little formant modulation, on first formant tracking calculated from the vocal tract estimates. Then the glottal flow results from inverse filtering with a vocal tract estimated derived from the covariance method within this interval. This method, despite being robust presents certain drawbacks. First of all the computational cost derived of using a one-sample shift sliding window for close-phase estimation. It also requires the use of multiple pitch cycles when dealing with high pitch speakers. In addition, the close-phase region detection is clearly affected by the fact that under some circumstances or some voice pathology the vocal folds may never completely close. Finally there is a need for a way to determine a stable and optimal vocal tract function in the close phase. Following the idea of close-phase analysis [Akande,2005] introduced the AEVT (adaptive estimation of the vocal tract transfer function) method which is focused more on precise vocal tract estimation rather than on exact source-tract separation. In this adaptive method, the first step consists in removing the glottal frequency by a frequency-selective, multi-pole, zero-phase lag high-pass filter whose role-off is adjusted to meet the low-frequency gain criterion. Using this high-pass filtered data and applying covariance linear prediction analysis and an adaptive algorithm that selects an optimum linear prediction order that satisfies the criteria for minimum phase systems,

the vocal tract filter parameters are estimated over a pitch cycle. Finally, removing the effects of the vocal tract and lip radiation from original speech by inverse filtering and subsequent integration provides an estimate of the glottal volume velocity. Again, this method is limited by some problems pointed out by the authors. First, it seems to work only in the cases in which the glottal frequency is lower and sufficiently separated from the first formant which is not always the case. Additionally and probably more important is the fact that although removing the influence of the glottal frequency the method does not remove glottal contributions over the entire spectrum. Again high pitch voices seem to be problematic.

[Gudnason,2008] applied the DYPSA algorithm [Naylor,2007] to identify glottal closure instants in each cycle. Using multicycle close-phase analysis an autoregressive model of the vocal tract is estimated and used to produce vocal-tract mel cepstrum coefficients (VTCC). In order to obtain a representation of the voice source (different from glottal source), the VTCC are subtracted from the mel-frequency cepstrum coefficients (MFCCs) of the speech frame. As already pointed out, accurate detection of closed phase can be difficult under some circumstances such as presence of noise or soft phonation. To overcome this problem, in recent work [Kinnunen,2009] proposed an approach similar to ours. Again the IAIF method is applied to extract an estimation of both the vocal tract and glottal source, from this last signal, the source mel-frequency cepstral coefficients are evaluated to capture the frequency-domain characteristics of voice.

Moreover the differences with respect to these methods not only exist in the separation algorithm but in the features derived from the source and tract estimations.

## 2.4 BIOMETRIC CHARACTERISATION OF VOICE

Robust speaker recognition is based on the contextual use of a set of features which must be extracted at different levels: biometric, acoustic-phonetic, phonologic, prosodic and rhythmic, morphological-linguistic, dialectal, etc. However, if the speaker recognition system is supposed to be language independent then it must rely on acoustic and biometric features rather than in semantic, dialectal or pronunciation features. Although these last features provide alternative information for speaker characterisation they are more dependent on the socioeconomic, educative and linguistic level of the speaker. On the other hand aspects such as prosody, rhythm, speed and pitch, or modulation, along with issues such as nasality, depth or roughness and others associated with the anatomy of the vocal apparatus and laryngeal system (biometric), may constitute low level clues that can be easily characterised.

Despite this consideration the most important aspect regarding characterisation features is that they must provide high discriminative power between speakers, i.e., high inter-speaker variability and low intra-speaker variability. Characterisation features can be split into two main groups, labelled as low-level features and high-level features. Regarding low-level features we can perform an additional classification:

- Biometric level: It is based on the detection of a set of features during voice production that are not subject to imposture or at least difficult to forge, as they are linked with physiological and psychological aspects of the speaker. For instance, jitter (short-term perturbation in the fundamental frequency, F0), shimmer (perturbations on the cycle-to-cycle phonation amplitude) or glottal source characteristics.

- Spectral level: It has been extensively used in speaker recognition systems for feature extraction. Typically this method consists in identifying spoken message fragments in an utterance and then estimate the power spectral density of the voiced fragments following standard methods (such as FFT or LPC) evaluated in overlapped sliding windows in order to achieve a good temporal monitoring while warranting the pseudo-stationarity of the evaluated frame. Then, the PSD is parameterised in order to obtain MFCC (Mel-Frequency Cepstral Coefficients) patterns and their temporal derivatives (ΔMFCC), which aligned in streams, will serve as templates in the classification process.

Biometric and spectral levels are considered as the lowest levels while the remainders are considered as high level features. As the use of high level features is beyond the scope of the present dissertation, only a brief description has been already presented in Chapter 1. Moreover, speaker recognition systems that use low level features usually provide better results than those using high-level features [Reynolds,2003].

The aim of the present proposal is to study the use of the information obtained from the glottal source correlates which result from removing the influence of vocal tract on the speech signal (to incorporate speaker information relying exclusively in biometry) together with parameters obtained from the vocal tract transfer function (to take into account the acoustic-phonetic character of the speech signal), and finally fusing these results in a non-competitive way with classical parameters. As we have already established, both glottal source and vocal tract systems are involved in speech production processes. Most of the previous works in the area and even current state-of-the-art systems still use classical parameterisation techniques that consider the power spectral density of speech as a whole or just parameterisation techniques that only take into account the vocal tract information. This can be easily checked by reviewing the description of the systems submitted to the NIST 2010 SRE (URL: NIST SRE 2010), in which almost 100% of those systems used MFCC parameters extracted from the voice signal. Additional examples can be found, for instance in [Kinnunen,2009], [Gudnason,2008], [Ferrer,2008], [Nickel,2006], [Bimbot,2004]. It may be expected that glottal information will be more influenced by the speaker's phonation habits, while the description of the vocal tract will be more conditioned by the phonetic structure of the message. On its turn the power spectral density of the glottal source is strongly conditioned by the biomechanics of the vocal folds. The hypotheses that support this approach are:

- Voice biometrics based on the use of the glottal trace is hardly controllable by the speaker, hence making it hard to forge. In this way the false-acceptance produce by wolf-like speakers is expected to be reduced.
- The separation of biometric (glottal) and articulatory (vocal tract) features into two independently-treated sets, will reduce the dependence degree of intra-speaker statistical distributions, thus producing less variability speaker descriptions.

In this way, it is necessary to define two processing cores (for glottal parameterisation and vocal tract parameterisation) which will cooperate with another core working on classical spectral parameters extracted from the original voice signal. Figure 2-26 provides an analytical description of voice biometry from the production model point of view, and also provides a graphical interpretation of the different cores that will be presented: glottal, vocal tract and whole voice cores.

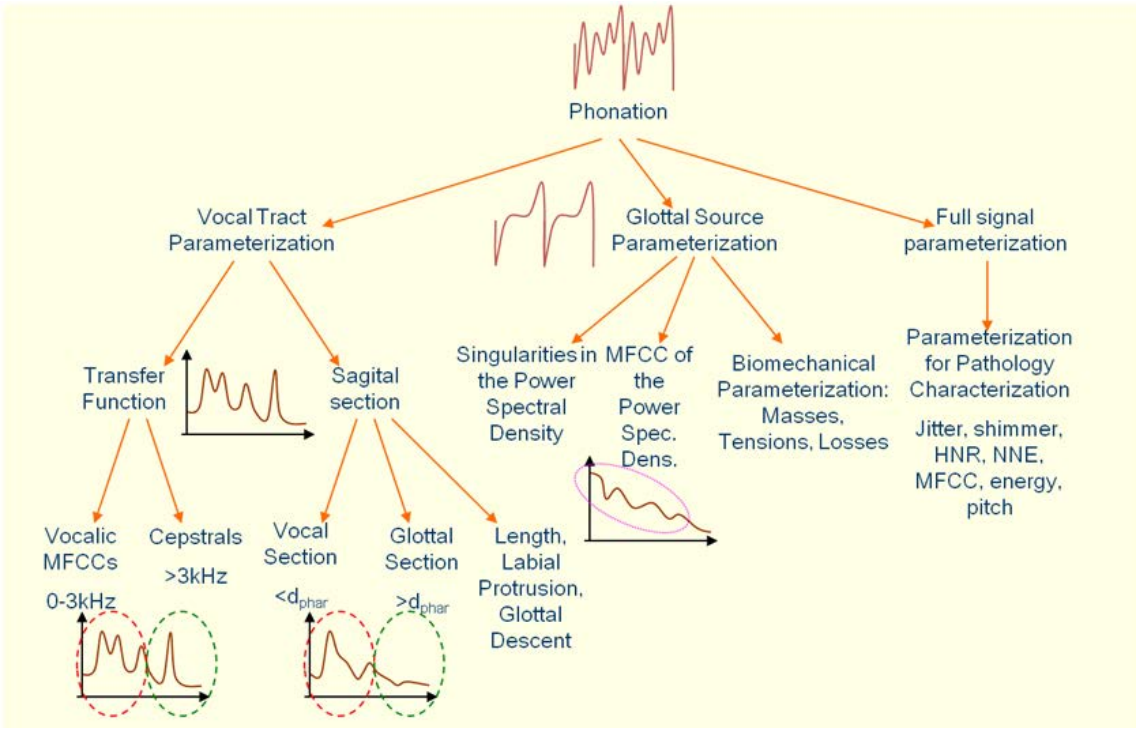



**Figure 2-26** Analytic description of voice biometry from the production model in terms of vocal (mainly message dependent) and glottal (mainly biometric) characteristics

### 2.4.1 Feature extraction

#### *2.4.1.1 Voice signal feature extraction*

A wide variety of parameters has been investigated for its application in automatic speaker recognition, based on both frequency and time domain analysis. Many of these parameters are related to some property of the short-time power spectrum, as it has been shown to be very effective in this area because the spectrum reflects somehow the speaker's vocal tract structure. The short-time spectrum provides a complete although not compact description of the acoustical characteristics of speech, being a three-dimensional representation with the coordinates being time, frequency and energy.

Among the different parameters proposed, the so-called mel-frequency cepstral coefficients (MFCCs) [Davis,1980] introduced in early 1980s for speech recognition and then adopted in speaker recognition seem to be difficult to beat in practice. In this study, MFCCs have been used as the state-of-the-art speaker specific features. In what follows a brief description on the steps involved in the MFCC computation will be performed. Figure 2-27 provides a block diagram representation of MFCC parameterisation.

The speech signal continuously changes due to articulatory movements, i.e. the temporal variation of the vocal tract shape during the utterance. This temporal variation, due to vocal tract characteristics, is relatively slow; therefore the speech signal is assumed to remain stationary in short periods of time. In other words, the speech signal can be regarded as having nearly constant characteristics in short periods such as those 20-40 ms in length [Furui,2000]. Once the signal is broken down in short frames, a spectral feature is extracted from each frame.

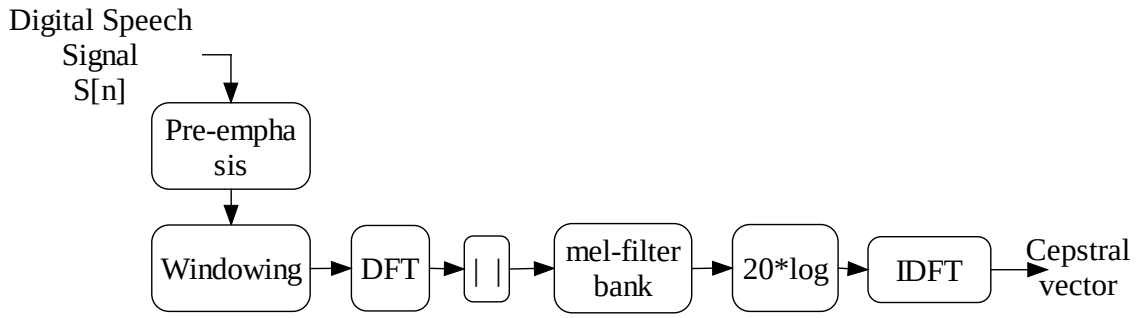



**Figure 2-27** Block diagram representation of the MFCC parameterisation

Prior to the frame decomposition, a pre-emphasis high-pass filter with transfer function:

$$H_{pre}(z) = 1 + a_{pre}z^{-1} \qquad \textbf{Eq. (2-54)}$$

is usually applied, where a typical range of values for $a_{pre}$ is [-0.98, -0.95].

The usefulness of this filter admits two different explanations. First, it serves to offset the negative spectral attenuation of the voiced sections of speech due to physiological characteristics of speech production, thereby improving the analysis efficiency. Alternatively, the pre-emphasis filter amplifies the spectrum region above 1kHz (area which is more sensitive in auditory processing), thus assisting the spectral analysis algorithm, in modelling the most perceptually important aspects of speech spectrum [Picone,1993]. However this pre-emphasis filter is not always applied and no definite answer on the usefulness of it has been found but empirical experimentation.

No matter whether a pre-emphasis filter has been applied, the next step consists in multiplying each frame by a smooth window function. The window function is needed because of the finite-length effects of the discrete Fourier transform (DFT). Different window functions have been proposed including the rectangular, Hamming, Hanning, Blackman, etc.; however, in practice, the choice of the window function is not critical (refer to [Harris,1978] for further analysis on window functions). Usually, the window function most used in speaker recognition is the Hamming window (a specific case of the Hanning window), as it helps to produce a smooth estimate of power through regions where power changes rapidly. A generalised Hanning window can be characterised as:

$$w(n) = \begin{cases} \dfrac{\alpha_w - (1-\alpha_w)\cos\left(\dfrac{2\pi n}{N_f} - 1\right)}{\beta_w}; 0 \le n \le N_f \\ 0; otherwise \end{cases} \qquad \textbf{Eq. (2-55)}$$

where $N_f$, is the window length and $\alpha_w$ is defined as a window constant in the range of [0,1]. In the specific case of implementing a Hamming window, $\alpha_w$=0.54. $\beta_w$ is a normalisation constant defined as:

$$\beta_w = \sqrt{\frac{1}{N_s}\sum_{n=0}^{N_s-1} w^2(n)} \qquad \textbf{Eq. (2-56)}$$

From the windowed frame, the well-known Fast-Fourier Transform (FFT) – a fast implementation of the DFT – is applied in order to decompose the signal into its

frequency components. A deep review of the DFT theory is beyond the scope of this study; however the interested reader can find the definition and their properties in [Cooley,1969] where the FFT is also described. From a practical point of view, we can define the DFT of a finite sequence s[*n*] / *n*=0,...,*N-1* as the sequence S[*f*] / *f*=0,...,*N-1*, such that:

$$S[f] = \frac{1}{N}\sum_{n=0}^{N-1} s[n]e^{-j\left(\frac{2\pi nf}{N}\right)} \qquad \textbf{Eq. (2-57)}$$

The corresponding inverse transform can be defined as:

$$s[n] = \sum_{f=0}^{N-1} S[f]e^{j\left(\frac{2\pi nf}{N}\right)} \qquad \textbf{Eq. (2-58)}$$

As already said, the FFT can also be used as a computationally efficient method, under the constraint that the spectrum is to be evaluated at a discrete set of frequencies multiple of $f_s/N$ (where $f_s$ denotes the signal sampling frequency), to compute the spectrum of a signal [Cooley,1965]. Although different variants exist, probably the simplest FFT algorithm is the known as decimation in time. The main idea behind this algorithm is the fact that a DFT of an *N*-point sequence can be rewritten in terms of two *N/2*-points DFT. In this way, if *N* is a power of two, then it is possible to apply the following decomposition until we reach a single-point DFT.

The decomposition consists in splitting Eq. (2-57), where the normalisation factor has been omitted for practical reasons, into two terms, one with even indices and the other one with odd indices, yielding to the following expression:

$$S[f] = \sum_{n=0}^{N/2-1} s[2n]e^{-\frac{j2\pi(2n)f}{N}} + \sum_{n=0}^{N/2-1} s[2n+1]e^{-\frac{j2\pi(2n+1)f}{N}} \qquad \textbf{Eq. (2-59)}$$

Using the fact that: $e^{\frac{j2\pi 2}{N}} = e^{\frac{j2\pi}{(N/2)}}$, then we can rewrite the previous equation as:

$$\begin{aligned} S[f] &= \sum_{n=0}^{N/2-1} s[2n]e^{-\frac{j2\pi nf}{(N/2)}} + e^{-\frac{j2\pi f}{N}} \sum_{n=0}^{N/2-1} s[2n+1]e^{-\frac{j2\pi nf}{(N/2)}} \\ &= S_{11}[f] + e^{-\frac{j2\pi f}{N}} S_{12}[f] \end{aligned} \qquad \textbf{Eq. (2-60)}$$

where $S_{11}[f]$ and $S_{12}[f]$ are the N/2-point DFT of the even and odd terms of s, respectively.

Usually only the magnitude of the spectrum is retained, based on the facts that the phase has little perceptual information, and additionally because the global shape of the magnitude spectrum (known as spectral envelope) contains information about the resonance properties of the vocal tract which have excelled as one of the parts of the spectrum that provides more information in speaker recognition. Figure 2-28 provides a representation of both the FFT magnitude spectrum and the spectral envelope. The FFT spectrogram is composed mainly by horizontal bands of energy spaced by a common interval in frequency, which is the fundamental frequency *f0* or pitch. These are known as *harmonics* and convey information about the timbre of speech, which is ultimately related with the speaker's identity as well as with prosody.

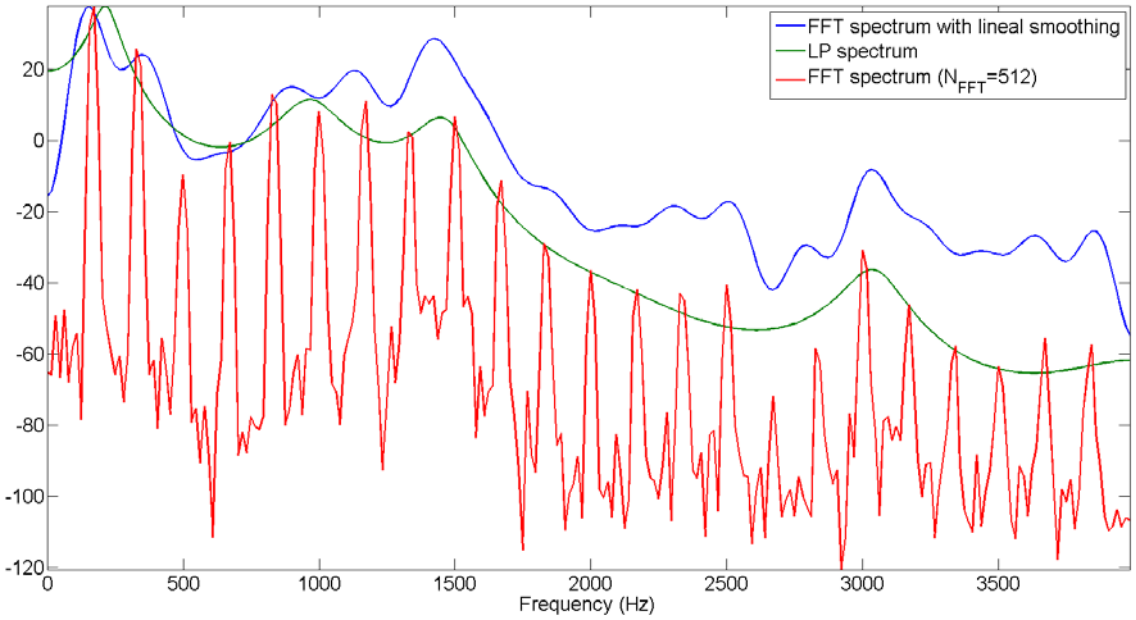



**Figure 2-28** Extraction of spectral envelope using cepstral analysis (blue solid line) and linear prediction (green solid line) from the FFT spectrum with N=512

A simple model of spectral envelope uses a set of band pass filters to do energy integration over neighbouring frequency bands. Probably one of the most popular filter banks used to smooth and get the envelope from the spectrum is the Mel-scale filter bank (see Figure 2-29 for an example of triangle shape Mel-scale filter bank). The Mel-scale is an auditory scale motivated by psycho-acoustic studies, where the lower frequency range is represented with higher resolution by allocating more filters with narrow bandwidth. More specifically, the Mel frequency scale is linear up to 1000Hz and logarithmic thereafter. For the specific filter bank, a set of overlapping Mel filters are made such that their centre frequencies are equidistant on the Mel scale.

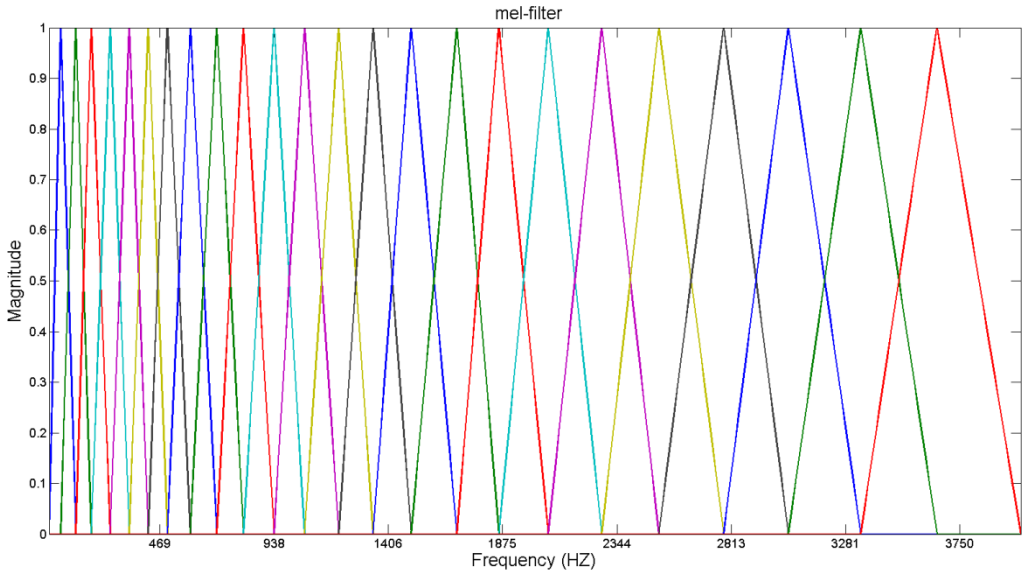



**Figure 2-29** Mel-scale filter bank with 24 triangle-shape filters

Another reason for the smoothing of the spectrum is the reduction of the size of the spectral vectors. For instance, if we multiply the spectrum previously obtained by the FFT process and plotted in Figure 2-28, by the filter bank shown in Figure 2-29, we get the spectral envelope shown in Figure 2-30 (solid blue line).

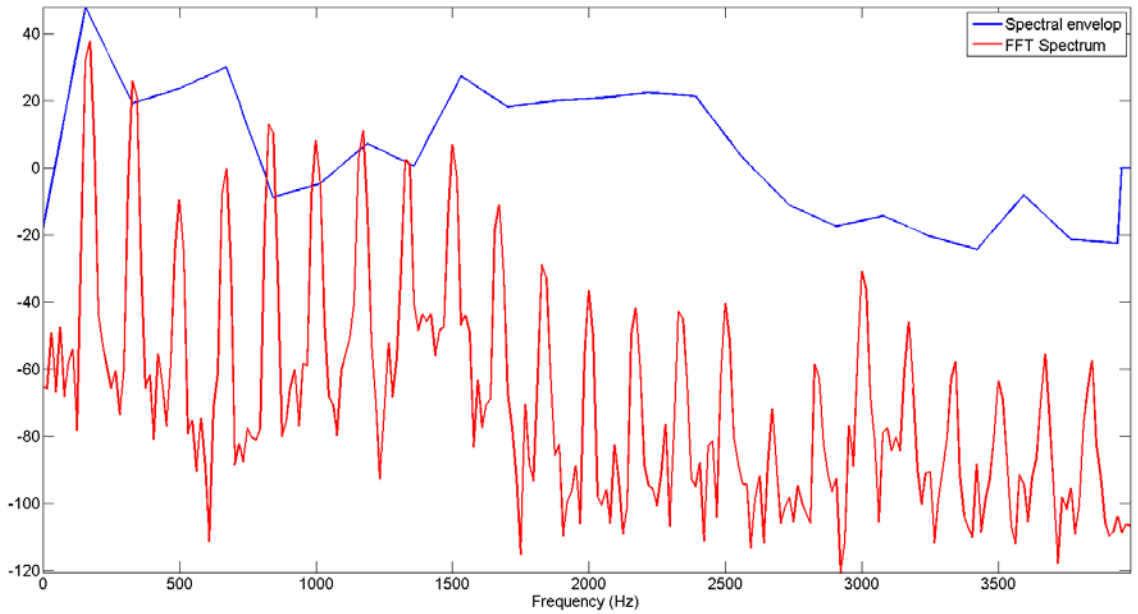



**Figure 2-30** Envelope spectrum using a 24-band Mel filter bank

Next, we take the log of this spectral envelope and multiply each coefficient by 20 in order to obtain the spectral envelope in dB. Finally an additional transformation is needed in order to obtain the cepstral coefficients, which is known as inverse DFT. However, as the logarithmic spectrum is a symmetric real function, the DFT can be replaced by the Discrete Cosine Transform (DCT) which only operates on the real part of the FFT. The DCT can be characterised as:

$$MFCC(i) = \sum_{j=1}^{P} m_j \cos\left(\frac{\pi i}{P}\left(j - \frac{1}{2}\right)\right); 0 \leq i \leq K \qquad \textbf{Eq. (2-61)}$$

where $P$ is the number of filters used in the filter bank, $K$ is the number of cepstrum coefficients and $m_j$ are the log-spectral coefficients. Only the first cepstral coefficients are used (typically K≤20). It should be noted that *MFCC(0)* in Eq. (2-61) represents the average value of the spectrum, or the root mean square (rms) value of the signal. As alternative measures of power are explicitly added to the final parameter vector, the $0^{th}$ MFCC is usually discarded.

Following this procedure, we finally obtain cepstral vectors for each analysis window. Once the cepstral coefficients have been computed, an additional processing step can be performed in order to remove the contribution of slowly varying convolutive noises form the cepstrum. This processing step, also known as cepstral mean subtraction (CMS) consists in subtracting the long-term average cepstral vector from each cepstral vector. An alternative processing method extensively used [Kinnunen,2009] is the RASTA (Relative Spectral) filtering [Hermansky,1994]. RASTA filtering which consists in applying a band-pass filter on the time series of each coefficient obtained from the spectral analysis, aims at reducing the distortions in the communication channel taking advantage of the fact that the rate of change of the short-term spectrum of non-linguistic components often lies outside the rate of change of the linguistic components. In other words, RASTA filtering suppresses the spectral components that vary more slowly or quickly than the typical range of change of speech. Regarding to CMS, which compares each analysis frame against the average of the whole utterance,

RASTA only uses a relatively short history of the signal to attenuate the low and high modulation frequencies.

- **LFCCs**

  An interesting variation to the MFCC is the linear frequency cepstrum coefficients also known as LFCC. These parameters seem to be more robust when dealing with telephone signals [Charbuillet,2007]. The main difference respect to MFCC approach is the filter bank used in the spectral envelope extraction process. In this case, the filter bank used is a linear scale filter bank, instead of the Mel-scale filter bank, in which the set of filters that define the filter bank are uniformly spaced over the frequency range of the speech signal, and additionally share the same bandwidth.

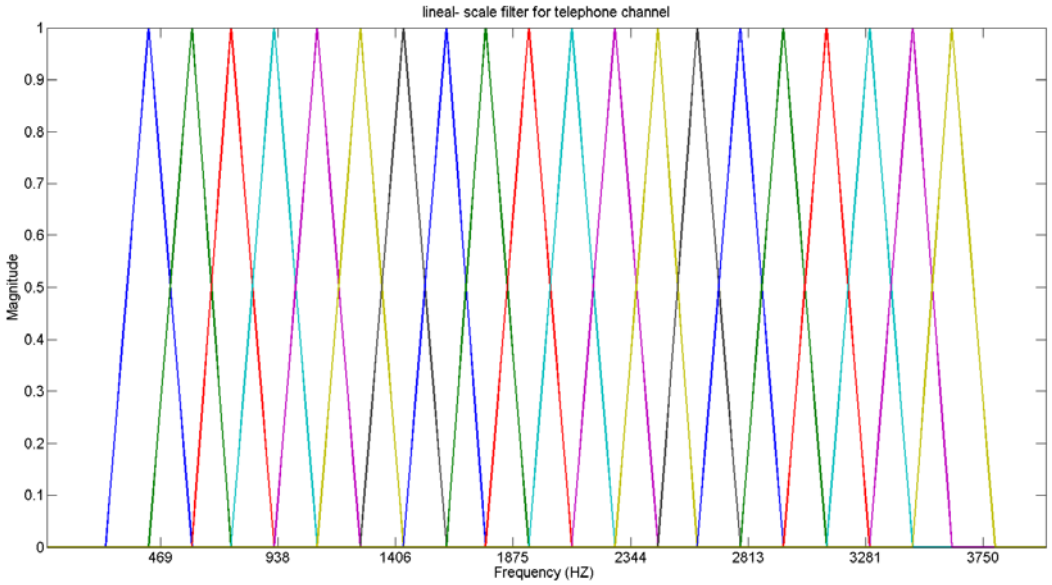



**Figure 2-31** Linear scale filter bank for telephone channel voice processing

Figure 2-31 shows an example of a typical linear scale filter bank for telephone speech signals in which frequencies below 250Hz and above 3600Hz are discarded.

- **On the use of dynamic information**

  Through the use of short-time spectral features, we are not taking into account the relation between neighbour windows or in other words, how these cepstral vectors vary over time. Dynamic information which plays a major role in perception and cognition of human speech [Gómez,2010], has been also used to improve the performance of both speech and speaker recognition systems [Furui,1986], [Soong,1988]. Regarding the speaker recognition area (and especially when dealing with cepstral coefficients), dynamic information has been classically introduced by using delta-cepstrum (Δ) and delta-delta cepstrum (ΔΔ). The delta-cepstrum and delta-delta-cepstrum are polynomial approximations of the first-order and second-order derivatives of the static cepstrum. From a practical point of view, they are computed as the time differences between neighbour vectors feature coefficients:

$$\Delta MFCC(i) = \frac{\sum_{k=-l}^{l} kMFCC(i+k)}{\sum_{k=-l}^{l} |k|} \qquad \textbf{Eq. (2-62)}$$

$$\Delta\Delta MFCC(i) = \frac{\sum_{k=-l}^{l} k\Delta MFCC(i+k)}{\sum_{k=-l}^{l} k^2} \qquad \textbf{Eq. (2-63)}$$

Additionally, the log energy and the Δlog energy can be also added as an additional feature. According to [Bimbot,2004], the former is often discarded while the latter is usually kept.

However, since these dynamic features do not have quite optimal performance when used alone, they are typically appended to the static cepstral features.

#### *2.4.1.2 Vocal Tract feature extraction*

The main objective when selecting a set of parameters to model the vocal tract is that it can capture the specificities of the vocal tract shape, regarded as a concatenation of tubes (see Figure 2-14). From this point of view, and taking into account that the only information available is the speech waveform, the most widely features used to model the vocal tract are those derived from the short-time spectral analysis of voice. The short-time spectral analysis aims at capturing the spectral envelope and thus the set of resonances (also known as formants) whose locations depend upon the vocal tract shape.

Since the source-tract separation algorithm not only provides a glottal residual but a trace almost free from the glottal influence and associated to the unit impulse response of the vocal tract, we will use this estimate to obtain the vocal-tract related features. Figure 2-32 depicts the vocal tract and glottal source estimates obtained from a female speech utterance of vowel /a/.

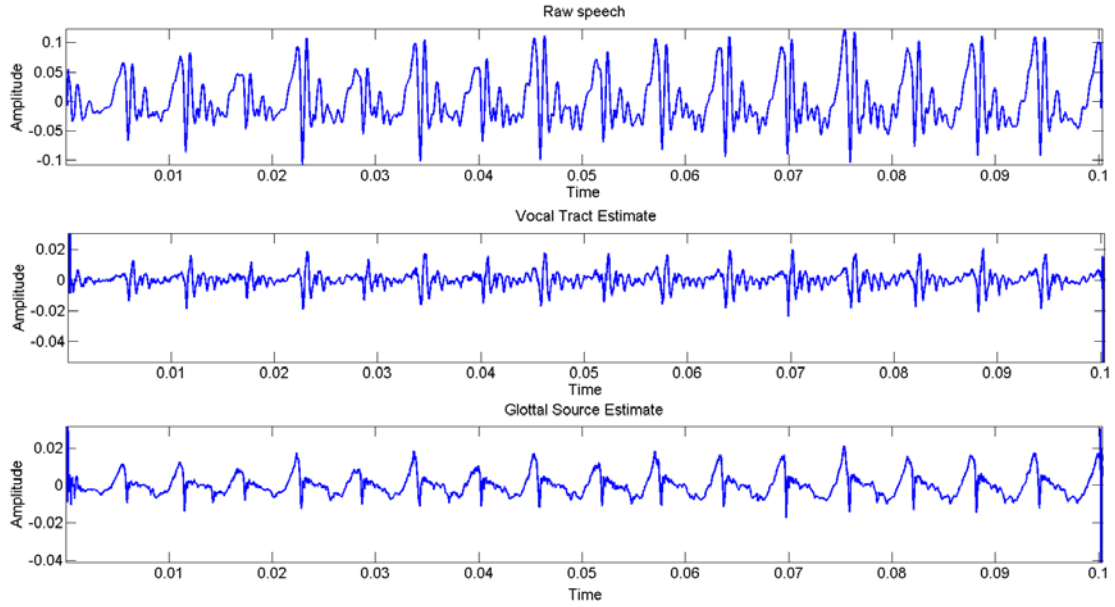



**Figure 2-32** Vocal tract (middle) and Glottal source (lower) estimates for a female vowel an utterance (upper)

This vocal tract estimate is particularly relevant in the accurate estimation the formants (resonances) of the vocal tract, as it is almost free from the influence of the glottal dynamics in the vocal tract pattern, as oppose to whether this estimate was made on the

original speech signal. Figure 2-33 shows how the formants, F1 and F2 located at 7000Hz and 1250Hz approximately, are clearly highlighted when the signal used is the vocal tract estimate.

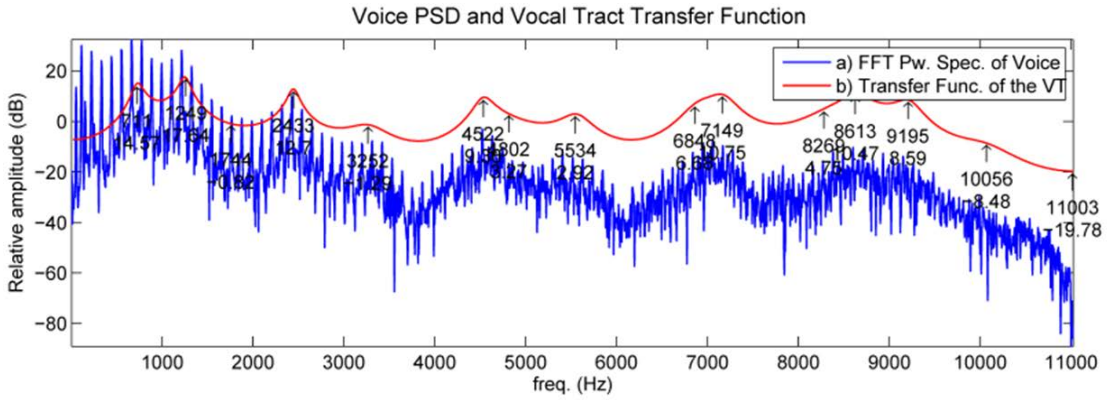



**Figure 2-33** FFT Power Spectrum of the original voice signal (solid blue line) and transfer function of the Vocal tract estimate (red solid line)

No matter whether the original speech signal or the vocal tract estimate are used to obtain the vocal tract related features, there are mainly two major approaches to this feature extraction problem: LP analysis and short-time Fourier transform (STFT). Depending on the selected approach, different types of features may be used to characterise the vocal tract. For instance, in the case of LP analysis, several derivations of the LP coefficients have been proposed: Reflection Coefficients (RC), Log Area Ratios (LAR) [Campbell,1997], [Pandey,2010], Line Spectral Pair Frequencies (LSP) [Kinnunen,2009], [Sahidullah,2010], etc. However the most prevalent parameters among them are the Linear Predictive Cepstral Coefficients (LPCC). In the case of the STFT approach, the prevalent features extracted are MFCC which can be extracted as described in the previous section (2.4.1.1).

Although in the present study we are not going to use the LP approach to extract the vocal tract related features, it is worth making a brief presentation of the LPCC parameters (as they are the most extensively used) and of the LSP parameters (as they are reaching some popularity in the last years and often show better performance than LPCC while retaining exactly the same information).

- LP Cepstral Coefficients:

  Figure 2-34 shows a block diagram representation of an LPC-based cepstral parameterisation.

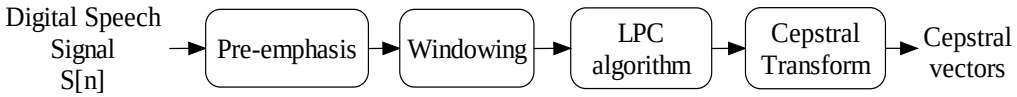



  **Figure 2-34** Block-diagram representation of an LPCC extraction procedure

  The extraction procedure as depicted in Figure 2-34 reflects certain similarity with the MFCC extraction process. More specifically, the pre-emphasis and the windowing steps are common in both extraction processes so the same explanation applies to this case. The next step differs from the MFCC process. In this case, LP analysis is used to estimate a set of predictive coefficients also known as LP coefficients. In section 2.3.1 a review of different LP methods

have been already made. This set of predictive coefficients can be used as a parameter vector; however, in this case, the cepstral coefficients are extracted thorough the cepstral transform block to build the LPCC vectors. The Cepstral transform can be implemented using the following recursion, taken into account that the set of prediction coefficients estimated in the previous block are {a1,…,$a_p$}, where p (which denotes the number of poles in the autoregressive filter) should be at least:

$$p \geq \frac{2 * f_c * l}{c} \qquad \textbf{Eq. (2-64)}$$

where $f_s$ is the sampling frequency, $l$ is the vocal tract length (typically of 17cm. in adult men while in adult woman is 14cm), and c is the speed of sound (speed of sound in cm. at 37 ºC and 1 atm. is 35400 cm/s)

$$LPCC_0 = \log(1) = 0\,;$$

$$LPCC_i = -a_i - \sum_{j=1}^{i-1} \left(1 - \frac{j}{i}\right) a_j LPCC_{i-j}\,; 1 \leq i \leq N_c \qquad \textbf{Eq. (2-65)}$$

Historically, $LPCC_0$ has been defined as the log power of the LP error. However, we can consider the cepstral model as a normalised model in which $LPCC_0$=0 [Picone,1993]. Regarding the number of cepstral coefficients, although an infinite number of cepstral coefficients can be computed, $N_c$ is usually comparable to the number of LP coefficients: $0.75p \leq N_c \leq 1.25p$.

- LSP

  The LSP representation originally proposed by Itakura [Itakura,1975], is a frequently used LP parametric representation. Figure 2-35 provides a modular representation of an LSP-based representation

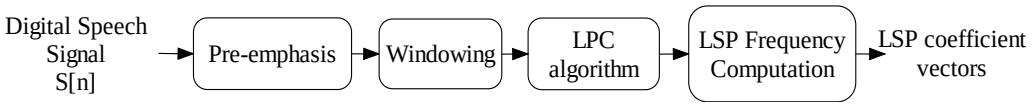



**Figure 2-35** Block-diagram representation of an LSP Coefficients extraction procedure

- The main difference compared to the LPCC extraction process lies in the last block which processes the LP coefficients to generate the LSP coefficients. We assume that the LP coefficients $\{a_i;\ 1 \leq i \leq p\}$ which define a linear time invariant all-pole filter, $H(z)=1/A(z)$ for a given frame are known; where p is the order of the $A(z)$ filter, characterised by:

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i} \qquad \textbf{Eq. (2-66)}$$

The LSP frequencies are representation of the predictor coefficients of the $A(z)$ filter. More precisely we can state that the LSP are the roots of the polynomials $P(z)$ y $Q(z)$ of order $p+1$, in which $A(z)$ is decomposed:

$$A(z) = \frac{P(z) + Q(z)}{2}$$
$$P(z) = A(z) - z^{-(p+1)}A(z^{-1}) \qquad \textbf{Eq. (2-67)}$$
$$Q(z) = A(z) + z^{-(p+1)}A(z^{-1})$$

Since the $(p+1)^{th}$ reflection coefficient is either -1 or 1, the order of the filter *A(z)* is extended without introducing any new information. In theory, by applying this decomposition, *P(z)* represents the specific vocal tract configuration in which the glottis is completely closed, while *Q(z)* represents the situation in which the glottis is open.

The polynomials *P(z)* and *Q(z)* present some relevant properties:

- The roots of *P(z)* and *Q(z)* occurs in symmetrical pairs, that is why these parameters receive the Line Spectrum Pairs (LSP) name.
- As *A(z)* is a minimum phase filter (with all its poles lying inside the unit circle in the z-plane), then in the presented decomposition all zeros of *P(z)* and *Q(z)* lie on the unit circle. The zeros of these polynomials can be represented only by their angles, as they lie on the unit circle.
- Zeros of *P(z)* and *Q(z)* are interlaced with each other, i.e. the LSP are in ascending order.

Figure 2-36 provides a graphical representation of the main properties of the LSP decomposition.

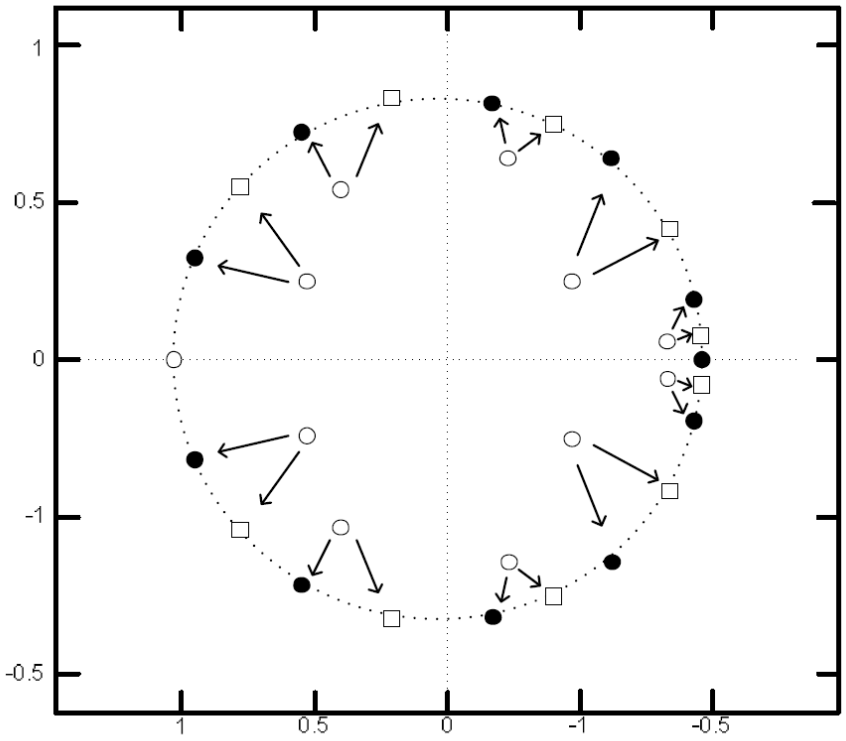



**Figure 2-36** Zeros of the LSP-process involved polynomials. White circles: original LP polynomial zeroes. Squares: *P(z)'s* zeroes. Filled circles: *Q(z)'s* zeroes (Extracted from [Onshaunjit,2008]).

The conventional methods for computing the roots of the polynomials may be complex root, real root [Lee,2006], ratio filter, Chebyshev series [Kabal,1986], and adaptive sequential LMS (Least Mean Square).

The LSF (Line Spectral Frequency) coefficients, as they are loosely related to the formant frequencies (more specifically, the vocal tract resonance frequencies fall between the two pairs of LSF frequencies), retain sufficient sensitive speaker dependent information. In the literature we can find several examples [Liu,1990], [Kim,2007], [Pandey,2010], [Sahidullah,2010] in which LSF parameters have been successfully applied to speaker recognition.

#### *2.4.1.3 Glottal source feature extraction*

As mentioned early, although glottal source contributes little to linguistic distinguishing of phonemes, it bears rich speaker specific information both in time and in frequency domain. This section provides a deep description of the different kind of features which can be extracted from the glottal source. Prior to the presentation of the different sets of features that can be extracted from the glottal source, it is convenient to introduce a new concept, the Power Spectral Density Profile (PSD profile), as it will be used to extract different parameters from the glottal source.

##### 2.4.1.3.1 Power Spectral Density Profile of the glottal source

It refers to the envelope of the power spectral density of the glottal source. If $s_g(n)$ represents the glottal source, then its Discrete Fourier Transform will be defined as:

$$S_g(m) = \sum_{n=0}^{N-1} S_g(n) e^{jm\Omega n\tau} \qquad \textbf{Eq. (2-68)}$$

where n represents the temporal index of the vector $s_g(n)$ inside a temporal window of $N$ samples, $0 \leq n \leq N-1$, taken every $\tau$ sec. The frequency index is given by the integer variable m, which corresponds to an impulse given by $m\Omega$, with frequency resolution $\Omega$.

$$\Omega = \frac{f_s}{2N}; f_s = \frac{1}{\tau}; 0 \leq m \leq \frac{N}{2} - 1 \qquad \textbf{Eq. (2-69)}$$

where $f_s$, represents the sampling frequency and $j$ denotes the imaginary unit. Under these assumptions the power spectral density of the glottal source will be represented by:

$$T_g(m) = \left\| S_g(n) \right\|^2 \qquad \textbf{Eq. (2-70)}$$

Figure 2-37 depicts the power spectral density of the glottal source. If we evaluate the power spectral density synchronously with a phonation cycle:

$$N_{k-1} \leq n \leq N_k - 1 \qquad \textbf{Eq. (2-71)}$$

where $N_k$ represents the start of the closure on the $k^{th}$ phonation cycle inside a fragment of voice given in dB:

$$T_g(m)_{dB} = 20log_{10}\|S_g(m)\| \quad \textbf{Eq. (2-72)}$$

We will obtain an estimate of power spectral density harmonic envelope as depicted in Figure 2-11.

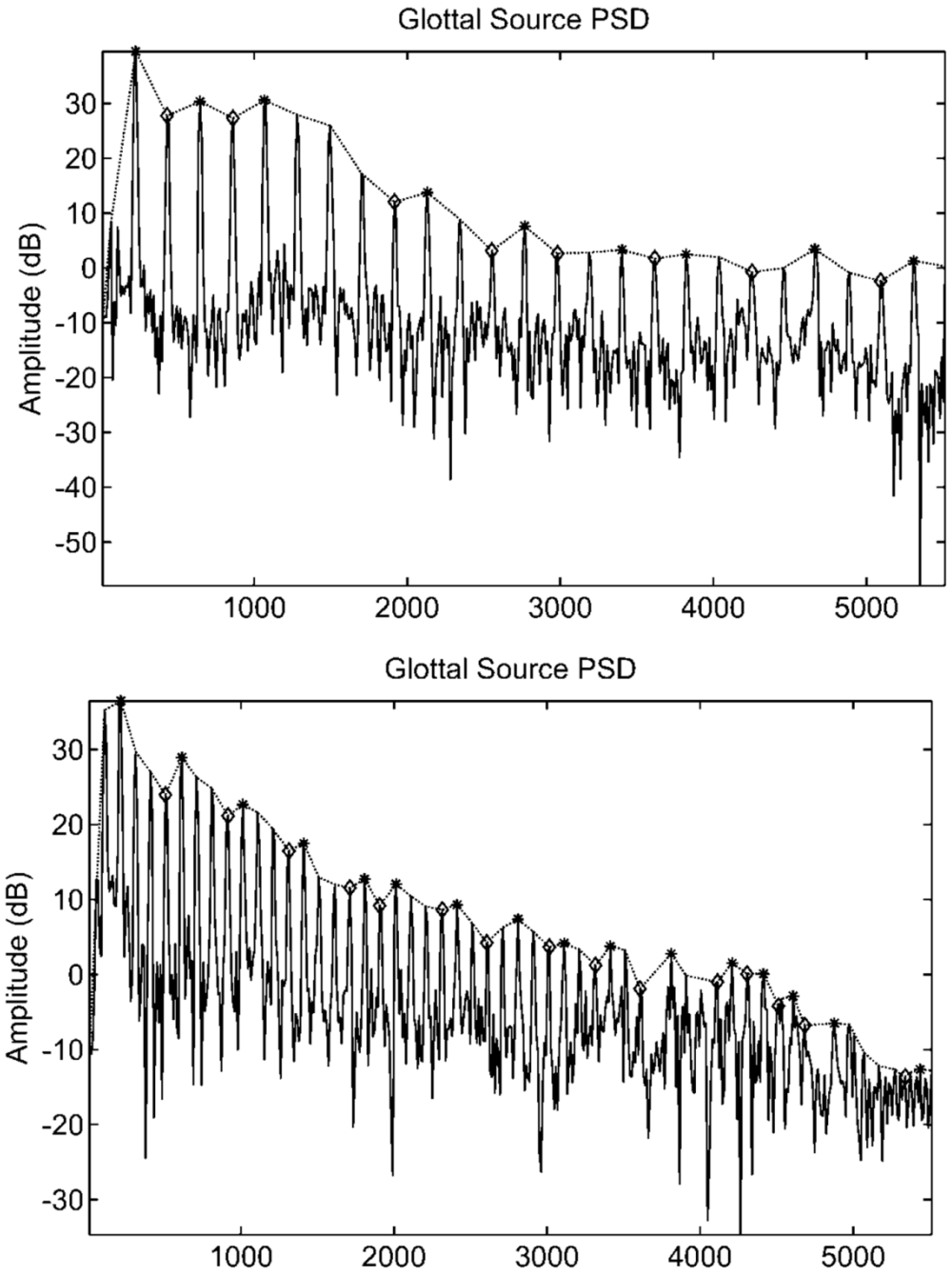



**Figure 2-37** Power spectral density of the glottal source evaluated over a temporal window which includes multiple glottal cycles. Upper – male voice. Lower – female voice. The relative maxima of the distribution are marked by the harmonics present in the signal. The interconnection of these maxima is known as Harmonic Envelope or Power Spectral Density Profile

Once the PSD profile of the glottal source has been presented, we continue with the glottal source feature extraction process.

- Biometric features:

  The biometric parameters of the glottal source are composed of the values and positions in frequency domain of the singularities on the PSD profile in dB (PSDdB). The first four maxima peaks of the PSDdB will be characterised by the pairs $\{T_{Mi}, f_{Mi}\}$; $1 \leq i \leq 4$:

$$T_{Mi} = \max_{m \in W_i} \{T_g(m)\} \quad \textbf{Eq. (2-73)}$$

$$f_{Mi} = \frac{f_s}{2N} arg\left[\max_{m \in W_i} \{T_g(m)\}\right] \quad \textbf{Eq. (2-74)}$$

  where $W_i$ is the fragment of the spectrum over which the relative maximum is estimated. No normalisation is applied to the first maximum, what is more, it is used in the normalisation of the others maxima and minima, as $\tau_{M1}$=0 and $\varphi_{M1}$=1. The equivalent normalised parameters of 2-4 maxima are:

$$\tau_{Mi} = T_{Mi} - T_{M1} \quad \textbf{Eq. (2-75)}$$

$$\varphi_{Mi} = \frac{f_{Mi}}{f_{M1}} \quad \textbf{Eq. (2-76)}$$

  Correspondingly, the first and second minima will be defined by the pairs $\{T_{mi}, f_{mi}\}$; $1 \leq i \leq 2$:

$$T_{mi} = \min_{m \in W_i} \{T_g(m)\} \quad \textbf{Eq. (2-77)}$$

$$f_{mi} = \frac{f_s}{2N} arg\left[\min_{m \in W_i} \{T_g(m)\}\right] \quad \textbf{Eq. (2-78)}$$

  where $W_i$ is the fragment of the spectrum over which the relative minimum is estimated.

  The resulting list of biometric features that can be extracted from the glottal source is as follows:

  - $p_{G1}$: Maximum value of the PSDdB
  - $p_{G2}$: Value of the first minimum after the maximum related to this maximum in dB
  - $p_{G3}$: Value of the second maximum of the PSDdB related to the first maximum.
  - $p_{G4}$: Value of the second minimum of the PSDdB related to the first maximum.
  - $P_{G5}$: Value of the third maximum of the PSDdB related to the first maximum.

- $P_{G6}$: PSDdB value at the maximum Nyquist value related to the first maximum.
- $p_{G7}$: Position on the frequency axis of the first minimum after the first maximum related to the position of the maximum.
- $p_{G8}$: Position on the frequency axis of the second maximum related to the position of the first maximum.
- $p_{G9}$: Position on the frequency axis of the second minimum related to the position of the first maximum.
- $p_{G10}$: Position on the frequency axis of the third maximum related to the position of the first maximum.
- $p_{G11}$: Position on the frequency axis of the end to the Nyquist frequency related to the position of the first maximum.
- $p_{G12}$:Slenderness factor of the first "V" profile, which is characterised by the first maximum, the first minimum and the second maximum, and can be defined a:

$$\sigma_{m1} = \frac{f_{Mm(2T_{m1}-T_{M2}-T_{M1})}}{2(f_{M2}-f_{M1})} \qquad \textbf{Eq. (2-79)}$$

- $p_{G13}$:Slenderness factor of the second "V" profile, which is characterised by the third maximum, the second minimum and the fourth maximum, and can be defined a:

$$\sigma_{m2} = \frac{f_{Mm(2T_{m2}-T_{M3}-T_{M4})}}{2(f_{M3}-f_{M4})} \qquad \textbf{Eq. (2-80)}$$

- Biomechanical features:

  The estimation of the biomechanical parameters of the vocal fold body is close related with the inversion of the integral-difference equations from the 1-mass model given by Eq. (2-2), through the synthesis of the equivalent elements from the PSD profile of the average acoustic wave (AAW) in a cycle-by-cycle evaluation basis. Figure 2-38 depicts a real case in which the decomposition of glottal source in its AAW component and the MWC (as described in Eq. (2-49) to Eq. (2-52)) is applied. An estimate of the three parameters involved in Eq. (2-2) can be obtained indirectly, just by performing an optimum fit of the PSD profile of the AAW by an equivalent electromechanical RLC (see Figure 2-39). From a practical point of view, the major difficulty when applying this process is the establishment of a robust and accurate practical method for profile fitting.

  Reliable estimates of the relative values of vocal fold body masses and tensions can be obtained from the PSD of the AAW, as it has been shown in [Gómez,2005-A]. The estimation process consists in performing an adaptive fitting of the AAW PSD against the transfer function of the 1-mass model. The work hypothesis is based on the assumption that the AAW is determined by the vocal fold dynamic component; therefore the envelope of the AAW PSD of the body dynamic component is directly related with the square modulus of the input admittance derived from the 1-mass model as:

$$T_b(\omega) = |Y_b|^2 = \left|\frac{V_x(\omega)}{F_x(\omega)}\right|^2 = [(\omega M_b - \omega^{-1}K_b)^2 + R_b^2]^{-1} \qquad \textbf{Eq. (2-81)}$$

where $M_b$, $K_b$ and $R_b$ are respectively the parameters associated with the lumped mass, elasticity and looses of the 1-mass model when only the body of the vocal fold is taken into account following the dimensionality reduction of the Story-Titze's model. $F_x(\omega)$ and $V_x(\omega)$ are the short-time estimates of the Fourier Transforms' of the speed and force dynamic variables on the x-axis, respectively.

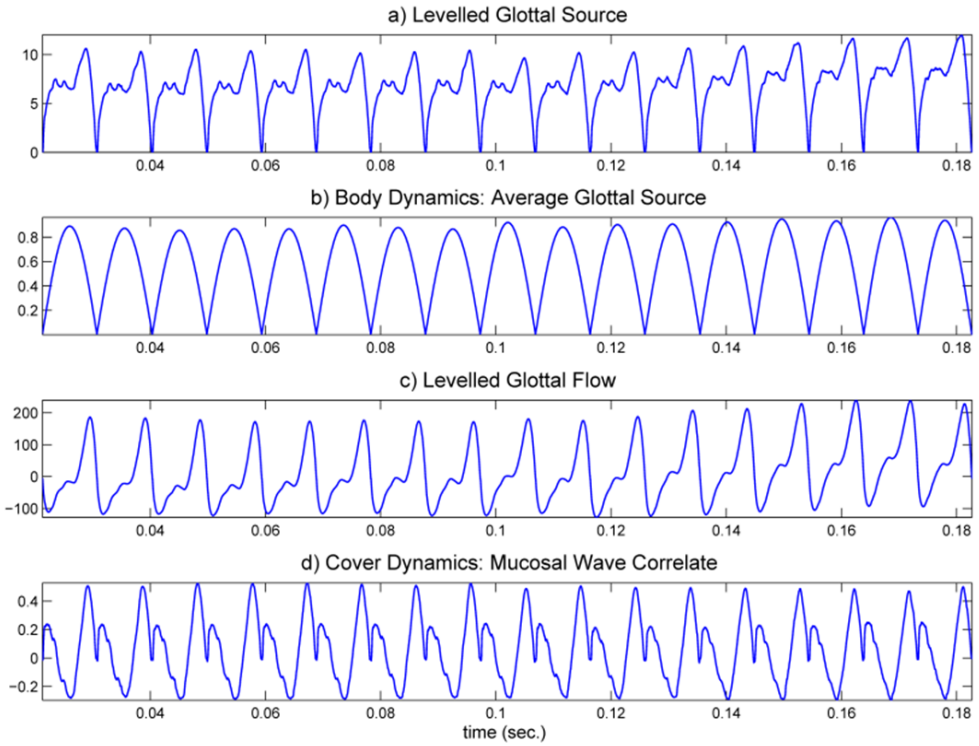



**Figure 2-38** AAW and MWC real case estimation: a) Levelled Glottal Source. b) Average Acoustic Wave also known as Average Glottal Source. c) Levelled Glottal Flow. d) Mucosal Wave Correlate

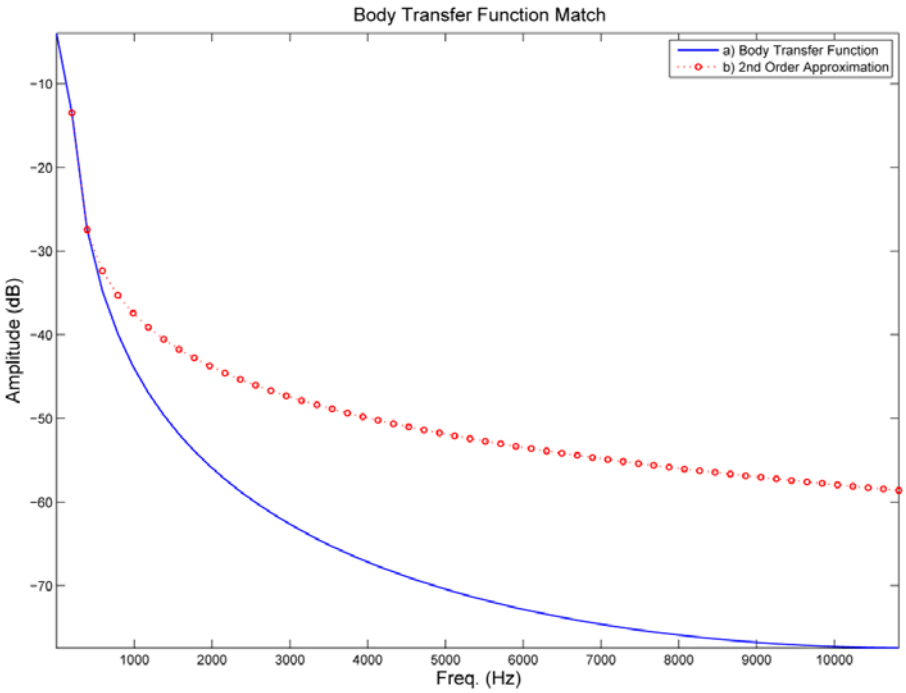



**Figure 2-39** Approximation of the PSD profile of the AAW (solid line) through the transfer function of a second order electromechanical equivalent (RLC: dotted line). As it is shown, both curves match almost completely until -30dB.

The $M_{bl,r}$, $K_{bl,r}$ and $R_{bl,r}$ parameters are estimate in the following way:

- Mass parameter estimate:

  The robust estimation of the model parameters is based in the precise determination of two points in the PSD of the dynamic component, these being $\{T_{b1},\ \omega_1\}$ and $\{T_{b2},\ \omega_2\}$. From these points the dynamic mass of the vocal fold body can be estimated as:

$$M_b = \frac{\omega_2}{\omega_2^2 - \omega_1^2}\sqrt{\frac{T_{b1} - T_{b2}}{T_{b1}T_{b2}}} \qquad \textbf{Eq. (2-82)}$$

  The selection of the most adequate points for $\{T_{b1},\ \omega_1\}$ and $\{T_{b2},\ \omega_2\}$ is highly related with the robustness of the estimation procedure. A good candidate for $\{T_{b1},\ \omega_1\}$ is the position of the main peak (resonant peak or glottal formant) in the amplitude of the PSD of the AAW. The peak position is apparently easy to determine, but is contaminated by the sampling resolution in the frequency domain of the Short-Time Discrete Fourier Transform (STDFT). The use of the STDFT is required if a cycle-by-cycle study is to be performed. Increasing the duration of the analysis window to several cycles would increase the resolution, but at the cost of concealing the imbalances and temporal variability of the parameters. An alternative to this would consist in interpolating the STDFT by splines.

  A good candidate for $\{T_{b2},\ \omega_2\}$ is the position of the third harmonic from the peak position, as the time series shows odd symmetry. These points

have shown to be robust enough in the studied cases, although other strategies are also possible.

- Elasticity parameter estimate:

  Once the body mass has been estimated, the elastic parameter (body stiffness) $K_b$, may be obtained from the precise determination of the position of the maximum (assuming this to be associated to the resonant peak) this being $\{T_r, \omega_r\}$, as:

$$K_b = M_b \omega_r^2 \qquad \textbf{Eq. (2-83)}$$

  where $T_r$ stands for the value of the square modulus of the input admittance in Eq. (2-81) at the frequency of resonance $\omega_r$ associated to the first maximum in the glottal source PSD.

- Loss parameter estimate:

  The parameter of body losses (which includes the non-conservative effects due to the viscosity of the fluids, to the turbulence of the air flow through the acoustic tube and to the inelastic collision of the folds) can be estimated (but for a scale factor $G_b$) as:

$$R_b = \frac{G_b}{\sqrt{T_r}} \qquad \textbf{Eq. (2-84)}$$

  The estimate of $G_b$ is performed by normalising the glottal source correlate with respect to the maximum amplitude of the AAW over all cycles of phonation under analysis. The value of $G_b$ is only relevant when the goal is to obtain absolute estimates of the parameters under analysis.

- Biomechanical parameter unbalance estimate:

  So far we have considered that both vocal folds were completely symmetrical. However, this assumption does not stand in most of the cases, as clear deviations are found in parameter estimates between neighbour phonation cycles that can only be explained by an asymmetry in the vocal folds. This unbalance naturally occurs both in normophonic and in dysphonic voices, although it seems reasonable to think that the unbalance will be larger in cases where voice pathology is present. Moreover, it is generally accepted that the presence of unbalance, which is related in certain way with *jitter* and *shimmer*, is a correlate to vocal fold pathology.

  No matter whether the root of the unbalance is physiological, psychological or dynamic, it will leave a fingerprint on the biomechanical parameter estimates in a cycle-by-cycle evaluation. Unbalance between neighbour phonation cycles may be appreciated in Figure 2-38 where the dynamic variables represented tend to vary slightly. The estimations of mass, stiffness and losses are produced on a phonation cycle-frame basis, therefore the (intra-speaker) unbalance of these parameters ($\mu_b$: Body Mass Unbalance; $\sigma_b$: Body Losses Unbalance; $\gamma_b$: Body Stiffness Unbalance) may be defined as:

$$\mu_{bn} = \left(\widehat{M}_{bn} - \widehat{M}_{bn-1}\right) \Big/ \left(\widehat{M}_{bn} + \widehat{M}_{bn-1}\right) \qquad \textbf{Eq. (2-85)}$$

$$\rho_{bn} = \left(\hat{R}_{bn} - \hat{R}_{bn-1}\right) \Big/ \left(\hat{R}_{bn} + \hat{R}_{bn-1}\right) \qquad \textbf{Eq. (2-86)}$$

$$\gamma_{bn} = \left(\widehat{K}_{bn} - \widehat{K}_{bn-1}\right) \Big/ \left(\widehat{K}_{bn} + \widehat{K}_{bn-1}\right) \qquad \textbf{Eq. (2-87)}$$

where $1 \leq n \leq N$, is the phonation cycle window index, while $\widehat{M}_{bn}$, $\hat{R}_{bn}$ and $\widehat{K}_{bn}$ are the *n*th cycle estimates of mass, losses and stiffness on a given voice sample (for a single specific subject).

- Mucosal wave correlate biomechanical parameters:

  Similar derivations (for mass, elasticity and looses) may be defined for the biomechanical parameters of the vocal fold cover, using in this case the MWC. As long as the influence of the body dynamics has been removed implicitly on separating the AAW from the MWC, then the problem is reduced to a single mass model, thus easing the application of the same methodology. The biomechanical parameters of the MWC are estimated by approximating the PSD of the MWC to the transfer function of an RLC circuit whose elements are in this case $M_{cl,r}$, $K_{cl,r}$ and $R_{cl,r}$

The resulting list of biomechanical features extracted from the glottal source and the mucosal wave correlate is as follows:

- $p_{G14}$*:* Equivalent dynamic mass of the left (l) / right (r) vocal fold's body, according to Figure 2-9. ($p_{G14} = M_{bl} = M_{br}$).
- $p_{G15}$*:* Parameter related to the losses due to frictions, viscosity and inelastic behaviour of the vocal fold's body, according to Figure 2-9. ($p_{G15} = R_{bl} = R_{br}$).
- $p_{G16}$*:* Equivalent elastic parameter of the left (l) / right (r) vocal fold's body, according to Figure 2-9 ($p_{G16} = K_{bl} = K_{br}$).
- $p_{G17}$*:* Body mass unbalance parameter ($p_{G17} = \mu_{bl} = \mu_{br}$).
- $p_{G18}$*:* Body losses unbalance parameter ($p_{G18} = \rho_{bl} = \rho_{br}$).
- $p_{G19}$*:* Body stiffness unbalance parameter ($p_{G19} = \gamma_{bl} = \gamma_{br}$).
- $p_{G20}$*:* Equivalent dynamic mass of the left (l) / right (r) vocal fold's cover, according to Figure 2-9. ($p_{G20} = M_{cl} = M_{cr}$).
- $p_{G21}$*:* Parameter related to the losses due to frictions, viscosity and inelastic behaviour of the vocal fold's cover, according to Figure 2-9. ($p_{G21} = R_{cl} = R_{cr}$).
- $p_{G22}$*:* Equivalent elastic parameter of the left (l) / right (r) vocal fold's cover, according to Figure 2-9 ($p_{G22} = K_{cl} = K_{cr}$).
- $p_{G23}$*:* Cover mass unbalance parameter ($p_{G23} = \mu_{bl} = \mu_{br}$).
- $p_{G24}$*:* Cover losses unbalance parameter ($p_{G24} = \rho_{bl} = \rho_{br}$).

- $p_{G25}$: Cover stiffness unbalance parameter ($p_{G25} = \gamma_{bl} = \gamma_{br}$).

NOTE: According to Figure 2-9, $K_{bcl,r}$ stands for the equivalent elasticity parameter between the vocal fold's cover and body (l→ left vocal fold and r→right vocal fold). However this elasticity parameter is not usually used as a biomechanical parameter. Instead, if we consider the vocal folds to be completely symmetric, then it will be sufficient to estimate the parameters described above (i.e., mass, loss and elasticity for both cover and body and its related unbalances).

- Temporal parameterisation over a phonation cycle.

The temporal parameterisation is based on three signals extracted from the glottal source (as described in section 2.3.2): the average acoustic wave (AAW – $s_g(n)$), the mucosal wave correlate (MWC – $s_c(n)$) and its derivative ($s_d(n)$). This last signal defined as:

$$s_{dk}(n) = \frac{s_{ck}(n) - s_{ck}(n-1)}{\tau} \qquad \textbf{Eq. (2-88)}$$

where k is the phonation cycle index under study.

The estimation is performed for each phonation cycle, and synchronous with its start and end, defined from minimum to minimum of the glottal source (clipping point or MFDR: Minimum Flow Declination Rate) as depicted in the lower representations of Figure 2-24 and Figure 2-25. Figure 2-7 provides a representation of the signals referenced above as well as classical points of reference:

- Glottal closure instant: This is the point at which the glottal source reaches its minimum value in the form of a negative sharp peak (MFDR, point 7 at Figure 2-5). This peak has its origin in the depression produced in the supraglottic area by the abrupt interruption of flow, while the column of air in the vocal tract follows its outgoing movement thanks to its inertial behaviour. It is taken as the origin of the glottal source cycle, $t$=0.
- Recovery instant: t=$T_r$ → Corresponds to the point in which the partial reversal of the air column in the vocal tract sets back the supraglottic pressure to reference or atmospheric pressure. It is located between points 8 and 9 at Figure 2-5.
- Start opening instant: t=$T_o$ → It is the point at which the vocal folds open again. It corresponds to point 3 at Figure 2-5.
- Start closure instant: t=$T_{cl}$ → Point at which the maximum flow is reached, if we manage to remove the influence of the vocal tract, the aperture between folds begins to decrease (points 4 and 5 at Figure 2-5.)
- Final glottal cycle instant: t=$T_c$ → At this point the minimum supraglottic pressure is reached; therefore it denotes the beginning of a new cycle.

However, the set of parameters that can be used and depicted, for a specific phonation cycle, in the upper representations of Figure 2-24 and Figure 2-25 are:

- $t=T_{d0} \rightarrow$ Instant at which the MWC derivative reaches its maximum in quadrant A.

$$T_{d0} = arg\left[\max_{0\leq n\leq n_0}\{s_{dk}(n)\}\right] \quad \textbf{Eq. (2-89)}$$

- $t=T_{mwc1} \rightarrow$ Instant at which the MWC reaches its maximum in quadrant A.

$$T_{mwc1} = arg\left[\max_{0\leq n\leq n_0}\{s_{ck}(n)\}\right] \quad \textbf{Eq. (2-90)}$$

- $t=T_{d1}=t_C \rightarrow$ Instant at which the MWC derivative reaches its minimum in quadrant A.

$$T_{d1} = arg\left[\min_{0\leq n\leq n_0}\{s_{dk}(n)\}\right] \quad \textbf{Eq. (2-91)}$$

- $s_{gB} \rightarrow$ Maximum value of the glottal source $u_g(n)$ in quadrant B.
- $T_{maxB} \rightarrow$ Instant at which the glottal source reaches its maximum in quadrant B.

$$T_{maxB} = arg\left[\max_{0\leq n\leq n_0}\{u_{gk}(n)\}\right] \quad \textbf{Eq. (2-92)}$$

- $t=T_{mwc2} \rightarrow$ Instant at which the MWC reaches its minimum.

$$T_{mwc2} = arg\left[\min_{0\leq n\leq N_k}\{s_{ck}(n)\}\right] \quad \textbf{Eq. (2-93)}$$

- $t=T_{d2}=t_O \rightarrow$ Instant of zero crossing of the MWC derivative previous to d3, which divides the phonation cycle in four quadrants: A, B, C, D (both in time and value). B and C are the quadrants above the O value, before and after instant O respectively; while A, and D are the quadrants below the O value, before and after instant O respectively.
- $t=T_{d3} \rightarrow$ Instant at which the MWC derivative reaches its maximum in quadrant D.
- $s_{gC} \rightarrow$ Maximum value of the glottal source $u_g(n)$ in quadrant C.
- $T_{maxC} \rightarrow$ Instant at which the glottal source reaches its maximum in quadrant C.

$$T_{maxC} = arg\left[\max_{n_0\leq n\leq N_k}\{u_{gk}(n)\}\right] \quad \textbf{Eq. (2-94)}$$

- $t=T_{sC} \rightarrow$ Instant at which the glottal source falls below the O value in quadrant C.

$$T_{sC} = \min_{0 \le n \le N_k} \left| arg\{u_{gk}(n) < u_{g0}\} \right| \quad \textbf{Eq. (2-95)}$$

- $t=T_{mwc3} \rightarrow$ Instant at which the MWC reaches its maximum in quadrant D.

$$T_{mwc3} = arg\left[ \max_{n_0 \le n \le N_k} \{s_{ck}(n)\} \right] \quad \textbf{Eq. (2-96)}$$

- $EA \rightarrow$ Aperture or injection efficiency. This parameter estimates the average flow in the glottal aperture phase (from $T_0$ to $T_{cl}$)

$$EA = (T_{mwc2} - T_{mwc1}) arg\left[ \max_{0 \le n \le n_0} \{u_{gk}(n)\} \right] \quad \textbf{Eq. (2-97)}$$

- $DC \rightarrow$ Closure or escape deficiency. This parameter estimates the average flow in the glottal closure phase (from $T_r$ *to* $T_0$).

$$DC = (T_{mwc3} - T_{mwc2}) arg\left[ \max_{n_0 \le n \le N_k} \{u_{gk}(n)\} \right] \quad \textbf{Eq. (2-98)}$$

The resulting list of temporal parameters extracted over a phonation cycle is as follows:

- $p_{G26} = \frac{T_{mwc2}}{T_c}$
- $p_{G27} = \frac{T_{sC}}{T_c}$
- $p_{G28} = \frac{T_{d1}}{T_c}$
- $p_{G29} = \frac{T_{mwc1}}{T_c}$
- $p_{G30} = \frac{T_{mwc3}}{T_c}$
- $p_{G31} = \frac{T_{d0}}{T_c}$
- $p_{G32} = \frac{T_{maxB}}{T_c}$
- $p_{G33} = \frac{T_{d3}}{T_c}$
- $p_{G34} = \frac{T_{maxC}}{T_c}$
- $p_{G35} = EA$
- $p_{G36} = DC$
- $p_{G37} = EA - DC$

- Spectral parameterisation of the glottal source.

  Besides the set of parameters already presented, an alternative set of parameters which aims to capturing the frequency-domain characteristics of the glottal source are the Mel-frequency cepstral coefficients (or alternatively LFCCs). The

estimation of these MFCC/LFCC parameters follows the same processing steps as the ones already presented in 2.4.1.1, except that in this case the input signal is no longer the speech signal, but the glottal residual obtained in the source-tract separation process.

The main goal in using this set of features for speaker characterisation is the fact that it can be easily integrated into what we have called state-of-the-art speaker recognition system, i.e. UBM-GMM based speaker recognition system whose feature parameters are MFCC extracted from the voice signal. Additionally, by using the same system but changing the set of features that characterise each speaker, a comparison in terms of recognition accuracy can be performed, thus leading to a better understanding of the discriminative power of the glottal source. Last but not least, temporal, biomechanical and biometric parameters will be subject to greater distortion if a good estimation of the AAW is not reached, what is more likely to happen under no-supervised processing. Thus the spectral based coefficients may show a better performance under non-supervised conditions.

## 2.5 ALTERNATIVE PARAMETERISATIONS

Although not used in early experiments in speaker recognition, in the course of time many researchers have adopted glottal information to improve recognition rates. However, none of them have previously used, neither the specific algorithm described here for source-tract separation nor the set of features precisely describe for the glottal source characterisation. Moreover, based on literature, it has not been until more recently when the fusion of these complementary features has been applied in order to improve accuracy.

In section 2.3.3 a review of alternative source-tract separation algorithms have already been made. Therefore in this section we will review different parameterisation approaches used to characterise a speaker making use of glottal source information.

The first approaches that make use of glottal information propose the use parametric glottal flow model parameters. [Brookes,1994] proposed a modification of the LF model which includes an additional parameter that controls the skewness of the positive portion of the first difference of the glottal airflow waveform. From this model, four dimensionless parameters are extracted: the ratio of the peak rise and fall slopes of glottal airflow waveform, the fraction of the cycle corresponding to the open phase, the ratio of fall to rise time, and the vocal fold closure time as a fraction of the cycle. This set of features is supplemented with the frequencies and bandwidths of the first two formants, the pitch and the speech energy. Although the experiments were conducted using a limited set of speakers (10 – 5 male and 5 female) and therefore difficult to compare with other studies, they report certain improvement in recognition rates when the four glottal waveform parameters are used. [Plumpe,1999] presented a more comprehensive analysis of different types of glottal source based parameters using TIMIT and NTIMIT database. In this case the glottal flow has been parameterised both using a parametric approach and MFCC approach. In the parametric approach, twelve different parameters have been proposed and combined to improve the accuracy of the recognition system. Seven of these parameters are based on the LF coarse structure model, while the other five refer to the energy computed over 5 different within phonation cycle intervals:

- The time of glottal opening ($T_o$)

- The time of the maximum negative value of the glottal pulse ($T_e$)
- The time of glottal closure ($T_c$)
- An exponential time constant which determines how quickly the flow derivative returns to zero after time, $T_e(\beta)$
- The value of the flow derivative at time $T_e$ ($E_e$).
- Factor that determines the ratio of $E_e$ to the peak height of the positive portion of the glottal flow derivative.
- Frequency that determines flow derivative curvature to the left of the glottal pulse.
- 5 Energy measures evaluated over the following intervals:
    - Closed phase for the LF model
    - Open phase for the LF model
    - Return phase for the LF model
    - Closed phase for formant modulation
    - Open phase for formant modulation

In the case of cepstrum approach, 23MFCC have been computed from the glottal flow derivative waveform and from the waveform synthesised using the LF modelled glottal flow derivative. This set of features has been tested both alone and combined with LPC coefficients derived from the voice signal. When used alone, the results show that the set of features based on glottal information show some discriminative ability, whereas when used in conjunction with LPC coefficients, the recognition rates are again improved in the order of 5% on *EER*.

An interesting review on different parameterisation methods both on time and in frequency-domain, which have been applied to the estimated glottal flow or its derivative obtained by inverse filtering, can be found in [Alku,2003]. [Arroabarren,2003] also present a comparative study between two different glottal source parameterisation methods in time and frequency domain: Normalised Amplitude Quotient (NAQ) and Parabolic Spectral Parameter (PSP). More recently, different parameterisation techniques have been applied to the glottal source, ranging from residual phase of the glottal source, through wavelet analysis, cepstral parameters or even higher-order statistics.

In [Murty,2006] the MFCC extracted from voiced segments of the speech signal are combined with residual phase information, defined as the cosine of the phase function of the analytic signal derived from the LP residual of the speech signal, and extracted around glottal closure instants. These features are used to train two-different auto-associative neural networks (one for the MFCC and another for the residual phase) whose scores are linearly weight-combined. The experiments conducted on the NIST-2003 database show a significant improvement regarding recognition rates respect to MFCC features alone. However, the results are gender-bias as only male speakers were used in the verification studies.

Source Cepstral parameters have been also used by [Gudnason,2008] and [Kinnunen,2009] using two different approaches. The first computed three different sets of parameters from the voice signal: MFCC derived from the magnitude spectrum of

speech, the vocal-tract cepstrum coefficients which are computed by applying the DCT to the logarithm of the Mel-filtered Spectral envelope of the AR model of the vocal tract, and the voice source cepstrum coefficients that are computed by subtracting the VTCC from the MFCC extracted from the same frame. The set of features where tested on the TIMIT and YOHO databases using a GMM classifier. Clearly in this approach no glottal source reconstruction was carried out. On the other hand, [Kinnunen,2009] differ from previous works in the fact that source-tract separation has been carried out using IAIF algorithm. Once the source and the vocal tract estimates have been computed, a feature vector consisting of MFCC plus $\Delta$ and $\Delta\Delta$ extracted from the speech frames, MFCC extracted from the voice source and LSF parameters extracted from the vocal tract filter were used on a GMM-UBM speaker recognition system. Moreover the reported results were obtained using the NIST 2006 database which contains telephone recordings.

Wavelet analysis has also been proposed to characterise voice source. [Zheng,2007] proposed the use of what they called WOCOR coefficients (Wavelet Octave Coefficients of Residues). A confidence measure is proposed to improve the score-level fusion of two classification systems, one base on classical MFCC parameters and the other one based on this WOCOR coefficients that have been extracted by applying a pitch-synchronous wavelet transform to the LP residual signal. Finally, [Chetouani,2009] proposed three different parameterisation techniques to model the LP residual. Two of them were based on second (application of traditional LPC analysis to the LP residual) or higher order statistics and the other one based on the extraction of frequency information from the LP-residual. However, the best recognition rates were achieved when LPCC and Power difference of spectra in sub-band from the LP-residual were combined, i.e. when the frequency information from the LP-residual was used.

The main conclusion that can be extracted from all of these previous works is that glottal source information presents some discriminative capabilities that are not enough, when processed alone, to outperform the recognition rates of classical parameters. However, it provides additional and uncorrelated information which combined with classical parameters show an improvement in recognition rates.

## 2.6 GLOTTAL SOURCE ALTERNATIVE APPLICATION

Speech Processing Technologies are designed to provide not only information about the message convey in a specific utterance, but also to extract what Nickel in a recent review [Nickel,2006] defines as “contextual side” information. This concept includes meta-information present on the speaker’s voice, such as the identity of the speaker (as we have already presented), but also the speaker’s age, emotional state, voice conditions (healthy or pathological), language and dialect issues, function of prosody and intonation in the message, etc.

Among all these meta-informations, glottal source has proven its value, for instance, in pathology detection, gender and age characterisation. Although classical gender classification methods are based on pitch or formants [Abdulla,2001], [Whiteside,2001]; [Price,1989] proved that voice source of female and male voices present different characteristics both in time and frequency domain. More recently [Muñoz,2010] have also shown that MFCC extracted from glottal source estimates also provide high accuracy in determining the speaker’s gender.

Of greater interest, not only for professionals depending strongly on the use of their voice, but also for the general population due to the increase in the number of diseases

related with certain habits, as smoking, is the detection of certain pathologies in the voice. Although voice pathology detection has been based traditionally in visual inspection, in the last two decades, non-invasive methods are achieving greater interest. It is in this scenario in which the use of the glottal source has begun to prove its value. Good examples of this interest can be found for instance in [de Oliveira Rosa,2000], [Parsa,2000] or more recently in the works of the GIAPSI research group [Gómez,2009], [Gómez,2005-B], [Gómez,2007-B].

# 3 CLASSIFICATION METHODS

As previously said, speaker modelling refers to the process of building a specific representation of the speaker based on specific features extracted from the voice signal. Once a model is obtained for each speaker represented in the system, the classification step consists in evaluating the membership of a new utterance to a specific speaker. Class modelling and classification are critical issues when building a speaker recognition system, since the same set of parameters can lead to different classification results depending on the modelling method and classification used. This chapter provides a description of some of the most popular/classical methods used in text-independent speaker recognition. When available, open-source toolkits which implement these methods will be also presented.

## 3.1 VECTOR QUANTISATION

Vector Quantisation (VQ) [Gray,1984] is one of the simplest text-independent speaker's modelling technique. Close related with clustering, its roots are originally in the domain of data compression. In the specific case of speaker models, this technique provides a method not only to compress the extracted information during the feature extraction process, but a way to reduce the dispersion of the data.

In the training phase, a reference set or codebook is defined for each speaker ($S$) from the train feature vectors extracted from the voice signal. In the test phase, the average quantisation distortion is evaluated between the test utterance feature vectors, $T=\{t_1,...,t_M\}$, and the codebook, $C_S=\{c_1,...,c_K\}$. As far as a smaller value of the average quantisation distortion denotes higher likelihood for $T$ and $C$ originating from the same person, the smallest value achieved provides the target speaker. The average quantisation distortion can be defined as:

$$D_Q(T, C_s) = \frac{1}{M}\sum_{m=1}^{M} \min_{1\le k\le K} d(t_m, c_k) \qquad \textbf{Eq. (3-1)}$$

where $d(\cdot,\cdot)$ is a distance or distortion measure, for example, Euclidean or Mahalanobis distance.

In theory, it is possible to use the whole set of training vectors for a given speaker as the reference set. However, for computational reasons, the number of training vectors is reduced in order to get a codebook using a vector quantiser.

A vector quantiser $Q$, can be defined as a mapping $Q:R^p \rightarrow C$. Which means that the p-dimensional vectors in $R^p$, are mapped into a finite set $C$ of $p$-dimensional vectors, which is called reproduction set or codebook. Each code vector $c_k$ represents a specific region, $k$, of the vector space. The number of vectors in the codebook, $K$, is known as the codebook size. The optimisation of the codebook size is an important issue in VQ.

An additional problem that we have to face when building a vector quantiser is the selection of the code vectors. The objective in this case is reducing the distortion to the minimum. The usual approach is to minimise the overall average distortion:

$$D(Q) = \sum_{k=1}^{K} D_k(Q) \qquad \textbf{Eq. (3-2)}$$

where $D_k=f(d(\cdot,\cdot))$ is known as the average quantisation error for region *k*, and is function of the distance or distortion measure previously cited.

There are two different conditions that a given quantiser must fulfil to be optimal [Gersho,1992]. The first condition also known as nearest neighbour condition states that the quantifier, *Q*, must map any input vector to its closest code vector, *c*:

$$Q(x) = c_j, only\ if\ d(x, c_j) \leq d(x, c_k), k = 1, \dots, K, j \neq k \qquad \textbf{Eq. (3-3)}$$

The second condition also known as centroid condition specifies how the code vector is selected given a set of regions. The code vector, *c*, for a specific region, *k*, is evaluated to minimise the average quantisation error in region *k*.

$$D_k(Q) = \min_c \left[ f\left(d(x, c)\right) | x \in R_k \right] \qquad \textbf{Eq. (3-4)}$$

Different algorithms have been proposed to compute the code vectors. An easy approach to compute the code vectors is to establish an initial estimate of them among the input vectors and then iteratively apply the nearest neighbour condition and the centroid condition until a termination criterion is satisfied (for instance, when no difference is obtained between consecutive iterations). This algorithm is also known as Lloyd's algorithm [Lloyd,1982]. QccPack [Fowler,2000], an open-source package available on-line (URL: QccPack), provides an implementation of the Generalised Lloyd Algorithm for VQ-codebook design.

In the specific case in which the squared Euclidean distance is used as distortion measure, the centroid condition becomes:

$$c_k = \frac{1}{M_k} \sum_{x \in R_k} x \qquad \textbf{Eq. (3-5)}$$

where $M_k$ is the number of vectors in the region *k*, and $c_k$ is merely the average of the training vectors in the selected region. Under these circumstances, the described algorithm is known as LBG (Linde-Buzo-Gray) [Linde,1980] and was presented under the assumption that either the probabilistic model of the input data is known or a long training sequence of data is available. In [Soong,1985], this algorithm was tested for text-dependent speaker recognition under different conditions of codebook size and session variability.

Other methods have been proposed. For example, LVQ (Learning Vector Quantisation) is a supervised neural network algorithm aimed to determine the set of vectors that best represent each of the predefined classes in which the original data is classified. A public-domain software package [Kohonen,1996] (URL: LVQ_PAK) is available on-line to demonstrate the implementation of this algorithm and to ease experiments.

Fuzzy approaches have been also successfully tested especially when applied to reduce the dependence of the resulting codebook on the selection of the initial codebook. Although applied to image compression, a review of different Fuzzy Vector Quantisation algorithms can be found in [Karayiannis,1995].

The main advantage of VQ is that the regions or classes obtained in the process do not follow a specific distribution or probability density function. Moreover, each cluster can represent an acoustical class without any kind of labelling. Additionally, vector

quantisation can provide competitive accuracy when combined with background model adaptation as described for instance in [Kinnunen,2009].

A deeper review of these algorithms along with pseudo-code implementation can be found in [Gersho,1992], [Friedman,1999], [Theodoridis,2006].

## 3.2 HIDDEN MARKOV MODELS

Since they were presented in the mid-60s by Baum and his colleagues [Baum,1972], [Baum,1970], [Baum,1968], [Baum,1966-A], [Baum,1966-B] Hidden Markov Models (HMMs) have been one of the most used techniques both for speech and speaker recognition. Although extensively used for text-dependent speaker recognition applications, it also has been successfully applied to text-independent applications.

In standard Markov models, each state corresponds to a deterministic observable event, for instance to a phonetic category. In order to overcome the limitations of Markov models, they can be extended to HMMs, in the sense that we include the case in which the observable event is a probabilistic function of the state. Therefore the HMM is a doubly embedded stochastic process where the underlying stochastic process is hidden (not directly observable), but can only be observed through another set of stochastic processes that produce the sequence of observations [Rabiner,1989].

From this point of view, HMMs can be regarded as a stochastic model that can efficiently model temporal sequences of events. Actually, when applied to speech signals, the statistical variations in the speech signal parameters over time can be represented by stochastic Markovian transitions between states. In other words, HMMs model the feature space (using a set of states) of a specific speaker and the sequence (using transitions between states) in which the speaker is more likely to generate the specific feature vectors.

As an HMM is an extension of Markov models, a brief introduction to Markov models will be carried out before delving into HMMs.

- *First-order* discrete-time Markov model.

  Consider a system defined by a set of N distinct states, $S=\{S_1,\ldots,S_N\}$, which is in a certain state at any time, and with the system changing to another state (or remaining in the same state) at regularly spaced discrete times, $t=\{1,2,\ldots,T\}$. In a first-order discrete-time Markov model, at any time t, the system is in a particular state that we denote as $q_t$. Moreover, the Markov property states that the conditional probability distribution for the system at the next step given its current state depends only on the current state of the system, discarding all previous states. i.e.:

$$P\left[q_{t+1}=S_j|q_1=S_l,q_2=S_k,\ldots,q_t=S_i\right]= P\left[q_{t+1}=S_j|q_t=S_i\right] \qquad \textbf{Eq. (3-6)}$$

  Thus we can obtain the set of state transition probabilities, $a_{ij}$:

$$a_{ij}=P\left[q_{t+1}=S_j|q_t=S_i\right],1\leq i,j\leq N \qquad \textbf{Eq. (3-7)}$$

  where the following restriction applies:

$$a_{ij} \geq 0, \forall i, j \qquad \textbf{Eq. (3-8)}$$

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i \qquad \textbf{Eq. (3-9)}$$

Given this description, a Markov model can be represented by a directed graph in which each node represents a specific state and links between nodes represent transition probabilities (See Figure 3-1).

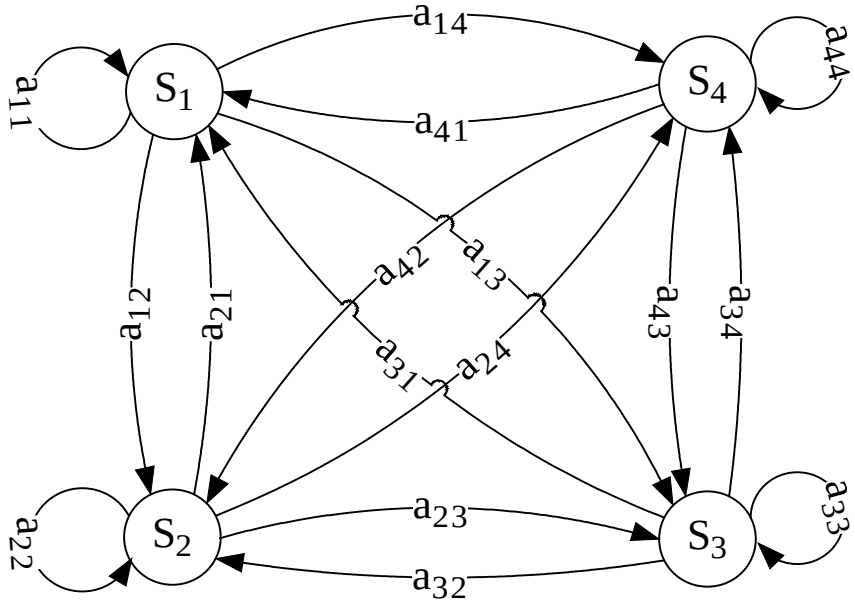



**Figure 3-1** A Markov chain with 4 states. The discrete states, labelled as $S_i$ are represented by nodes, and the transition probabilities, $a_{ij}$, are represented by links between nodes.

Since the output of the process is the sequence of observable states at each time, this stochastic process also receive the name of observable Markov model. An interesting question arising at this moment is how we could evaluate the probability that a particular sequence of observations has been generated by a particular model (where the full set of $a_{ij}$ is known). In this case, we simply multiply the successive state transition probabilities. In other words, assuming a particular model $S=\{S_1, S_2, S_3, S_4\}$, with transition probabilities $A=\{a_{ij}\}$, and a particular observation sequence O, for instance $O=\{S_3, S_2, S_3, S_1, S_4, S_2\}$, then

$$P(O|S) = a_{32}a_{23}a_{31}a_{14}a_{42} \qquad \textbf{Eq. (3-10)}$$

Additionally if the prior probability on the first state is known $P(q_1=S_i)$, we must take into account such factor as well. A deep review on both Markov models and HMMs can be found in [Bishop,2006], [Theodoridis,2006], while a comprehensive review for speech applications can be found in [Rabiner,1989].

Going back to HMMs, they can be regarded as a finite state machine, with a given number of states $S=\{S_1,..,S_N\}$ connected by a transition network $A=\{a_{ij};\ 1\leq i,j\leq N\}$. The transitions from one state to another depend on probabilities rather than in a

deterministic rule. As previously said, the set of states is not visible. Instead, a set of observation symbols (corresponding to the observations being modelled by the HMM) $V=\{V_1,..,V_M\}$ can be emitted from each particular state. Figure 3-2 shows a 3-state HMM, where all transitions between states are possible, and where the probability of emitting a specific symbol, $V_{k,}$ in a specific hidden state, $S_{j,}$ is denoted by $b_j(k)$. When every one of the hidden states have a non-zero probability of occurring given some starting state, the HMM is called *ergodic*.

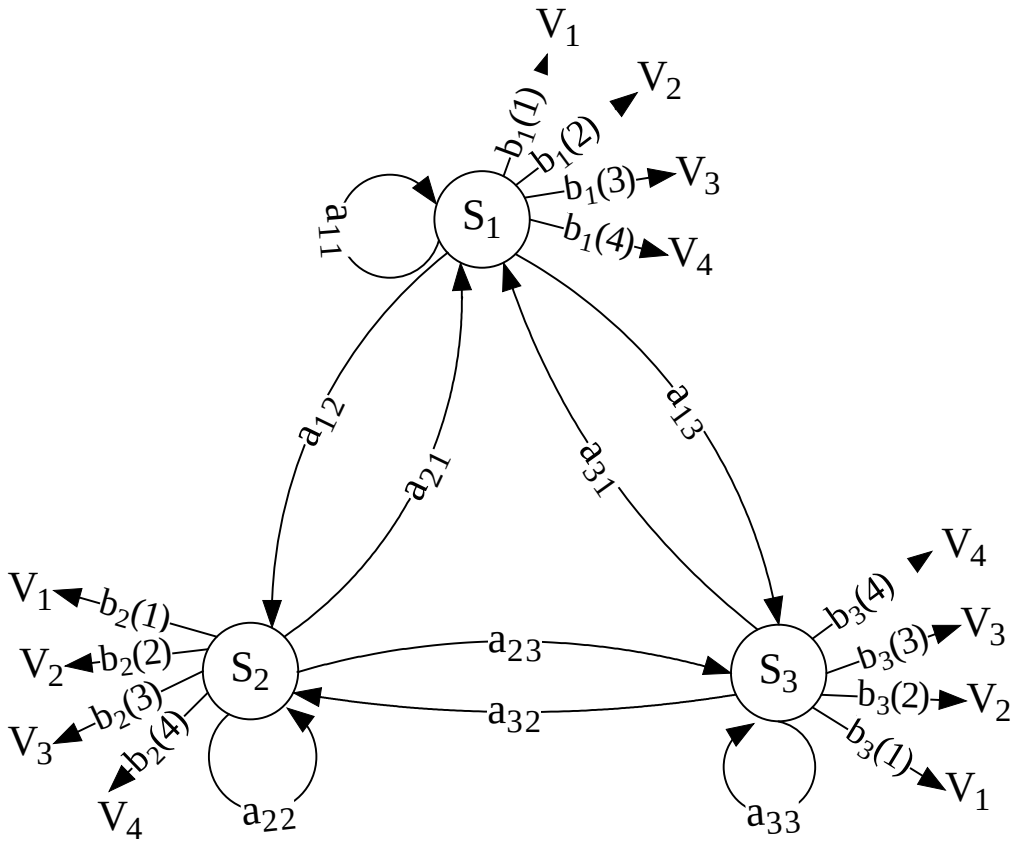



**Figure 3-2** 3-state HMM; where $a_{ij}$ represents the state transition probabilities, $V_k$ denotes the observation symbols also known as emitting or visible states and $b_j(k)$ represents the probability of the emission of a visible state.

An alternative to the ergodic HMM, is the Bakis or left-to-right HMM (see Figure 3-3) [Bakis,1976]. In this case, the Markov chain starts in a particularly state and goes through different intermediate states until it reaches a final state, moreover, when traversing the intermediate states no backward transitions are allowed.

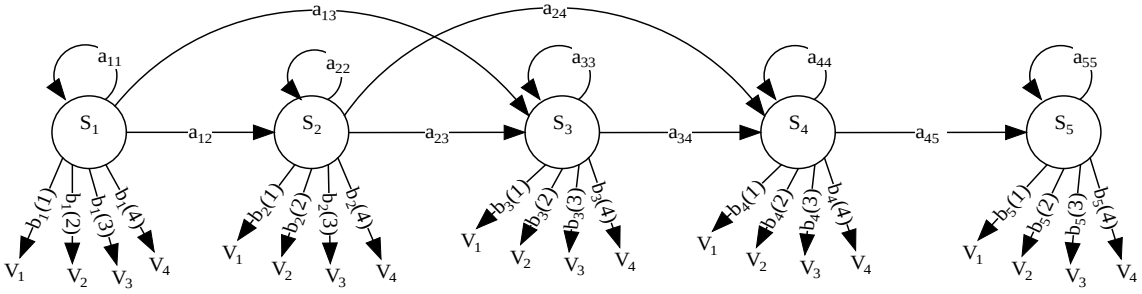



**Figure 3-3** Illustration of 5-state left-to-right HMM

A formal description of this type of HMM, establishes the following properties and/or restrictions:

- Restrictions over state transition coefficients:
  - No backward transition are allowed:

$$a_{ij} = 0, j < i \qquad \textbf{Eq. (3-11)}$$

  - An optional restriction that has been successfully applied, especially in speech recognition, is not allowing large changes in state indexes.

$$a_{ij} = 0, j > i + \Delta \qquad \textbf{Eq. (3-12)}$$

  - Left-to-Right models have a final state, *N*, for which the following restriction applies:

$$a_{Ni} = \begin{cases} 0, i \neq N \\ 1, i = N \end{cases} \qquad \textbf{Eq. (3-13)}$$

- Restrictions over state probabilities:
  - Left-to-Right models have an initial state, therefore, the initial state probabilities must fulfil the following restriction:

$$\pi_i = \begin{cases} 0, i \neq 1 \\ 1, i = 1 \end{cases} \qquad \textbf{Eq. (3-14)}$$

No matter which type of HMMs we are dealing with, the general problem addressed by the HMM is to build a probabilistic model that best explains an observation sequence. To solve this problem it is necessary to formally define the elements that characterise an HMM:

- $N \rightarrow$ Number of states (which usually have some physical meaning associated) in the model. As previously specified, we represent the set of states as $S=\{S_1,\ldots,S_N\}$, and the state reached at instant $t$ as $q_t$.
- $M \rightarrow$ Number of symbols that correspond to the distinct observation symbols being modelled by the system. We denote the individual symbols as $V=\{V_1,..,V_M\}$.
- $A=\{a_{ij}\} \rightarrow$ Set of transition probabilities which control the way the hidden state at instant $t$ is chosen given the hidden state at instant $t-1$. Typically, each $a_{ij}$ satisfies equations Eq. (3-7) and Eq. (3-8).
- $B=\{b_j(k)\} \rightarrow$ Set of symbol emission probabilities for each hidden state, $j$, where

$$b_j(k) = P\left[V_k | q_t = S_j\right]; 1 \leq j \leq N; 1 \leq k \leq M \qquad \textbf{Eq. (3-15)}$$

- $\pi=\{\pi_i\} \rightarrow$ Initial state distribution, where

$$\pi_i = P[q_1 = S_i], 1 \leq i \leq N \qquad \textbf{Eq. (3-16)}$$

Based on these definitions, we can assert that in order to fully characterise an HMM, it is necessary to specify the N and M model parameters, in addition to the set of observation symbols and the three probability measures, A, B, π. However, we can use the following compact notation:

$$\lambda = \{A, B, \pi\} \qquad \textbf{Eq. (3-17)}$$

With these preliminaries, we must focus now on three central issues that must be solved for the model to be useful in real-world applications:

- The Evaluation Problem: From a practical point of view, this is a scoring problem in the sense that the solution to this problem allows us to decide on the best model among a set of different models that best matches a sequence of observations. Formally, we are looking for an efficient way to compute *P(O, λ)*, where $O=\{o_1, o_2,\ldots,o_T\}$ is the observation sequence ($o_j \in V$) and $\lambda=\{A, B, \pi\}$.

  An efficient way to compute this probability is the forward-backward algorithm [Baum,1966-A], [Baum,1968], [Rabiner,1989], although in this section only the forward component will be used. The forward algorithm defines the following forward variable:

$$\alpha_t(i) = P(o_1, o_2, \ldots, o_t, q_t = S_i|\lambda) \qquad \textbf{Eq. (3-18)}$$

  Which is the joint probability of state $S_i$ and the partial observation sequence $\{o_1, o_2,\ldots,o_t\}$ until instant *t*. In other words, represents the probability that the model $\lambda=\{A, B, \pi\}$ is in state $S_i$ having generated the observation sequence, $\{o_1, o_2,\ldots,o_t\}$, at instant *t*.

  Given the initialisation for instant *t*=1,

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N \qquad \textbf{Eq. (3-19)}$$

  we can compute $\alpha_{t+1}(i)$, by induction, for all the states *j*, $1 \leq j \leq N$, and for all instants *t*=1,2,…,*T*.

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), 1 \leq t \leq T, 1 \leq j \leq N \qquad \textbf{Eq. (3-20)}$$

  The expression in brackets represents the probability of reaching $S_j$ at instant *t*+1, from all possible model' states at instant t, having observed the partial sequence $\{o_1, o_2,\ldots,o_t\}$. Summing the terminal forward variables, $\alpha_T(i)$, over all possible states of the model, then we can obtain the expected probability:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad \textbf{Eq. (3-21)}$$

- The Decoding Problem: When solving this problem, we are providing an optimal sequence of hidden states, $Q=\{q_1,\ldots,q_T\}$, (according to some optimality criterion) for a given sequence of observations, $O=\{o_1, o_2,\ldots,o_T\}$, and a specific model $\lambda=\{A, B, \pi\}$. It is important to notice that, except in the case of degenerated models, there is no correct/unique state sequence to be found, but the optimal one, i.e., the one that best explains the observation sequence. The Viterbi algorithm [Viterbi,1967], [Forney,1973] is one of the most used algorithms to maximise $P(Q|O,\lambda)$, which is equivalent to maximizing $P(Q,O|\lambda)$.

  The Viterbi algorithm defines the following variable:

$$\delta_t(i) = \max_{q_1 q_2,\ldots,q_{t-1}} P[q_1, q_2, \ldots, q_t = S_i, o_1, o_2, \ldots, o_t|\lambda] \qquad \textbf{Eq. (3-22)}$$

  which represents the sequence of highest probability at instant $t$, ending in state $S_i$ and generated by the first $t$ observations in the sequence. Given the initialisation for instant t=1,

$$\delta_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N \qquad \textbf{Eq. (3-23)}$$

  we can compute $\delta_{t+1}(i)$, through the following recursion for all the states $j$, $1 \leq j \leq N$, and for all instants $t$=2,…,$T$

$$\delta_t(j) = \left[\max_{1\leq i\leq N} \delta_{t-1}(i)a_{ij}\right] b_j(o_t), 1 \leq j \leq N, 2 \leq t \leq T \qquad \textbf{Eq. (3-24)}$$

  However, the solution to the decoding problem must be an optimal sequence of hidden states, so we must also define an array,$\psi$, which keeps track of the argument maximizing Eq. (3-24).

$$\psi_1(i) = 0, 1 \leq i \leq N \qquad \textbf{Eq. (3-25)}$$

$$\psi_t(j) = \operatorname*{argmax}_{1\leq i\leq N}\left[\delta_{t-1}(i)a_{ij}\right], 1 \leq j \leq N, 2 \leq t \leq T \qquad \textbf{Eq. (3-26)}$$

  In this way, the optimum path can be found using the following path backtracking:

$$q_T^* = \operatorname*{argmax}_{1\leq i\leq N}[\delta_T(i)] \qquad \textbf{Eq. (3-27)}$$

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \ldots, 1 \qquad \textbf{Eq. (3-28)}$$

- The Learning Problem: In this case we are facing a parameter optimisation problem. Formally, given a training sequence, $T=\{t_1,\ldots t_L\}$, and an initial model, $\lambda^I=\{A^I, B^I, \pi^I\}$, we need to find a procedure to optimise the model parameters, so as to best describe the training sequence. The optimised model, $\lambda^O=\{A^O, B^O, \pi^O\}$, must fulfil the following constrain:

$$(T|\lambda^O) \geq P(T|\lambda^I) \quad \textbf{Eq. (3-29)}$$

Probably one of the most used parameter re-estimation algorithm is the iterative "Baum-Welch" algorithm [Baum,1970]. This algorithm uses the same principles as the Expectation-Maximisation algorithm which we will discuss later (section 3.3.1) on this chapter. In addition, the backward variable which we have omitted in the analysis of the Evaluation Problem will now be used:

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \dots, o_T|q_t = S_i, \lambda) \quad \textbf{Eq. (3-30)}$$

Given Eq. (3-30) we can formally define $\beta_t(i)$ as the probability of generating the partial observation sequence $O=\{o_{t+1},o_{t+2},\dots,o_T\}$, from instant $t+1$ until $T$, assuming the model is in state $S_i$ at instant $t$. Like in the case of the forward variable, the backward variable can be evaluated through the following recursion:

$$\beta_T(i) = 1, 1 \leq i \leq N \quad \textbf{Eq. (3-31)}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)\,; t = T-1, T-2, \dots, 1; 1 \leq i \leq N \quad \textbf{Eq. (3-32)}$$

Additionally, we must define the probability of being in state $S_i$ at instant $t$ and in state $S_j$ at instant $t+1$(transition probability), given the model and the observation sequence:

$$\xi_t(i,j) = P\big(q_t = S_i, q_{t+1} = S_j|O, \lambda\big) \quad \textbf{Eq. (3-33)}$$

which can be rewritten in terms of forward and backward variables as:

$$\xi_t(i,j) = \frac{P\big(q_t = S_i, q_{t+1} = S_j, O|\lambda\big)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{k=1}^{N} \alpha_T(k)} \quad \textbf{Eq. (3-34)}$$

Given the above formulas, we can use the following equations to re-estimate the parameters of the initial HMM, $\lambda^I=\{A^I, B^I, \pi^I\}$,

$$\bar{\pi}_i = \sum_{j=1}^{N} \xi_1(i,j) \quad \textbf{Eq. (3-35)}$$

$\bar{\pi}_i$ can be regarded as the expected frequency in state $S_i$ at instant t=1. However, in voice applications $\bar{\pi}_i$ is usually set to 1 for the initial state

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} \xi_t(i,k)} \quad \textbf{Eq. (3-36)}$$

We may regard $\bar{a}_{ij}$ as the ratio between the average number of transitions from state $S_i$ to state $S_j$, and the number of transitions from state $S_j$.

$$\bar{b}_j(V_k) = \frac{\sum_{t=1,o_t=V_k}^{T} \sum_{i=1}^{N} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{i=1}^{N} \xi_t(i,j)} \quad \textbf{Eq. (3-37)}$$

In this case, $\bar{b}_j(V_k)$, may be regarded as the ratio between the expected number of times the symbol $V_k$ is observed while in state $S_j$, and the expected number of times that the model is in the state $S_j$.

### 3.2.1 Types of HMMs

So far we have considered the standard type of HMM, where the state space of hidden variables is discrete and the observations were characterised also as discrete symbols. However, in some applications, the observations can also be continuous signals (or vectors). To overcome this problem, we can quantise the continuous vectors, for instance via codebooks. However, the use of quantisation introduces certain problems, since a small codebook will lead to high quantisation noise, whereas a large codebook can lead to many centroids being not represented in the training data thereby resulting in degradation of the recognition rates. A feasible solution to this problem is the use of continuous density HMM (CD-HMM) [Liporace,1982].

In the case of CD-HMM, a set of Gaussians is usually used to statistically characterise the observed feature vectors as a multivariate distribution within each state:

$$b_j(o) = \sum_{k=1}^{K} c_{jk} G\left(o, \mu_{jk}, \Sigma_{jk}\right) \quad \textbf{Eq. (3-38)}$$

In this case, the set of parameters $b_j(o)$ (where o is the vector being modelled), depends on a set of mixture coefficients, $c_{jk}$ (mixture component for $m^{th}$ mixture in state $j$), that weights each Gaussian density with mean vector $\mu_{jk}$, and covariance matrix, $\Sigma_{jk}$. Additionally the mixture weights must satisfy the stochastic constraints:

$$\sum_{k=1}^{K} c_{jk} = 1, 1 \le j \le N \quad \textbf{Eq. (3-39)}$$

$$c_{jk} \ge 0, 1 \le j \le N, 1 \le k \le K \quad \textbf{Eq. (3-40)}$$

In [Liporace,1982] we can find a set of formulas which allow parameter re-estimation, in order to solve the HMM learning problem, similar to the ones presented above. One of the problems of this approach is the fact that the parameters are not shared by all states. This implies that a small value of $K$, i.e. the number of mixtures, would lead to the erroneous assumption that the distribution probability can be modelled by a Gaussian mixture. Correspondingly, a large value of $K$ would result in a lack of representation for many states.

An intermediate solution presented in [Huang,1989], consists in using semi-continuous density models. The basic idea behind this proposal is the use of a common codebook for all the states, which describes the set of acoustic classes. However, in this case classes are not represented by discrete centroids but continuous probability density functions, typically Gaussians.

A comparative study of the performance of discrete and continuous HMMs, as well as VQ distortion in text-independent speaker recognition can be found in [Matsui,1992]. Although the database used in the experiment is quite limited in the number of speakers, they conclude that continuous HMMs present better performance than discrete HMMs and even the VQ approach when there is enough training data.

No matter whether we are using discrete, continuous or semi-continuous HMMs, we can use each of the two different structure of HMMs presented above. In other words, ergodic HMMs, in which every state of the model could be reached in a single step from every other state in the model, and Left-to-Right HMMs, in which as time index increases, the state index also increases (or at least remains the same). However, in the literature we can find other kinds of HMMs regarding their structure.

For instance, we can find HMMs in which the observations are no longer associated with the states but rather with the transitions. In this type of HMMs it is useful to allow the use of transitions that produce no output, providing a powerful way to describe phonetic elision phenomena. Such transitions are called null transitions [Bahl,1983].

Another interesting variation is the incorporation of the parameter tying concept [Bahl,1983], which has been extensively used in speech recognition. The main idea behind this concept is to reduce the number of independent parameters in the model by setting up an equivalent relation between HMMs parameters in different states. Tied mixture systems use a single set of Gaussian distributions which is shared across all states. The different mixture weight distribution of the shared Gaussians is what differentiates each state. Some variations have been proposed in order to get more detailed models. For instance, [Paul,1991] proposed the used of phonetically tied mixture (PTM) systems, in which a separate Gaussian codebook is shared among all triphone states corresponding to the same base phone. Another variation consists in clustering HMM states according to acoustic similarity, leading to State-clustered HMMs. The states in each cluster either share the same Gaussian distribution [Woodland,1994], or only share the same set of Gaussians but with different weights for each state [Digalakis,1996]. This method is of particular interest in cases when the amount of training data available to build the model is not sufficient. It shall be noticed that the mathematical issues involved on the learning algorithm are not affected by the parameter tying [Bellegarda,1990], despite the parameter estimation problem becomes somewhat simpler.

Finally, we can cite the autoregressive (AR) HMMs, in which the observation vectors are drawn from an autoregressive process. In [Poritz,1982], AR-HMM were first introduced and applied to speaker recognition through a 5-state ergodic HMM. Lately, [Juang,1986] expanded this concept to richer class of mixture autoregressive HMMs, in which the states are described as a linear combination of AR sources, and applied it to speaker-independent recognition of isolated digits. [Tisby,1991], applied this last approach to speaker recognition and compared its performance with a VQ-system concluding that the improvement is not worthy if compared to the cost of training these models.

One important issue related to HMMs is the inclusion of state duration density in the model, although it seems to be more useful for continuous speech recognition rather than in speaker recognition systems. In this case, the probability density function of duration in a state must be evaluated from the training data. Although the theoretical framework to incorporate temporal information is developed in [Rabiner,1989], and when used is supposed to significantly improve the system, it presents a high

computational cost that discourages its use (especially in terms of the increase in the number of parameters to be estimated for each state, and therefore the computational load of re-estimation).

### 3.2.2 Implementation Issues

In this section we will introduce some practical implementation issues as well as limitations when dealing with HMMs.

- Optimisation Criterion: To lighten some of the problems related with the estimation of the HMM parameters different alternatives to the standard maximum likelihood (ML) optimisation procedure have been proposed:
  - Maximum Mutual Information (MMI): Originally presented in [Bahl,1986], it has been successfully applied to speech recognition, as the main idea behind this method is not just modelling the distribution of the data set of the target class, but all the classes (HMMs representing, for instance, different words to be recognised) in the system jointly. In the MMI approach, we assume that the speech units represented by the HMMs, $\lambda_v$, $v$=1, 2,…,$V$, are equally likely. Assuming the existence of a set of training sequence observations $O_v$, the objective is to maximise the average mutual information, I, between the observation sequence and the complete set of models:

$$I_i = \max_{\lambda}\left(\log P\left(O^i|\lambda_i\right) - \log\sum_{j=1}^{V} P\left(O^i|\lambda_j\right)\right) \qquad \textbf{Eq. (3-41)}$$

    where $\lambda_i$ is the target model and $\lambda_j$ the others. When applied to the whole set of training sequences a possible implementation would be:

$$I = \max_{\lambda}\left(\sum_{i=1}^{M}\left(\log P\left(O^i|\lambda_i\right) - \log\sum_{j=1}^{M} P\left(O^i|\lambda_j\right)\right)\right) \qquad \textbf{Eq. (3-42)}$$

    Clearly, MMI estimation is not explicitly designed to maximise the target model probability, but to maximise the mutual information between the target class and other classes.

    However, conventional MMI training tends to over train the models, thus decreasing the discriminative capability. To overcome this problem, [Kim,1998] proposed a cross-validation approach in which parameters are estimated on a subset of the data and the recognition performance is evaluated on the remainder.

  - Minimum Discriminant Information (MDI): [Ephraim,1987] proposed this iterative approach that aims at minimizing the discrimination information between the source (not necessary a Markov source but just a positive definite correlation function) and the model. In this case, the goal is to find the HMM parameters which minimise the discriminant information between the signal probability densities ($Q$) and the set of HMM probability densities ($P_\lambda$). The discriminant information between two probability distributions ($P$ & $Q$) with probability density functions p and q, respectively, can be expressed as:

$$D(Q||P) = \int q(y) \ln\left(\frac{q(y)}{p(y)}\right) dy \qquad \textbf{Eq. (3-43)}$$

In order to minimise the discriminant information (i.e., compensate for mismatches between measurements and model), thus finding optimum values of $\lambda=(A,B,\pi)$, a modified variant of the Baum-Welch algorithm is used, which actually is an iterative descent algorithm for the discriminant information measure with local convergence.

Broadly speaking, the procedure begins with the estimation of the HMMs according to the traditional ML criterion. Then, for a given HMM a probability distribution for the source is estimated by minimizing the discriminant information measure over all probability distributions of the source which agree with the given measures. Finally, given a probability distribution of the source, we estimate the HMM that minimises the discrimination information on the set of hidden Markov models

- Minimum Error Classification (MEC): Also known as corrective training or discriminative training, MEC can be regarded as a post-processing phase that aims to increase the discriminative power of the models by re-estimating its parameters. Proposed by [Juang,1992], it has been successfully applied to speaker recognition [Rosenberg,1998], [Siohan,1998], [Liu,1995]. A discriminant function is first described, so that the classifier makes a decision for each training observation sequence by choosing the largest of the discriminants evaluated on the training sequence. Then a set of misclassification measures is defined for each class, taking into account the models of other competing classes, which attempt to evaluate how likely an observation is misclassified. Finally, a gradient descendent algorithm is used to derive all model parameters so that the total loss (cost function which embeds misclassification measures), for a given set of observations is minimised.
- Maximum Likelihood Linear Regression (MLLR): [Leggetter,1995-A] proposed the MLLR method as an alternative adaptation method for continuous density HMMs. In the experiments carried out, an initial speaker-independent system is adapted to a specific speaker by applying a set of linear transformations to the Gaussian mean vectors, i.e. updating the HMM parameters. Each transformation is used for a number of Gaussian distributions, and the number of transformations is determined by the amount of available training data, making this adaptation method suitable for cases in which training data is limited. The parameters of the transformation matrices are estimated to maximise the likelihood of the speaker specific data. MLLR will be briefly presented in the next section for the case of background model adaptation in GMM based systems.

- Limitations of the HMMs:

  HMMs are a powerful tool that theoretically allows modelling any probability distribution over sequences. Moreover, they have become the standard in automatic speech recognition and have proven its value in speaker recognition as well, thanks to the availability of efficient training and decoding algorithms. However, it is claimed that HMMs present some limitations (although they may

be solved, if feasible in practice, with large number of model parameters and a large amount of training data):

- Duration modelling issues: In the traditional HMM approach, state duration is modelled by the state transition probabilities. However, as reported by [Matsui,1992] the contribution of transition probabilities to the likelihood score in text-independent speaker recognition is often insignificant. Moreover, a recent study [Deshpande,2008] concluded that the speaker identification rates using CDHMMs are strongly correlated with the number of mixtures per state and the amount of training data. In contrast, as reported above, other authors have addressed the problem by introducing models with explicit state duration distributions [Russell,1985], [Levinson,1986]. The major drawback, which is also the solution, is the increase in the number of model parameters (therefore an increase in the computational load) and the need for large training data.
- Conditional independence of observations: A major limitation when dealing with speech/speaker recognition is the poor modelling of the acoustic context, i.e., the assumption of conditional independence of observations (frames of speech) given the state sequence. Thanks to this assumption, the probability of a sequence of observations can be expressed as:

$$P(O_1, O_2, \dots, O_T) = \prod_{i=1}^{T} P(O_i) \qquad \textbf{Eq. (3-44)}$$

  One possible solution to this limitation is the inclusion of short-term dynamic properties in the observation space. Other approaches proposed the use of HMM variations in which the observation correlation is explicitly modelled. For instance, the statistical dependence between the current observation vector and the last observed one can be modelled explicitly by conditional Gaussian HMMs [Wellekens,1987]. Buried Markov Models (BMM) [Bilmes,1999], which are standard HMMs to which extensions in the form of conditional dependencies between components of different features vectors have been added, are another possible solution. The observations are no longer considered conditionally independent of past ones, but the most important dependencies are considered explicitly. Segmental HMMs [Russell,1993], [Yun,2000] are also related with this limitation. In this approach, the observations are conditionally dependent both on the current HMM state and on the mean of a sequence of speech frames to which they belong. [Ostendorf,1996] present a review and general stochastic framework for segmental HMMs as well as the analogies between them and traditional HMMs.

  In addition, and also related with the segmental HMM approach, we are assuming that speech can be well represented by frame-based observations, with instantaneous transitions between them. However, a voice trace can be represented as a moving point in the parameter space as articulation changes occur, forming the speech trajectory. Given that a point may belong to different paths, models for speech recognition should be based on paths and not on the individual geometric positions

of the parameter space. The independence of observations present in the HMMs does not preserve the trajectory information.

- Finally, the Markov assumption itself constitutes another limitation [Rabiner,1989]. The modelling of first order hidden Markov models assumes that the transition probability between states at time *t+1* depends only on the state of the Markov chain at time *t*, which is clearly inappropriate for speech sounds. One possible solution, although also subject to the limitations already presented, are the second-order HMMs. In the second-order models HMM2, the transition probability of state in the instate t +1 depends on the states of the chain at times t and t-1. Experiments carried out with connected digit recognition show an improvement in performance.

- Using HMMs for speaker recognition:

  Like in any other recognition system, the use of HMMs for speaker recognition purposes involves two different stages: training phase and testing phase. During training phase an HMM is created for each speaker enrolled in the recognition system. In other words, for each speaker, *k*, the HMM model parameters $\lambda_k=\{A,B,\pi\}$, are estimated so that the likelihood of the training observation sequence for each speaker is optimised. The test phase differs depending on whether the task is identification or verification. In speaker identification for each input waveform, the speaker whose model likelihood is highest is selected as the identification result. In the case of verification, if the likelihood of the input waveform for the claimed speaker's HMM is larger than a specify threshold, the claimant is accepted by the system while otherwise is rejected. And additional issue that must be taken into account is whether the recognition system is text-dependent or text-independent. Although it is not a rule that must be followed strictly, in the first case Left-to-Right HMMs (with self-loops and no skips) are usually used, thus imposing a time sequence structure, while in the later ergodic HMMs are usual the choice.

  While the recognition phase is easily performed by a Viterbi process that calculates the log probability of each utterance, the training phase is probably the most difficult problem when dealing with HMMs. A re-estimation algorithm such as Baum-Welch training algorithm must be applied to an initial model estimate. However, according to [Rabiner,1989] there is no simpler or straightforward way of choosing the initial estimates of the HMM parameters, although based on experience A and $\pi$ can be either a random or uniform distribution whereas B parameters need a good initial estimation to get rapid and proper convergence (especially when dealing with multiple mixtures in the CDHMM). In spite of this, a two-step initialisation procedure has been successfully used for speaker recognition as reported in [Yu,1995], [Boakye,2005], [Deshpande,2008].

  In the first stage, a k-means procedure can be used to initialise the output probabilities of each state, based on a preliminary segmentation of the training observations into a number of segments equal to the number of states. Form these segments, the mixture components (means and variances) of each state are computed in an iterative process. The transition probabilities can be set to reasonable default values (for instance, time counts of state occupation or $1/D_i$, where $D_i$, is the number of transitions allowed out of state ***i***) and then updated

iteratively along with the output probabilities. This initial estimation is followed by the application of the Viterbi segmentation algorithm to find the most likely state sequence for the observation sequence. Re-computation of the mixture components occur during realignment of the observations. In the second stage, the Baum-Welch re-estimation algorithm is applied until convergence on the HMM parameters using the same training data. Diagonal covariance matrices are assumed in the estimation of the output distributions.

### 3.2.3 Available software tools

Probably one of the most known and extensively used toolkits for building and manipulating HMM is HTK (URL:HTK) (Hidden Markov Model Toolkit). Originally developed at the Machine Intelligence Laboratory of the Cambridge University Engineering Department, it consists of a set of library modules and tools available in C source form. HTK provides specific tools for speech analysis, HMM training and testing, supporting both continuous density mixture Gaussian HMMs and discrete distributions as well. Although Microsoft retains the copyright to the original code, HTK is available for teaching and academic research at no cost. This free availability has promoted the use of this toolkit not only in speech recognition applications, but also in speech synthesis, speaker recognition, character recognition or even DNA sequencing.

Another C-library that implements data structures and algorithms for basic and extended HMMs is the General Hidden Markov Model library (GHMM). The GHMM (URL:GHMM) is developed by the Algorithmics Group at the Max Planck Institute for Molecular Genetics and is freely available under LGPL license. GHMM toolkit provides support for non-homogenous Markov chains, clustering and mixture modelling for HMMs, discrete and continuous emissions among others.

Jahmm (URL: JAHMM) provides a java implementation of HMM related algorithms (Viterbi, Forward-Backward, Baum-Welch and k-means among others) under the BSD license. According to the author, Jahmm is not computationally efficient as focus was on code readability for academic purposes.

MATLAB (URL: MathWorks) provides a statistical toolbox which includes a set of functions related to HMMs. Additionally, as MATLAB is one of the most widely used software products for algorithmic development, some other toolboxes have been made available on the web to deal with HMMs. For instance, Professor K. Murphy (Univ. of British Columbia) provides a toolbox (URL: HMMTM) that supports inference and learning for HMMs with discrete outputs, Gaussian outputs (ghmm's), or mixtures of Gaussians output. H2M (URL: H2M) provides a set of functions that implement the EM algorithm with multivariate Gaussian state-conditional distribution. Three special cases have been considered: Gaussian mixture models, Ergodic (or fully connected) Gaussian hidden Markov models and Left-to-Right Gaussian hidden Markov models.

## 3.3 GAUSSIAN MIXTURE MODELS

A Gaussian Mixture Model (GMM) is a probabilistic model which has become the *de facto* reference method in text-independent speaker recognition. This can be confirmed, for example, if we take a look at the systems presented at the 2008 NIST SRE (URL: NIST SRE 2008), in which most of them used this kind of models alone or combined with other models. Gaussian mixture models are not restricted to the speaker recognition area; they are also used in data mining machine learning and statistical analysis.

It is important to notice that GMMs do not explicitly use any phonetic information to build the speaker model. However, each Gaussian in the model can be seen as a model of an underlying phonetic event which pooled together with all the phonetic events characterises a person's voice. Trying to establish a relation between this model (GMM) and the ones previously reviewed, GMMs can be regarded as a hybrid approach between a unimodal Gaussian classifier and a VQ codebook, in the sense that each feature vector is not assigned to a specific class but it has a nonzero probability of being originated from each of the available classes. Moreover, it can be viewed as a single-state HMM with a Gaussian mixture observation density or an ergodic Gaussian observation HMM with fixed, equal transition probabilities.

A GMM for a specific speaker is composed of a set of multivariate Gaussian components. If a sufficient number of Gaussians are considered and a fine-tuning of means, covariances and mixing weights is carried out, almost any continuous density can be approximated. The goal, when training the speaker model characterised by $\lambda_S=\{w_i,\mu_i,\Sigma_i\}_{i=1,\dots M}$, from a set of D-dimensional training vectors $X=\{x_1,\dots,x_N\}$, is to determine the $w_i$ (weights), $\mu_i$ (means) and $\Sigma_i$ (covariances) values that best represents the speaker given a specific number of Gaussians $M$. For the set of training vectors, the mixture density for speaker $S$, can be defined as:

$$P(X|\lambda_s) = \prod_{n=1}^{N} p(x_n|\lambda_s) \qquad \textbf{Eq. (3-45)}$$

Moreover, for a specific feature vector $x_n$, we can define:

$$p(x_n|\lambda_s) = \sum_{i=1}^{M} w_i p_i(\mathrm{x}_n) \qquad \textbf{Eq. (3-46)}$$

That is to say, the density is a weighted linear combination of Gaussian densities, $p_i(\mathrm{x_n})$, each one characterised by a *Dx1* mean vector, $\mu_i$, and a *DxD* covariance matrix, $\Sigma_i$.

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}, 1 \le i \le M \qquad \textbf{Eq. (3-47)}$$

The mixing weights also satisfy the following constrains:

$$\sum_{i=1}^{M} w_i = 1, \qquad 0 \le w_i \le 1, \forall i \qquad \textbf{Eq. (3-48)}$$

The basic approach when training a GMM is *Maximum Likelihood* (ML) estimation. Assuming that feature vectors in $X$ are independent, the average log-likelihood of $X$ with respect to model $\lambda_S$, can be defined as:

$$LL_{avvg}(X|\lambda_s) = \frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\lambda_s) = \frac{1}{N}\sum_{n=1}^{N} \log \sum_{i=1}^{M} w_i p_i(x_n) \qquad \textbf{Eq. (3-49)}$$

Dividing by $N$ allows us to normalise the duration effects from the log-likelihood value. The higher the $\mathrm{LL_{avg}}$ value, the higher the indication that the vectors in $X$ where originated from the model $\lambda_S$. From this point of view, this condition can also be used not only in the train phase but in the test phase as well. Assuming that the GMM models $\lambda=\{\lambda_1,\dots,\lambda_L\}$ for a set of known speakers $S=\{S_1,\dots,S_L\}$ are available, and given a set of

test vectors from a target speaker $T=\{t_1,...,t_N\}$, we can determine which is the speaker originating the test utterance by computing:

$$\hat{S} = arg \max_{1 \le s \le L} P(\lambda_s|T) = arg \max_{1 \le s \le L} \frac{P(T|\lambda_s)P(\lambda_s)}{P(T)} \quad \textbf{Eq. (3-50)}$$

Assuming $P(\lambda_s) = {}^{1}/_{S}$ and *P(T)* equal for all speakers, then:

$$\hat{S} = arg \max_{1 \le s \le L} P(T|\lambda_s) \quad \textbf{Eq. (3-51)}$$

And finally, assuming that feature vectors in *T* are independent, we get:

$$\hat{S} = arg \max_{1 \le s \le L} P(T|\lambda_s) = arg \max_{1 \le s \le L} LL_{avg}(T|\lambda_s) \quad \textbf{Eq. (3-52)}$$

Going back to the training phase, one approach extensively used to maximise the average log-likelihood function is the *Expectation Maximisation* (EM) algorithm [Dempster,1977], [Jeff Wu,1983].

### 3.3.1 The expectation maximisation algorithm

In this section we are not going to provide a deeper review or a theoretical background of this algorithm, but just a valid approximation for the particular case of the Gaussian mixture model. The interested reader can find a deeper review, for instance, in [Bishop,2006].

As previously stated, given a Gaussian mixture model $\lambda_s = \{w_i, \mu_i, \Sigma_i\}_{i=1,...,M}$, the goal of the EM algorithm is to iteratively maximise the likelihood function with respect to the mean, covariance and weights parameters. In each iteration *k*, the log probability is increased with respect to the previous iteration:

$$LL_{avg}(X|\lambda_s^k) > LL_{avg}(X|\lambda_s^{k-1}) \quad \textbf{Eq. (3-53)}$$

After providing an initialisation for the target parameters, $\lambda_s^0 = \{w_i^0, \mu_i^0, \Sigma_i^0\}_{i=1,...,M}$, the iterative algorithm consists in applying the following steps:

1. Expectation step or E step → In the expectation step, the a posteriori probability of a specific acoustic class, *i*, or mixture is evaluated for each feature vector $x_n$:

$$Pr(i|x_n) = \frac{w_i^k p_i^k(x_n)}{\sum_{j=1}^{M} w_j^k p_j^k(x_n)} \quad \textbf{Eq. (3-54)}$$

2. Maximisation step or M step → In the maximisation step, we re-estimate the model parameters using the a posteriori probabilities evaluated in the expectation step.

$$w_i^{k+1} = \frac{1}{N} \sum_{n=1}^{N} Pr(i|x_n) \quad \textbf{Eq. (3-55)}$$

$$\mu_i^{k+1} = \frac{\sum_{n=1}^{N} Pr(i|x_n) x_n}{\sum_{n=1}^{N} Pr(i|x_n)} \quad \textbf{Eq. (3-56)}$$

$$\Sigma_i^{k+1} = \frac{\sum_{n=1}^{N} Pr(i|x_n)\left(x_n - \mu_i^{k+1}\right)\left(x_n - \mu_i^{k+1}\right)^T}{\sum_{n=1}^{N} Pr(i|x_n)} \quad \textbf{Eq. (3-57)}$$

In the particular case in which diagonal covariances are used, obviously only diagonal elements of the covariance matrices need to be updated. Under this assumption, the $d^{th}$ diagonal element, $\sigma_i^2(d)$, of the covariance matrix of the $i^{th}$ Gaussian component, can be defined as:

$$\sigma^2{}_i^{k+1}(d) = \frac{\sum_{n=1}^{N} Pr(i|x_n) x_n^2(d)}{\sum_{n=1}^{N} Pr(i|x_n)} - \mu^2{}_i^{k+1}(d) \quad \textbf{Eq. (3-58)}$$

Where $x_n(d)$ and $\mu_i(d)$ refers to the $d^{\text{th}}$ element of $x_n$ and $\mu_i$, respectively.

3. Log likelihood evaluation→ The log likelihood is evaluated using the new model parameters, and if the convergence criterion is not satisfied a new iteration is executed.

$$LL_{avg}(X|\lambda_s^{k+1}) = \frac{1}{N}\sum_{n=1}^{N} \log \sum_{i=1}^{M} w_i^{k+1} p_i^{k+1}(x_n) \quad \textbf{Eq. (3-59)}$$

where

$$p_i^w(x_n) = \frac{1}{(2\pi)^{D/2}|\Sigma_i^w|^{1/2}} e^{-\frac{1}{2}(x-\mu_i^w)^T \Sigma_i^{w-1}(x-\mu_i^w)} \quad \textbf{Eq. (3-60)}$$

D being the dimension of the feature vector used for training the model, and $N$ the number of feature vectors in the training set.

Although extensively used, the EM algorithm has also its drawbacks. The log likelihood function could have multiple local maxima, but the EM algorithm does not guarantee that the largest of these maxima is going to be found. Additionally, severe over-fitting may occur in the cases where a Gaussian of the mixture collapses onto a data point. This problem arises most of times under two specific circumstances: when clusters contain a few observations or when too many components are used to fit a data set which can actually only be split in fewer clusters. There are some approaches that can be used to avoid these singularities, such as the use of heuristics to reset the collapsed Gaussians or to adopt a Bayesian approach. For instance, [Ormoneit,1995] proposed the use of a Bayesian regularisation method to deal with singularity condition, in which the covariance matrix is evaluated in the maximisation step as:

$$\Sigma_i^{k+1} = \frac{\sum_{n=1}^{N} Pr(i|x_n)\left(x_n - \mu_i^{k+1}\right)\left(x_n - \mu_i^{k+1}\right)^T + \beta I^d}{\sum_{n=1}^{N} Pr(i|x_n) + 1} \quad \textbf{Eq. (3-61)}$$

Where $I^d$, is a d-dimensional identity matrix and $\beta$ is a regularisation parameter that can be obtained using a validation set or by appropriate Bayesian methods. [Reynolds,1995-A] proposed the use of a minimum value for each variance element after each EM-iteration, just in case the variance element had a lower value.

Two additional and important issues related with this algorithm remain unsolved. The first one refers to the initialisation of the target parameters. One approach can be the use of the k-means algorithm to find an initial set of clusters given a specific set of training data [Chu,2001]. The main reason to do that is because the k-means algorithm uses less number of iterations to reach convergence and is, computationally speaking less demanding than EM. Another solution can be the use of a set of randomly selected training vectors as initial model means, and an identity matrix for the initial covariance matrix. HMMs have also been used to split training data into phonetic classes. Class means and variances of these phonetic classes are used as initial model for EM iteration. However, according to an experiment reported in [Reynolds,1995-B], no significant performance differences have been found between these different initialisation methods.

In relation to the number of iterations, no theoretical solution has been proposed. Although it is not usual to find more than ten EM iterations in state-of-the-art systems, the best practice is to optimise the number of EM iterations for each task, as stated by [Kinnunen,2010].

### 3.3.2 Background model adaptation

As previously said, the goal in speaker recognition is to determine whether a given segment of speech *X*, represented as a set of D-dimensional training vectors $X=\{x_1,...,x_N\}$ has been produced by a specific speaker *S*, represented in this case by a GMM ($\lambda_S$). This problem can be reformulated as a basic hypothesis test between $H_T$ (*X* was generated by *S*) and $H_F$ (*X* was not generated by *S*). In order to decide between these hypotheses, an average log-likelihood ratio test can be performed:

$$LL_{avg}(X,H_T,H_F)=\frac{P(X|H_T)}{P(X|H_F)}=\log P(X|\lambda_S)-\log P(X|\lambda_{\bar{S}})$$
$$=\frac{1}{N}\sum_{n=1}^{N}\left(\log\left(p(x_n|\lambda_s)\right)-\log\left(p(x_n|\lambda_{\bar{s}})\right)\right) \qquad \textbf{Eq. (3-62)}$$

where $\lambda_{\bar{S}}$ can be regarded as the model that represents the entire space of possible alternatives to the target speaker. From this test one conclusion can be extracted: we need to create a model (a GMM in this case) for the corresponding anti-target model, universal background model (UBM) or also known as world model. The process of training the background model is analogous to the process described for a single speaker, but in this case the training data is a set of speech pronounced by people who, in theory, will not be present on the system. The training data must be large enough to represent the speaker-independent distribution of the feature vectors used to create the model. According to [Reynolds,2000] although there is no objective measure to determine the amount of speakers and speech from each speaker present in the UBM, it is clear that it must contain "*speech that is reflective of the expected alternative speech to be encountered during recognition*". For example, we have to take into account quality of speech, transmission type (telephone line, GSM, microphone, etc.), gender (male, female), age (whether the target speaker may be senior, middle age or youngsters), etc.

Many researchers ([Reynolds,2000], [Pelecanos,2000], [Hou,2003], [Liu,2006-A], [Karam,2007], [Gonzalez-Rodriguez,2007], [Kinnunen,2009]) have also proposed the used of adaptation methods, in order to use the trained UBM as starting point when training the target speaker. One of the reasons to do that relies on the lacking of training

material for target speakers, under specific circumstances. For example, user acceptance of speaker recognition systems will considerably decrease as enrolment time increases.

Different methods have been proposed to adapt the speaker's model from the UBM, such as Maximum a Posteriori (MAP), Maximum Likelihood Lineal Regression (MLLR), Reference Speaker Weighting (RSW) or eigenvoices, among others. In [Krishna,2008], [Mak,2006] and [Mariethoz,2001] a comparative study is carried out among these adaptation methods for speaker verification tasks. However, in the later this comparison is carried out in the specific case where limited amount of information is available for adaptation (speaker model training). In this case, MLLR or eigenvoices show better performance compared to other methods. Let's make a quick review of how the most popular adaptation methods work:

- MAP: As previously said, the goal of adaptation is to derive the speaker model from the UBM ($\lambda_{UBM}=\{w_i,\mu_i,\Sigma_i\}_{i=1,\dots M}$), given the training data set $X=\{x_1,\dots,x_N\}$. The adaptation can be formulated as a two-step process. In the first step, the probabilistic alignment of the training vectors into each of the UBM components are computed as:

$$Pr(i|x_n) = \frac{w_i p_i(x_n)}{\sum_{j=1}^{M} w_j p_j(x_n)} \qquad \textbf{Eq. (3-63)}$$

These values are then used to estimate the statistics as weights, mean and variance parameters:

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \qquad \textbf{Eq. (3-64)}$$

$$E_i = \frac{\sum_{t=1}^{T} Pr(i|x_t)x_t}{\sum_{t=1}^{T} Pr(i|x_t)} = \frac{1}{n_i}\sum_{t=1}^{T} Pr(i|x_t)x_t \qquad \textbf{Eq. (3-65)}$$

$$E_i^2 = \frac{\sum_{t=1}^{T} Pr(i|x_t)x_t x_t^{'}}{\sum_{t=1}^{T} Pr(i|x_t)} = \frac{1}{n_i}\sum_{t=1}^{T} Pr(i|x_t)x_t^2 \qquad \textbf{Eq. (3-66)}$$

These statistics are used to update the old background model parameters for each mixture i, through the use of the following equations:

$$w_i^{new} = [\alpha_i n_i/N + (1-\alpha_i)w_i]\gamma \qquad \textbf{Eq. (3-67)}$$

$$\mu_i^{new} = \alpha_i E_i + (1-\alpha_i)\mu_i \qquad \textbf{Eq. (3-68)}$$

$$\sigma_i^{new\,2} = \alpha_i E_i^2 + (1-\alpha_i)(\sigma_i^2 + \mu_i^2) - \mu_i^{new\,2} \qquad \textbf{Eq. (3-69)}$$

where $\gamma$ is a scale factor computed over all adapted mixture weights to ensure that they still satisfy the constraint: $\sum_{i=1}^{M} w_i = 1$. The adaptation coefficient, $\alpha_i$, can be defined as:

$$\alpha_i = \frac{n_i}{n_i + r} \qquad \textbf{Eq. (3-70)}$$

where *r* is a relevance factor that somehow controls how much new data should be observed in a mixture before the new parameters replace the old ones.

Although MAP allows the adaptation of all mixture coefficients, it has been empirically shown that the adaptation of only mean parameters provides the best performance [Reynolds,2000].

- MLLR: Originally developed for speech recognition and especially for HMM adaptation, this algorithm has also been successfully applied to speaker recognition. MLLR proposes to constrain the means and variances of the Gaussians that compose the GMM of a specific speaker to be linear combinations of the means and Gaussians of the UBM. However, for practical reasons, as in the case of MAP, most of the times only the means are adapted as proposed in [Leggetter,1995-A], [Leggetter,1995-B]. Assuming that $\lambda_{UBM}=\{w_i,\mu_i,\Sigma_i\}_{i=1,\ldots M}$ represents the UBM and given the training data set $T=\{t_1,\ldots,t_N\}$ where n is the dimensionality of the training vectors, the adaptation of the mean is achieved by applying the transformation:

$$\mu_i^{ss} = A_{in}\mu_i^{ubm} + b_{in} \qquad \textbf{Eq. (3-71)}$$

where the matrix $A_{in}$ ($n$ x $n$) and the vector $b_n$ are parameters to be found by maximizing the likelihood of the specific speaker data. This transformation can be also expressed as:

$$\mu_i^{ss} = W_i\xi_i^{ubm} \qquad \textbf{Eq. (3-72)}$$

where $W_i$ is the $\{n\ x\ (n+1)\}$ transformation matrix and $\xi_i^{ubm}$ is the extended mean vector $\xi_i^{ubm} = [1 \quad \mu_{i1} \quad \ldots \quad \mu_{in}]^T$. Obviously, $W_i=[b_i\ A_i]$. The aim of the MLLR adaptation is to find the transformation $W_i$ that maximises the likelihood of the training data. This transformation matrix can be obtained through the application of the EM algorithm as presented in [Leggetter,1995-A] or [Gales,1996]. In the last case an extension to adapt variances is also presented.

If we strictly apply MLLR, the number of parameters to be adjusted is higher than in standard Maximum Likelihood as for each mean vector of each Gaussian in the GMM, a matrix of size $n\ x\ (n+1)$ needs to be adjusted. However, for practical reasons, a typical approach to this problem consists in using a set of regression classes. In this case, a set of Gaussians of the model (regression class) are grouped together and forced to share the same transformation parameters. In order to establish these regression classes different methods have been proposed, for instance, regression class trees based on Euclidean similarity between clusters [Gales,1996], or regression classes based on the phonetic similarity of tri-phone models as in [Ferras,2009].

Both the MAP and MLLR adaptations form a basis for the recently proposed supervector classifiers.

An additional benefit of using the GMM-UBM framework is that the computational effort can be reduced if only top scoring Gaussians from the UBM are used to evaluate the score. As described in Eq. (3-62), to produce a score of whether the test utterance belongs to a specific speaker model, it is necessary to compute a frame-by-frame matching. However, as described in [Reynolds,2000], [Tydlitat,2007] if we evaluate the *LLR* for the world model, but only evaluate the corresponding top distributions of the UBM in the target speaker, performance can be greatly improved without punishing recognition rates.

### 3.3.3 Model Order

Another important issue in Gaussian mixture modelling, that still remains an open problem, is the selection of the number of components or mixture in the model. Choosing the adequate number of components is essential if a useful model is going to be created, as too many components will lead to an over-fitted model with singular covariance matrices, while on the other hand the use of too few components will lead to a model that do not capture accurately the structure of the data.

Different approaches have been tested over the last years to determine the number of components. For instance, [Akaike,1974] proposed an extension of the maximum likelihood principle for the general problem of selecting the appropriate model among a set of possible candidates. The main idea consist in finding the model $\lambda_{i \in 1..n}$ that maximises the function log $M_i$ - $k_i$ where $k_i$ is the dimension of the model and $M_i$ is the maximised value of the likelihood function for the estimated model $\lambda_i$. This method is also known as AIC or Akaike's Information Criterion.

Approximate Bayes factors [Kass,1995] can also be used to compare models when the mixture-model approach is used as clustering method. In the particular case in which the EM is used to find the maximum mixture likelihood, the Bayesian Information Criterion or BIC defined by Schwarz [Schwarz,1978] can be used to find the most suitable model. The BIC can be defined as:

$$BIC = 2ln\zeta(X,\lambda) - m_\lambda \log(n) \qquad \textbf{Eq. (3-73)}$$

where $\zeta(X,\lambda)$ is the maximised value of the likelihood for the model $\lambda$, given the observed data $X$, $m_\lambda$ is the number of independent parameters to be estimated in the model (or dimension of the model) and n is the number of observations in the observed data. The BIC assumes that the data points in the observed data are independent and identically distributed. However, this assumption may or may not be valid depending on the available dataset. The BIC value can be interpreted in the sense that the larger the obtained value the stronger the evidence of the model being originated from the observed data. Compared to AIC, the BIC tends to favour simpler models as it penalises more heavily model complexity.

The Bayesian Information Criterion has been used under different a priori assumptions. For example, [Fraley,1998] apply EM and BIC to determine the best model given a maximum number of mixture components, *M*. On the other hand [Cheng,2004], propose an iterative method starting with a single component and successively splitting a selected component into two new components until the appropriate number of components is found. In this case, the BIC is used to select the component that is going to be split in the next iteration.

In the same year in which Schwarz presented the BIC, Rissanen [Rissanen,1978] also presented the Minimum Description Length (MDL) criteria, which is formally identical to the BIC.

More recently [Abu El-Yazeed,2004], proposed the use of a goodness of fit (GOF) measure to decide whether the training data fits well into a GMM distribution or not for the specific case of text independent speaker identification. As in the case of [Cheng,2004], an iterative method is proposed starting with a model order *M=1*. In each step, the GOF for each of the clusters in the model is evaluated as well as the GOF of the model, each one defined as:

$$GOF_i = \frac{\left({}^{1}/_{T_i}\right)\sum_{x_t \in c_i}\left((x_t-\mu_i)^T\Sigma_i^{-1}(x_t-\mu_i)\right)^2}{D^2+2D} \quad \textbf{Eq. (3-74)}$$

$$GOF_{\lambda_M} = 1 - \frac{1}{M}\sum_{i=1}^{M}|1-GOF_i| \quad \textbf{Eq. (3-75)}$$

where,

$$D^2+2D = E\left\{\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)^2\right\} \quad \textbf{Eq. (3-76)}$$

The iterative algorithm finishes when the prior probability of at least one cluster of the model does not exceed a specific threshold. At this point the optimum model order is determined by:

$$M_{opt} = arg\max_{k=1,2,\ldots,M-1} GOF_{\lambda_k} \quad \textbf{Eq. (3-77)}$$

The performance of the proposed algorithm has been compared with MDL and AIC algorithms, achieving similar identification accuracy in text-independent speaker identification experiments.

An alternative method, kwon as Minimum Message Length or MML first described in [Wallace,1968], was compared in [Oliver,1996] to other methods such as Partition Coefficient (PC), AIC, ICOMP and BIC.

According to the MML criterion, we can select a specific model $\lambda_i$ among a set of models $\lambda_{k=1\ldots K}$, as the one that best describes the training data, *X*, if this model minimises the following value:

$$MML_{value} = MsgLen(\lambda_i) + MsgLen(X|\lambda_i) \quad \textbf{Eq. (3-78)}$$

where the first term represents the length of the asserted model, and the second term corresponds to the expected length of the training data under the assumption that the parameter estimates of the model are the true values. A deep review on MML applied to mixture models can be found in chapter 6 of [Wallace,2005].

Finally, [Tadj,1998], work on the idea that the model order used to describe a specific speaker is related with the amount of information available for training the model. Tadj reports some improvements when each speaker is modelled with a different model order according with the amount of data available for training, in contrast with the results

achieved when all the speakers in the experiment use the same model order. Two different methods were proposed to establish the relationship between the model order and the amount of available training data. One of the methods is based on the use of a nonlinear transformation with different parameters determined empirically, while the other method relies on exhaustive experimentation in order to establish a linear relation between speech signal duration and model order. However, in the case of text-independent speaker identification, the number of Gaussian components required to accurately model a speaker relies on the data distribution rather than its amount as stated in [Abu El-Yazeed,2004].

### 3.3.4 Phonetic GMM variations

One of the main problems when modelling speakers with GMMs is that, as long as this model does not explicitly use any phonetic information; the match score in the test phase may be biased due to the high probability of finding different phonetic classes in training and testing utterances. In order to provide a possible solution to this problem, different phonetic approaches have been proposed based on GMMs.

For instance, [Chaudhari,2003] proposed the use of a tree structure, where each speaker is modelled at three levels of detail. The root of the tree contains all the phones, the second level partitions the phones into six linguistic classes (vowels, nasals, voiced and unvoiced fricatives, plosives, liquids) plus silence, while the last level is comprised of the individual phones.

Another approach consists in modelling each speaker using different GMMs for different phonetic classes. Following this concept we can find in the literature different approaches. [Faltlhauser,2001] presented the so-called Phonetically Structured GMM, in which separate GMMs trained on separate phonetic classes (nasals, fricatives, different vowels, liquids, plosives or subclasses of them) are combined using specific weights. Although phonetic labelling is needed during training, in the test phase all speech frames from the test utterance are scored against the combined model. In a similar way [Castaldo,2007] described the phonetic GMM (PGMM). Like in the previous case, a GMM is trained for each of 11 language-independent broad phone classes, which together constitute the speaker model. However, during the test phase, each audio segment is first phonetically labelled and then scored against its corresponding GMM.

[Sturim,2002] and [Bocklet,2009] face the problem from a more general point of view. In the case of [Bocklet,2009], a GMM is trained for each of the eight syllable groups (syllable onsets, syllable nuclei, syllable codas, syllables following pauses, one-syllable words, syllables containing [N], syllables containing [T] and syllables containing [B], [P], [V], or [F]) and in the test phase, scores for each subsystem are combined by Linear Logistic Regression (LLR). [Sturim,2002] presented the text-constrained system, in which for each speaker a set of GMM-UBM system is defined, each of them representing a specific word.

The recognition rates achieved by the different studies already presented clearly indicate that phonetic modelling in GMMs is worth of further study.

### 3.3.5 Available software tools

Like in the case of Vector Quantisation and HMMs, there are some open-source software tools available on-line that implement different algorithms for building GMMs.

Probably one of the most known and extensively used open-source software tool for Gaussian Mixture Modelling in speaker recognition applications is ALIZE [Bonastre,2005], [Bonastre,2008]. Developed within the project MISTRAL funded by the French National Research Agency (ANR), ALIZE library is available on different operating systems (Linux, Windows, Mac OSX) and provides a modular and easy to use platform capable of managing different biometric tasks (feature extraction, model training, test, normalisation, etc.). The training process is carried out using the EM algorithm in the case of the train background task, while the adaptation of the background model to a specific target model can be carried out through two different Maximun A Posteriori algorithms (MAPOccDep/ MAPConst). ALIZE is widely used both in the academic and industrial community and has obtained good results in major international evaluation campaigns NIST SRE (Speaker Recognition Evaluation) and ESTER (evaluation of automatic radio broadcast transcription systems for French language). The ALIZE library is available for download at (URL:ALIZE), whereas additional information can be found at the project's official web page (URL:MISTRAL) and the recently created wiki (URL:wiki-mistral).

MATLAB (URL: MathWorks) is one of the most widely used software products for algorithmic development not only in academic and research institutions but in the industry world as well. Although not computationally efficient for massive data processing, MATLAB provides a powerful programming language and easy to use environment for rapid prototyping and visualisation. Moreover, they include in the software package the *gmdistribution* class, which allows us to define a Gaussian mixture distribution.

Other solutions can be found in the Web as open-source projects. For instance, the Gaussian Mixture Model and Regression project (*GMM-GMR*) registered on SourceForge.net on 2008 (URL: GMM-GMR). As described by the project team, *GMM-GMR* is a light package of functions in C/C++ to compute Gaussian Mixture Models and Gaussian Mixture Regression, allowing the user to evaluate a GMM from any dataset, and to retrieve partial data by specifying the desired inputs (GMR).

## 3.4 SUPPORT VECTOR MACHINES

After its introduction as maximum margin classifier in 1992 [Boser,1992], Support Vector Machines (SVM), which are based upon the structural risk minimisation principal of Vapnik [Vapnik,2000], have become more and more popular for solving classification problems, especially due to robustness and good generalisation performance to classify unseen data. Particularly, they have been applied in speaker recognition with different features such as spectral [Fauve,2007], or high-level [Ferrer,2007-A] – syllable-based prosodic features – or [Campbell,2007] – n-gram frequencies.

An SVM is a discriminative two-class classifier whose purpose is to model the boundaries between two classes as a separating hyperplane. In the specific case of speaker recognition, a subtle distinction must be made between speaker verification and speaker identification. In the first case, one class (the target speaker class) consists of the target feature vectors, while the other class (the background class) consists of the training feature vectors from a set of impostors (see Figure 3-4). Using these labelled training vectors, an SVM will find a separating hyperplane between these two classes. Usually, for system simplicity, the impostor population is kept constant for every speaker target enrolled in the system. In the case of speaker identification, the focus is on determining the identity of a target speaker from a group of speakers. From this point

of view, two different approaches can be followed. The one-versus-all strategy, consist in considering the target speaker as one class, and the remaining non-target speakers as the alternative class. This operation is performed for all speakers to obtain a set of speaker identification models. However, this approach presents the problem that every new speaker incorporation involves retraining the whole set of speaker models. To overcome this problem an alternative approach consists in building as many classifiers as speakers using a fixed background set for all the speakers. In both cases, if there are *n* speakers then *n* classifiers must be constructed.

In the following sections a brief review on SVM theory will be presented. The interested reader may refer to one of the following references for a more technical review of Support Vector Machines for pattern recognition [Burges,1998], [Cristianini,2000], [Schölkopf,2001], [Vapnik,2006]. Additionally [Vapnik,2000] provide a deeper insight on Support Vector Machines and other kernel-based learning methods.

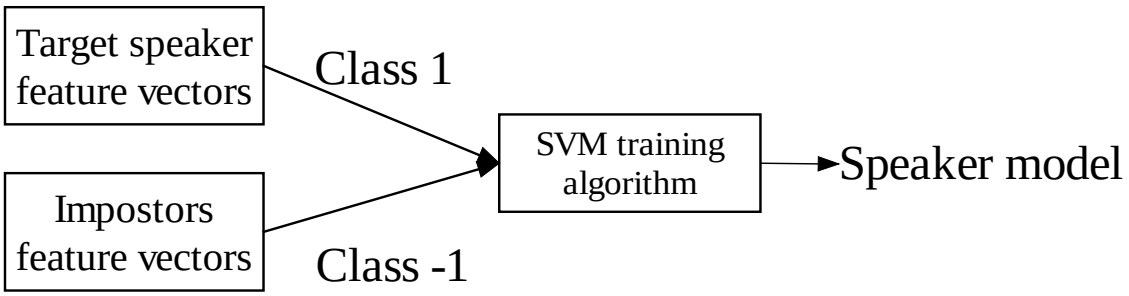



**Figure 3-4** SVM training strategy for speaker recognition

### 3.4.1 Basic SVM theory

At this point we assumed that we have a training set, $T = \{(x_i, y_i)\}_{i=1}^{l}$, where $x_i$ denotes the training vectors and $y_i$ denotes the corresponding label/class ({-1,+1}) assigned to the feature vector. On the simplest case, when dealing with linearly separable classes, the objective is to find the hyperplane that provides the maximum margin between the two classes [Vapnik,2000].

The separating hyperplane will have the form: $w \cdot x + b = 0$, where *w* is the normal vector to the hyperplane, and the training data will satisfy the following constraints:

$$\left.\begin{matrix} x_i \cdot w + b \geq 1, y_i = +1 \\ x_i \cdot w + b \leq 1, y_i = -1 \end{matrix}\right\} \rightarrow y_i(x_i \cdot w + b) \geq 1, \forall i \qquad \textbf{Eq. (3-79)}$$

So in this context, the decision function of the SVM is given by:

$$f(x) = sgn(w \cdot x + b) \qquad \textbf{Eq. (3-80)}$$

and the margin is thus:

$$m = \frac{1}{\|w\|^2} \qquad \textbf{Eq. (3-81)}$$

As a result, we can find the maximum margin by minimizing $\|w\|^2$, subject to constraints defined by Eq. (3-79), i.e.

$$\min_{w,b} \frac{1}{2}\|w\|^2, subject\ to\ y_i(w \cdot x_i + b) \geq 1, \forall i \qquad \textbf{Eq. (3-82)}$$

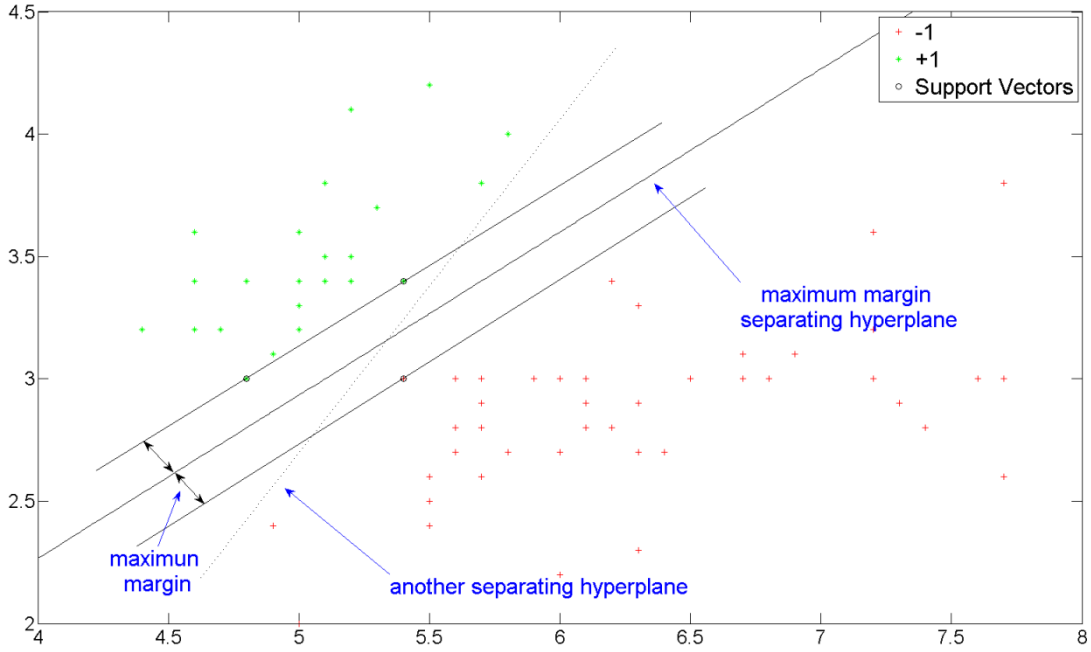



**Figure 3-5** Maximum margin hyperplane that separates two linearly separable classes

In order to ease the practical use of constraints while forcing the training data to appear only in the form of dot products between vectors, the Lagrangian formulation (with positive Lagrange multipliers $\alpha_i$, for each of the inequality constraints, $i=1,..,l$) of the problem can be introduced, resulting in:

$$L_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^{l} \alpha_i \qquad \textbf{Eq. (3-83)}$$

In this Lagrangian formulation, the aim is to minimise $L_P$ with respect to w, b, and subject to the constraints that derivatives of $L_P$ with respect to all $\alpha_i$ vanish and $\alpha_i \geq 0, \forall i$. This optimisation problem can be transformed into a dual form, where the objective is to maximise:

$$L_D \equiv \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \qquad \textbf{Eq. (3-84)}$$

Subject to the constraints:

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \textbf{Eq. (3-85)}$$

$$\alpha_i \geq 0, \forall i$$

Once the Lagrange multipliers have been determined, the normal vector w, can be derived (the threshold value b will be analise later on):

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \qquad \textbf{Eq. (3-86)}$$

In this context, the hyperplane decision function can be written as:

$$f(x) = sgn(\sum_{i=1}^{m} y_i \alpha_i (x \cdot x_i) + b) \qquad \textbf{Eq. (3-87)}$$

From Eq. (3-87) we can state that the normal vector has an expansion in terms of a subset of the training patterns, $T = \{(x_i, y_i)\}_{i=1}^{l}$, specifically those patterns whose Lagrange multiplier is non-zero. This subset of training patterns are called *Support Vectors*, which actually are the data points that lie at the border between the two classes. The support vectors, $x_i$, the Lagrange multipliers, $\alpha_i$, for each support vector and the bias term, $b$, are all obtained from the training set solving a Quadratic Programming (QP) problem.

However, in real-life classification problems, data is not always linearly separable. In this case, [Vapnik,2000] introduce the use of slack variables, $(\xi_i, i = 1, \dots, l)$, which somehow measure the distance between a non-correct classifying training feature-vector and a margin. Using these slack variables we are allowing misclassification cases during the training phase. This approach is also known as soft-margin SVM.

The slack variables relax the hard-margin constraints, resulting in:

$$y_i(x_i \cdot w + b) \geq 1 - \xi_i, \forall i = 1, \dots, l \qquad \textbf{Eq. (3-88)}$$

In this case, the maximum margin classifier can be found by minimizing $\|w\|^2$, and an upper bound on the number of training errors, $\sum_{i=1}^{l} \xi_i$, i.e.:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i \, , subject\ to\ y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall i = 1 \dots l \qquad \textbf{Eq. (3-89)}$$

where $C > 0$ is a regularisation constant that trades-off margin width vs. the number of misclassifications, a larger $C$ corresponding to a higher misclassifications penalty. Rewriting again in terms of Lagrange multipliers, this leads to the dual problem in Eq. (3-84) subject to the following constraints:

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \textbf{Eq. (3-90)}$$

$$0 \leq \alpha_i \leq C, \forall i = 1, \dots, l$$

Thus the only difference is that we have established an upper bound for the Lagrange multipliers. An alternative to the C-SVM (already presented) is the so-called *v*-SVM [Chen,2005], [Schölkopf,2000]. This is a soft margin variant using *v*-parameterisation, where the primal problem can be formulated as minimising:

$$L_P \equiv \frac{1}{2}\|w\|^2 - v\rho + \frac{1}{l}\sum_{i=1}^{m} \xi_i \qquad \textbf{Eq. (3-91)}$$

subject to

$$y_i(w \cdot x_i) + b \geq \rho - \xi_i, i = 1, \dots, l \qquad \textbf{Eq. (3-92)}$$

And

$$\rho \geq 0, \xi_i \geq 0 \qquad \textbf{Eq. (3-93)}$$

where *v* is the upper bound on the fraction of errors and at the same time the lower bound on the fraction of *Support Vectors*.

An additional problem appears when the decision function is not a linear function of the data. In order to make this classification method useful for real-life applications, we need to find a way to deal with non-linear separable data. To address this problem, the kernel concept was introduced [Boser,1992]. Intuitively, the main idea behind this concept is that using a mapping from the input space to a high-dimensionality space, two classes (non-linearly separable in the input feature space) can be easily separable with a linear hyperplane in the high-dimensional space. This high-dimensional mapping could be problematic due to the curse of dimensionality; however, the SVM training criterion deals effectively with the problem [Campbell,2007].

Let $\Phi: \mathbb{R}^N \to \mathcal{F}$ be a nonlinear mapping, so that data $x_1 \dots x_n \in \mathbb{R}^N$ is mapped into a potentially much higher dimensional feature space $\mathcal{F}$(also referred as Hilbert space)*:*

$$\begin{aligned} \Phi: \mathbb{R}^N &\to \mathcal{F} \\ x &\to \Phi(x) \end{aligned} \qquad \textbf{Eq. (3-94)}$$

In this way, the learning algorithm, will work on the training set, $T = \{(\Phi(x_i), y_i)\}_{i=1}^{l}$, instead on the one presented above: $T = \{(x_i, y_i)\}_{i=1}^{l}$. Taking into account that the training data only appears in the training procedure in the form of dot products, then in this new scenario, the training algorithm would only depend on the data through dot products in $\mathcal{F}$, i.e. $\Phi(x_i)\Phi(x_j)$. For certain feature spaces $\mathcal{F}$, and corresponding mappings Φ, we can compute the scalar product using *kernel functions*:

$$K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \qquad \textbf{Eq. (3-95)}$$

So we do not need to explicitly know Φ. In this new context, the hyperplane decision function can be written as:

$$f(x) = sgn(\sum_{i=1}^{m} y_i \alpha_i K(x, x_i) + b) \qquad \textbf{Eq. (3-96)}$$

The next question that must be answered is how this non-linear mapping affects the solution to the optimisation problem of finding *w*. Although the nonlinearities alter the quadratic form, the dual optimisation problem is still quadratic in $\alpha$, thus the objective will be still to maximise:

$$L_D \equiv \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \quad \textbf{Eq. (3-97)}$$

Subject to constraints defined by Eq. (3-85).

### 3.4.2 Kernel functions

We have introduced the concept of kernel function to deal with non-linear separable data. However, we must consider now how to determine the kernel functions, *K*, which corresponds to a dot product in some feature space, $\mathcal{F}$. In this case, a function, *K*, which satisfies Mercer's condition is supposed to be a valid kernel function [Mercer,1909], [Vapnik,2000]. Mercer's condition establishes that a space $\mathcal{F}$ and a mapping $\Phi: \mathbb{R}^N \rightarrow \mathcal{F}$ such that Eq. (3-95) holds, exists if given a Hilbert space, $L_2(C)/C \subset \mathbb{R}^N$:

$$\forall f \in L_2(C): \int_{CxC} K(x,y) f(x) f(y) dx dy \geq 0 \quad \textbf{Eq. (3-98)}$$

We should point out that Mercer's condition does not specify how to construct Φ or even what $\mathcal{F}$ is. However, some simple rules for composing kernels, also satisfying Mercer's condition, can be derived [Smola,2004]:

- Positive linear combinations of kernels result in an admissible kernel, i.e., given two valid kernels $k_1$, $k_2$:

$$K(x,y) = c_1 k_1(x,y) + c_2 k_2(x,y), \forall c_{1,} c_2 \geq 0 \quad \textbf{Eq. (3-99)}$$

  is a valid kernel.

- Product of admissible kernels results in an admissible kernel, i.e., given two valid kernels $k_1$, $k_2$:

$$K(x,y) = k_1(x,y) k_2(x,y) \quad \textbf{Eq. (3-100)}$$

  is a valid kernel.

- If there exists a function, *s(x,y)*, on $\chi$ x $\chi$, such that:

$$K(x,y) = \int_{\chi} s(x,z) s(y,z) dz \quad \textbf{Eq. (3-101)}$$

  Then *K* is an admissible kernel.

- A translation invariant kernel *K(x,y)=K(x-y)* is an admissible kernel if and only if

$$F[K](\varpi) = (2\pi)^{-\frac{d}{2}} \int_{\chi} e^{-i\langle \varpi, x \rangle} K(x) dx \geq 0 \quad \textbf{Eq. (3-102)}$$

The simplest kernel function satisfying Mercer's condition is the linear kernel, which is the identity mapping for the feature space mapping ( $\Phi(x) = x$), leading to:

$$K(x,y) = x^T y \qquad \textbf{Eq. (3-103)}$$

However, there are numerous forms of kernel functions commonly used [Vapnik,2000], [Vapnik,1996], apart from the linear kernel, such as:

- The polynomial kernel:

$$K(x,y) = \left((x^T y) + 1\right)^d \qquad \textbf{Eq. (3-104)}$$

- Full polynomial kernel:

$$K(x,y) = \left(\frac{(x^T y)}{a} + b\right)^d \qquad \textbf{Eq. (3-105)}$$

- Radial Basis Function (RBF):

$$K(x,y) = exp(-\gamma|x-y|^2) \qquad \textbf{Eq. (3-106)}$$

- Two layer neural network, also kwon as sigmoidal kernel:

$$K(x,y) = \tanh(a(x^T y) - b) \qquad \textbf{Eq. (3-107)}$$

- Linear splines: The inner product that generates splines of n-order on one dimension is:

$$K(x,y) = \sum_{r=0}^{n} x^r y^r + \sum_{s=1}^{N} (x - t_s)_+^n (y - t_s)_+^n \qquad \textbf{Eq. (3-108)}$$

  where:

$$(x-t)_+ = \max\{(x-t), 0\}, t_1, \ldots, t_N \in [0,1] \qquad \textbf{Eq. (3-109)}$$

  For the specific case of linear splines of 1st-order, with an infinite number of points, the following generating kernel is derived:

$$K(x,y) = 1 + xy + xy\min(x,y) - \frac{(x+y)}{2}\left(min(x,y)\right)^2 + \frac{\left(min(x,y)\right)^3}{3} \qquad \textbf{Eq. (3-110)}$$

  In the case of dealing with *m*-dimensional splines, the generating kernel is the product of *m* one-dimensional generating kernels:

$$K(x,y) = \prod_{k=1}^{m} K_k(x^k, y^k) \qquad \textbf{Eq. (3-111)}$$

- Kernels generating Fourier expansions:

The inner product in the following 2N+1 dimensional feature space:

$$\frac{1}{\sqrt{2}}, \cos x, \sin x, \ldots, \cos Nx, \sin Nx$$

Is defined by:

$$K(x,y) = \frac{1}{2} + \sum_{r=1}^{N} (\cos rx \cos ry + \sin rx \sin ry) = \frac{\sin(N+1/2)(x-y)}{\sin\frac{x-y}{2}} \quad \textbf{Eq. (3-112)}$$

However, two alternatives to this kernel have been proposed:

- Weaker mode regularised Fourier kernel:

$$K(x,y) = \frac{\pi}{2\gamma} \frac{\cosh\frac{\pi - |x-y|}{\gamma}}{\sinh\frac{\pi}{\gamma}} \quad \textbf{Eq. (3-113)}$$

  where $0 \leq |x\text{-}y| \leq 2\pi$ and $\gamma$ is user defined.

- Stronger mode regularised Fourier kernel:

$$K(x,y) = \frac{1-\gamma^2}{2(1-2\gamma\cos(x-y)+\gamma^2)} \quad \textbf{Eq. (3-114)}$$

  where $0 \leq |x\text{-}y| \leq 2\pi$ and $\gamma$ is user defined.

In the specific case of its application to speaker recognition, there are other alternatives based on sequence kernels, which have provided interesting results [Campbell,2007]. These sequence kernels, instead of modelling features from individual frames, aim at comparing speech utterances, i.e., model the entire sequence of feature vectors. Different approaches have been considered in the last years, which include the generalised linear discriminant sequence kernel [Campbell,2002-A], Fisher kernel methods [Fine,2001], [Wan,2003], [Jaakkola,1998], n-gram kernels [Campbell,2004-A], maximum-likelihood linear regression (MLLR) transform kernels [Stolcke,2005] and GMM supervector kernels [Campbell,2006], [Campbell,2006-A].

### 3.4.3 Training algorithms

So far we have presented the SVM basic theory for both linear and non-linear classifiers as well as for the separable and non-separable case, represented by the primal ($L_P$) and dual ($L_D$) functionals through Lagrangian formulation. No matter which case we are dealing with, the solution to the classification problem will be found by minimizing the $L_P$ or by maximizing $L_D$, subject to their respective constraints. More specifically, in order to solve the support vector optimisation problem, a convex quadratic problem (QP) must be solved, i.e., it is necessary to solve:

$$\max_{\alpha} \sum_{i} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad \textbf{Eq. (3-115)}$$

Under the following constraints:

$$0 \leq \alpha_i \leq C \qquad \textbf{Eq. (3-116)}$$

$$\sum_i \alpha_i y_i = 0 \qquad \textbf{Eq. (3-117)}$$

When the amount of training data is small enough or when the support vectors from the training data can be known in advance, this optimisation problem can be solved analytically. However, this situation does not usually occur on a regular basis; and for most real cases, these equations must be solved numerically. In this last case, it is important to note that as the objective function is convex, then every (local) maximum is already a global maximum. Moreover, there can be several optimal solutions which may lead to different testing performances [Muller,2001]. For this reason, [Burges,1998] introduced three basic aspects that must be taken into account when addressing a solution to this problem:

- The solution must satisfy the Karush-Kuhn-Tucker (KKT) optimality conditions, as these are necessary and sufficient conditions for the solution to be optimal. The KKT conditions for the dual SVM QP problem are particularly simple:

$$\begin{aligned} \alpha_i = 0 &\rightarrow y_i f(x_i) \geq 1 \; and \; \xi_i = 0 \\ 0 < \alpha_i < C &\rightarrow y_i f(x_i) = 1 \; and \; \xi_i = 0 \\ \alpha_i = C &\rightarrow y_i f(x_i) \leq 1 \; and \; \xi_i \geq 0 \end{aligned} \qquad \textbf{Eq. (3-118)}$$

  where *f(x)* is the decision function characterised by Eq. (3-96).

  From the analysis of these conditions we can state that only the Lagrange multipliers, $\alpha_i$, associated with training data located either on the margin or inside the margin area are non zero. Additionally, taken into account that for all support vectors inside the margin area the associated slack variable, $\xi_i$, is zero, then the KKT conditions provide a way of deriving the threshold $b$, in the decision function. For any support vector, $x_i / i \in I = \{i: 0 < \alpha_i < C\}$, holds

$$y_i \left( b + \sum_{j=1}^{l} y_j \alpha_j K(x_i, x_j) \right) = 1 \qquad \textbf{Eq. (3-119)}$$

  Averaging over these patterns a numerically stable solution can be found:

$$b = \frac{1}{|I|} \sum_{i \in I} \left( y_i - \sum_{j=1}^{l} y_j \alpha_j K(x_i, x_j) \right) \qquad \textbf{Eq. (3-120)}$$

  However, an alternative threshold [Platt,1998] can be used:

$$b = w \cdot x_k - y_k, for \; some \; \alpha_k > 0 \qquad \textbf{Eq. (3-121)}$$

- Define a strategy for approaching optimality by uniformly increasing the dual objective function subject to the constraints.
- As the amount of training data to handle can be quite large, it will be necessary to define a decomposition algorithm so that at any given time, only a small

fraction of these data will be considered in the search for a solution. Although, obviously, all training data must be taken into account in the process.

A review of literature on solving quadratic problems is out of the scope of this chapter; however, the interested reader may find a good starting point in [Smola,2004] where some free or commercial packages and corresponding references, are introduced. Nevertheless, some classical approaches will be presented:

- Chunking: This method proposed by [Vapnik,2006], is based on two important facts: only support vectors are important in the final solution which must satisfy the KKT conditions. Additionally, it uses the fact that the value of the quadratic form is the same if zero Lagrange multipliers are removed from the process. The algorithm starts selecting a small and arbitrary subset (chunk) of the training data, and solves the QP problem on that subset. The remaining training data are tested on the resulting classifier. In the next step, a new chunk is selected which contains all the support vectors already found (i.e. all non-zero Lagrange multipliers) and a set of the *M* (value decided heuristically) worst training cases that violate the KKT conditions. The process is continued until all data points are found to satisfy the KKT conditions. It should be noted, that the size of QP problems may vary through the process, but at the last iteration the entire set of non-zero Lagrange multipliers are identified and the QP problem is solved.

- Decomposition methods: The decomposition methods are similar to chunking in the sense that they decompose the original problem into a sequence of small QPs, but in this case the size of the sub-problems is fixed. One example of decomposition method was presented in [Osuna,1997]. Like in the case of chunking, only a subset of the training data is evaluated at each step, moreover; in theory only one training point is deleted and added (one that does not hold KKT conditions) at each step, thus keeping the size of the QP constant. This algorithm is proved to converge to the global optimum in a finite number of iterations, as at each step at least one training sample violating the KKT is added to the QP sub-problem. Thus solving the QP sub-problem will reduce the overall objective function. However, as pointed out by [Platt,1999-A], replacing just one training sample at each time will lead to a very slow and computationally inefficient solution. To overcome this problem multiple examples are replaced at each step.

- Sequential minimal optimisation (SMO): The SMO algorithm [Platt,1999-A] represents an extreme case of the decomposition method already presented, in which at each iteration only a QP problem of size two is solved. This fact constitutes both its main advantage and its main drawback. The main advantage because at each step the two-variable sub-problem can be analytically solved, thus no quadratic optimisation software is required. The main drawback, as the selection of two adequate variables in each iteration is not straight-forward. An additional aspect that must be taken into account is that SMO re-computes the threshold *b* after each step of the algorithm, based on the two Lagrange multipliers, which have been previously selected.

  Since at each step the algorithm selects and optimises two Lagrange multipliers, of which at least one does not satisfy the KKT conditions, the convergence is guaranteed as each step decreasing the objective function. The heuristics for selecting the two Lagrange multipliers for joint optimisation are based both on the KKT conditions and on the positive progress on the solution.

Figure 3-6 compares the size of the sub-set of training data used at each step of the different algorithms presented so far.

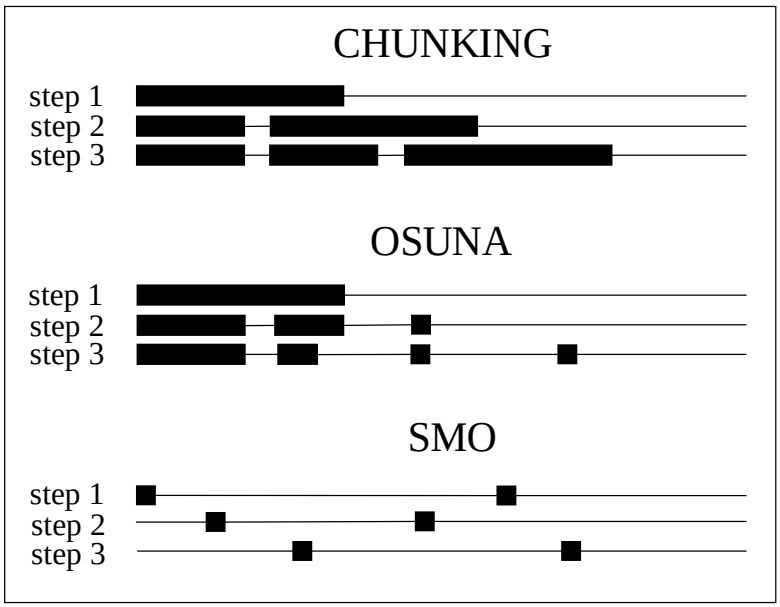



**Figure 3-6** Comparative of three steps of the Chunking, Osuna and SMO training algorithms (extracted from [Platt,1999-A]). The horizontal solid line represents the training set, while the black boxes correspond to the Lagrange multipliers being optimised.

#### 3.4.4 Multi-class classification in SVMs

As we have already presented, an SVM is a powerful discriminative binary classifier. However, this classifier can be modified to handle an N-class (N > 2) classification problem. In this case, the aim of the classifier is to assign a specific label, *l*, drawn from a finite set, *L*, of several elements from the test data. In other words, the multi-class classification problem can be defined as follows:

- Given a training set, $T = \{(x_i, y_i)\}_{i=1}^{l}$, where $x_i$ denotes training vectors and $y_i$ denotes corresponding class labels ($\{1,...,n\}$) assigned to the feature vector, find a classifier with a decision function, $f(x)$, such that $y=f(x)$, where $y \in \{1,...,n\}$ is the class label for $x$.

Over the last years, different solutions have been proposed to solve this problem.

When dealing with speaker recognition problems, the most used approach involves the decomposition of the multi-class problem into multiple binary classification problems and in the application of a decision strategy to decide on the class of the input pattern. Two-different strategies prevail, One-Against-One and One-Against-All, although the last one seems to be the preferred one [Karam,2007], [Campbell,2006], [Fauve,2007].

- One-Against-One strategy:

  In this approach, also known as one-versus-one or pairwise SVM [Kressel,1999] a binary classifier for each possible pair of classes is build. In other words, for a pair of classes, ($c_1$, $c_2$), an SVM is trained on the data from these two classes to discriminate between them. The training data must be relabelled so that labels +1 and -1 are assigned to the training data instead of labels $\{1,...,n\}$:

$$y_i = \begin{cases} +1 \; if \; c_i = c_1 \\ -1 \; if \; c_i = c_2 \end{cases} \qquad \textbf{Eq. (3-122)}$$

In this approach, for an N-class problem, the total number of binary SVM classifiers that must be trained is $N * (N-1)/2$. During the test phase different strategies have been applied. The simplest and most commonly used strategy is the max-wins voting [Friedman,1996]. The class assigned to a test pattern, x, is determined by testing the pattern into all the pairwise SVMs, and deciding on the class which has been selected more times in the binary comparisons. More specifically, given $f_{c_i c_j}$, the decision function for the pairwise SVM for the pair of classes ($c_i$,$c_j$), is defined as:

$$f_{c_i c_j}(x) = \begin{cases} 1 & if \; x \; is \; classified \; as \; belonging \; to \; c_i \\ 0 & otherwise \end{cases} \qquad \textbf{Eq. (3-123)}$$

The class label assigned to the test pattern can be expressed as:

$$f(x) = \max_i \sum_{j \neq i} f_{c_i c_j}(x) \qquad \textbf{Eq. (3-124)}$$

In case two classes reach the same number of votes, an additional strategy must be defined to decide between them. An alternative to the voting-for approach is the vote-against principle presented by [Cutzu,2003], which allows the classification of an input in an unknown class, thus avoiding false acceptance errors. In this approach, if a pair wise classifier, $f_{c_i c_j}$, scores a test pattern as belonging to one of the classes, for instance $c_i$, then they conclude $x \notin c_j$. Therefore, the voting scheme provides a vote against $c_j$.

More advanced methods include the use of decision graphs to determine the class that must be assigned to a test pattern. For instance, [Platt,1999-B] proposed the use of Directed Acyclic Graphs (DAG) for multiclass classification (see Figure 3-7).

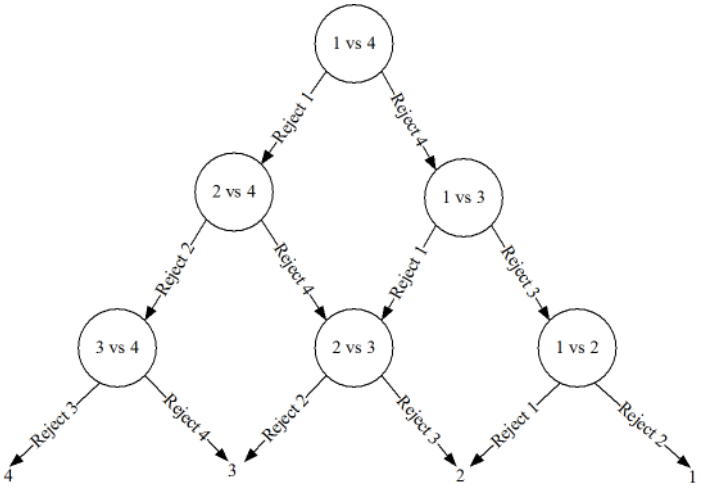



**Figure 3-7** DAG for a 4-class classification problem, where each node represents a binary classifier. (Extracted from [Platt,1999-B]).

In this approach, the class assigned to a test pattern is obtained using a rooted binary directed acyclic graph, where each node represents a binary classifier of

two different classes. In this case, only N-1 comparisons are required to provide the test pattern with a classification label. The test phase starts at root node where a decision is made between the two classes, rejecting one of them. Depending on the selected class, the process moves towards right or left node to a new node, where a new partial classification is performed. The process continues until a specific leaf, representing a specific class, is reached.

- One-Against-All strategy:

  In this approach, also known as one-versus-rest, N binary classifiers will be trained [Vapnik,1998]. However, each binary classifier is trained so that it can discriminate between a specific class and the remaining class (N-1). That is to say, for each class n, the SVM is trained on the whole training data set, where n-class data will be labelled as +1 whereas all the other training data (belonging to other classes) will be labelled as -1. In the test phase, a test pattern x, can be classified according to the winner-takes-all strategy. In other words, the label assigned to a test pattern will correspond to the class which obtains the highest value for the decision function (regardless of sign). Given the following decision function for class n,

$$f_n(x) = \sum_{i=1}^{m} y_i \alpha_i K(x, x_i) + b \qquad \textbf{Eq. (3-125)}$$

  The test pattern is classified according to the following expression:

$$f(x) = \underset{k}{\operatorname{argmax}} f_k(x) \qquad \textbf{Eq. (3-126)}$$

  One important difference between the One-Against-One strategy and the One-Against-All strategy resides in how efficiently a new class can be added to the classifier. This fact is important, for instance, when adding a new user to a speaker recognition system. In the first case, N new binary classifiers must be re-trained, while in the later all the binary classifiers already present in the system must be re-trained to include the new class in them, and additionally each of them must be trained on the whole training data.

### 3.4.5 Available software tools

In the case of support vector machines, there is a web-resource (URL: kernel-machines) devoted to learning methods among which are SVM. On this web site, we can find not only tutorials and publications related to SVM but also a list of SVM-related software tools. Some of the most commonly used or cited in the literature are the following.

SVM$^{light}$ [Joachims,1998] is an implementation of Vapnik's SVMs in C. The software is available for free for scientific use and can be found in the author's website (URL: SVMLight). This software tool is suitable for classification, regression and ranking problems.

Libsvm is open-source integrated software for support vector classification, regression and distribution estimation, supporting also multi-class classification. The software as well as the sources in java and C++ are available for direct download at the author's website (URL: LIBSVM)

It is worth citing the LNKnet software package, as it has recently added SVM and naïve Bayesian Classifiers. This modular software package integrates different neural networks, statistical and machine learning classification, clustering, and feature selection algorithms. The software package as well as the source code for individual programs can be downloaded from the website (URL: LNKnet). What is also interesting about this software package is that it integrates an easy-to-use GUI for running experiments.

Torch (URL: TORCH) software represents state-of-the-art machine learning algorithms. Developed in C++, and distributed under BSD license, it provides support for neural networks, support vector machines (in classification and regression), hidden Markov models, Gaussian mixture models, etc. The new version, Torch5 (URL: TORCH5), provides also a Matlab-like environment.

Again, MATLAB (URL: MathWorks) also provides a set of functions for statistical learning in the Bioinformatics toolbox. This set of functions allows the train of an SVM classifier, to classify data using an SVM or even evaluate the performance of the classifier among others. Different kernel functions are supported for training: Linear kernel, Quadratic kernel, Gaussian Radial Basis Function kernel with a default scaling factor (sigma) of 1, Polynomial kernel with a default order of 3 or Multilayer Perceptron kernel with default scale and bias parameters of [1, -1].

## 3.5 NEW TRENDS IN CLASSIFICATION METHODS

### 3.5.1 Supervector methods

In section 3.4.2, we have briefly introduced sequence kernels, also referred as dynamic kernels, as an alternative to classical kernels in SVM. This concept is actually closely related with a new trend in speaker recognition known as supervectors. The aim of supervectors is to provide a robust representation of utterances through the use of just a single vector in a high-dimensional space, allowing sequences of different durations to be compared and classified directly using traditional machine learning approaches, such as SVMs. Moreover, as pointed out by [Kinnunen,2010], a conventional adapted Gaussian mixture speaker model can be also regarded as a supervector. A more formal definition of supervector is also formulated as: any high- and fixed-dimensional representation of an utterance.

Speech utterances are usually parameterised as a sequence of observations $O=\{o_1,\ldots,o_N\}$, where different utterances result in different length observation sequences. In this context, dynamic kernels can be characterised as:

$$K(O_i, O_j) = \langle \phi(O_i, \lambda), \phi(O_j, \lambda) \rangle \qquad \textbf{Eq. (3-127)}$$

Where $\phi(O, \lambda)$ is a function that maps a speech utterance into a fixed- and usually high-dimensional space, and the kernel defines a distance between two different points (utterances) in the SVM feature space.

With this kernel, a speaker-decision boundary (i.e. a speaker model) can be trained given a data set of speech utterances $\{O_{tgt_1}, O_{tgt_2}, \ldots, O_{tgt_N}, O_{Nontgt_1}, O_{Nontgt_2}, \ldots, O_{Nontgt_M}\}$, where $O_{tgt}$ represents target speaker utterances and $O_{Nontgt}$ represents alternative speakers. Given this speaker model, $S$, a test utterance $O_T$ can be scored using:

$$S(O_T,S)=\sum_{i=1}^{N}\alpha_i y_i K(O_i,O_T)+b \rightarrow \begin{cases} < threshold \rightarrow impostor\ (reject) \\ \geq threshold \rightarrow target\ speaker\ (Accept) \end{cases} \quad \textbf{Eq. (3-128)}$$

Therefore, the issues to be addressed when using supervectors in the context of sequence kernel SVMs applied to speaker recognition can be summarised in Figure 3-8.

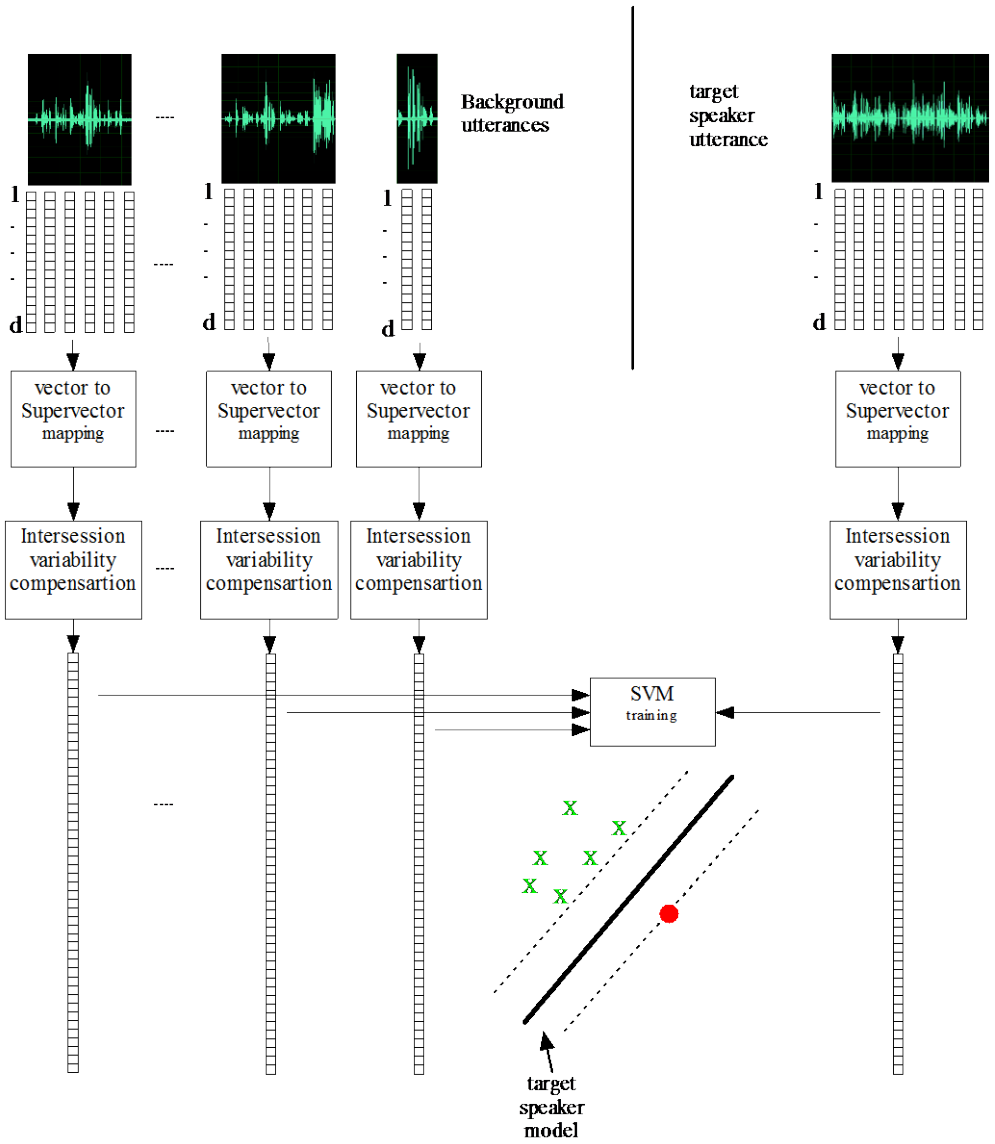



**Figure 3-8** Block diagram description showing the main concepts involved in the sequence kernel SVM modelling approach.

Depending on the function $\phi(O,\lambda)$, the dynamic kernels can be characterised into two different classes: parametric kernels and derivative kernels [Longworth,2009]. In the first case, an utterance will be used to train a generative model, whose parameters form a feature vector which is the representation of the utterance in a fixed- and high-dimensional representation space. Different approaches have been used to build the generative model on which this approach depends, for instance, [Stolcke,2005] introduced the MLLR kernel, [Ferras,2007] proposed the CMLLR kernel, [Campbell,2006-B], [Campbell,2006] presented the GMM-supervector kernel whereas [Yang,2007] applied the CAT kernel. In the case of derivative kernels, the partial

derivatives of the utterance log-likelihood respect to individual model parameters are used instead of the model parameters. [Wan,2005] proposed the use of score-space kernels, which are a generalisation of the Fisher kernels [Jaakkola,1998], for speaker verification.

- **GMM-supervector (GSV) kernel:** As previously said, the GMM-UBM has become the *de facto* reference in speaker recognition systems. The existence of the background model provides a straightforward way to map a new utterance into a high-dimensional vector. According to theory a new GMM speaker model can be trained, given a new speaker utterance, by MAP adaptation of the means of the UBM, $\lambda_{UBM} = \{w_i; \mu_i; \Sigma_i\}_{i=1}^{N}$. Thus the GMM supervector can be obtained by stacking the means of the adapted mixture components into a single vector.

  In order to apply an SVM classifier, the next step consists in defining an SVM sequence kernel that allows comparing two speech utterances, *a* and *b*, represented by their corresponding supervectors: $\lambda_a = \{w_i, \mu_i^a, \Sigma_i\}_{i=1}^{N} \rightarrow SV_a = \{\mu_i^a\}_{i=1}^{N}$ and $\lambda_b = \{w_i, \mu_i^b, \Sigma_i\}_{i=1}^{N} \rightarrow SV_b = \{\mu_i^b\}_{i=1}^{N}$. This sequential kernel is referred as *GMM SuperVector linear kernel* (also known as GSV kernel). The main idea consists in bounding the Kullback-Leibler (KL) divergence using the log-sum inequality between the two utterances. Assuming diagonal covariances, the kernel function defining the inner product can be expressed as [Campbell,2006-B]:

$$K(a,b) = \sum_{i=1}^{N} w_i (\mu_i^a)^t \Sigma_i^{-1} \mu_i^b = \sum_{i=1}^{N} \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^a\right)^t \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^b\right) \quad \textbf{Eq. (3-129)}$$

  Following this idea, an alternative method to find distances between GMM supervectors is proposed in [Campbell,2006-A]. In this case, using the standard inner product in function spaces and considering that means from different mixture components are far apart, we can derive the following kernel:

$$K(a,b) = \sum_{i=1}^{N} w_i^2 \mathcal{N}\left(\mu_i^a - \mu_i^b; 0, 2\Sigma_i\right) = \sum_{i=1}^{N} w_i^2 \frac{1}{(2\pi)^{D/2}\sqrt{2\Sigma_i}} e^{-\frac{1}{2}\left(\mu_i^a-\mu_i^b\right)^t (2\Sigma_i)^{-1}\left(\mu_i^a-\mu_i^b\right)} \quad \textbf{Eq. (3-130)}$$

  Working in the same direction of using KL divergence between GMMs, [Dehak,2006] proposed a probabilistic distance kernel, but in this case applying an exponential version of the distance. The derived kernel aims at finding the similarity between probability densities:

$$K\left(P(x|\lambda_a), P(x|\lambda_b)\right) = e^{-D^2\left(P(x|\lambda_a), P(x|\lambda_b)\right)} = e^{-\sum_{i=1}^{N}\sum_{d=1}^{D} w_i \frac{\left(\mu_{i,d}^a - \mu_{i,d}^b\right)^2}{\Sigma_{i,d}^2}} \quad \textbf{Eq. (3-131)}$$

where *D()*, corresponds to the Euclidean distance between two GMM models in the model space. Additionally, they performed a normalisation of all the speaker adapted GMM supervectors to have a constant distance from the UBM.

- **MLLR kernels:** As previously described in section 3.3.2 Maximum Likelihood Linear Regression (MLLR) has been used in speaker recognition as an alternative to MAP adaptation of UBM. In the MLLR approach, a single affine transform is applied to the mean vectors of a speaker-independent model to obtain the adapted speaker-dependent model means:

$$\mu_i^{speaker} = A\mu_i^{UBM} + b \qquad \textbf{Eq. (3-132)}$$

where *A* and *b* are the transform parameters which can be estimated so as to maximise the likelihood of the training data. These MLLR transformation parameters can be used as inputs to an SVM, as proposed in [Karam,2007] or [Stolcke,2007-A]. Although the MLLR adaptation method is used in both cases, the main difference lies in the universal model used (GMMs versus HMMs respectively). In [Stolcke,2007-A], a set of different transforms corresponding to context-dependent phone (triphones) classes, represented by their corresponding HMMs, are estimated for each speaker. In the case of [Karam,2007], MLLR adaptation is applied to adapt the means of an UBM represented by $\lambda_{UBM} = \{w_i; \mu_i; \Sigma_i\}_{i=1}^{N}$, given a training utterance *a*. As a result a MLLR transform vector, $\tau_a$, is obtained, by stacking the transposed rows of parameter *A* separated by the corresponding entries *b*, which is the representation of the input utterance in a high-dimensional space. Additionally, a nonlinear kernel in MLLR transform-vector space is introduced. This kernel is characterised as:

$$K(a,b) = \sum_{i=1}^{N} \left(\sqrt{\Delta_i}(A_a\mu_i + b_a)\right)^t \left(\sqrt{\Delta_i}(A_b\mu_i + b_b)\right) = \tau_a^t Q \tau_b \qquad \textbf{Eq. (3-133)}$$

where, $\Delta_i = w_i\Sigma_i^{-1}$, and *Q* is a block diagonal matrix consisting of *M* blocks $B_k$ of size (*M*+1)x(*M*+1),

$$B_k = \begin{pmatrix} R_k & r_k \\ r_k & \delta_k \end{pmatrix} \qquad \textbf{Eq. (3-134)}$$

where *M* is the number of rows in *A* and:

$$r_k = \sum_{i=1}^{N} \Delta_{ik}\bar{\mu}_i$$

$$\delta_k = \sum_{i=1}^{N} \Delta_{ik} \qquad \textbf{Eq. (3-135)}$$

$$R_k = \sum_{i=1}^{N} \Delta_{ik}\mu_i\mu_i^t$$

- **CMLLR kernel:** A variation of the MLLR kernels can be found in [Ferras,2007]. In this case, Constrained MLLR (CMLLR) is applied to adapt the

GMM/UBM to a specific speaker. CMLLR forces the transformation parameters estimated for means and variances to be the same, i.e.:

$$\mu_i^{speaker} = A\mu_i^{UBM} + b$$
$$\Sigma_i^{speaker} = A\Sigma_i A^t \qquad \textbf{Eq. (3-136)}$$

Additionally, the transform estimated via CMLLR can also be applied at feature level:

$$\hat{o}_t = A^{-1}o_t + A^{-1}b \qquad \textbf{Eq. (3-137)}$$

where $o_t$ is the feature vector observed at instant $t$.

After the CMLLR transforms are estimated for each speaker, and rearranged as vectors, Principal Component Analysis [Schölkopf,1997] is applied to reduce dimensionality and a liner kernel is lately applied to train the SVM.

- **CAT Kernel:** The CAT (Cluster Adaptive Training) kernel, proposed in [Yang,2007] actually makes use of a variation of the kernel in [Campbell,2006-B], but with a different set of features. In this approach, each speaker is represented by a feature vector which results from stacking the cluster weights extracted during the cluster adaptive training process. In order to characterise a speaker, $\lambda_s = \{\lambda_{s1}^g \ldots \lambda_{sC}^g\}_{g=1}^G$, Gaussian components are classified into multiple groups ([Yang,2007] reported results taking $G$=1, 4, 8, 12), and a set of separate set of cluster weights ([Yang,2007] reported results taking $C$=100, 200, 300, 400, 500}) are calculated for each group. The cluster weights are estimated through the EM algorithm.

  In this case, assuming that two utterances $a$ and $b$ are represented by their respective CAT weights $\{\vec{\lambda}_a^g\}_{g=1}^G$ and $\{\vec{\lambda}_b^g\}_{g=1}^G$, they define the kernel:

$$K(a,b) = \sum_{g=1}^{G}\left(\sqrt{w_g}\mathrm{C}_g^{-\frac{1}{2}}\vec{\lambda}_a^g\right)^t\left(\sqrt{w_g}\mathrm{C}_g^{-\frac{1}{2}}\vec{\lambda}_b^g\right) \qquad \textbf{Eq. (3-138)}$$

  where $w_g$ is the weight for the $g^{\text{th}}$ group (equal for all the groups) and $C_g$ is the covariance of the weight vectors in the $g^{\text{th}}$ group.

- **GLDS kernel**: The Generalised Linear Discriminant Sequence (GLDS), proposed in [Campbell,2006], uses an explicit mapping of a speaker utterance, $x$, into the kernel feature space using a monomial expansion (actually monomials up to degree 3 are used), $\boldsymbol{b}(x)$ defined as:

$$\boldsymbol{b}(x) = [b_1(x)\ b_2(x) \ldots\ b_k(x)]^t \qquad \textbf{Eq. (3-139)}$$

  where $b_i(x)$: $\mathbb{R}^D \rightarrow \mathbb{R}$ with $D$ being the dimension of $x$, and typically $b_1(x) = 1$. During training, each utterance, $x=\{x_1 \ldots x_{Nz}\}$, (either from target or from background speakers) is represented by the average expanded feature vector, and variance-normalised using just the background utterances:

$$x \to \bar{b}_x = VN\left(\frac{1}{Nz}\sum_{i=1}^{Nz} b(x_i)\right) \qquad \textbf{Eq. (3-140)}$$

This set of supervectors, $\{\bar{b}_x\}$, with their corresponding classification labels $l_x \in \{+1,-1\}$, is used to train an SVM using a standard linear kernel, leading to a speaker model that can be expressed as:

$$w_{speaker} = \sum_{i=1}^{sv} \alpha_i l_i \bar{b}_i + \boldsymbol{d} \qquad \textbf{Eq. (3-141)}$$

where *sv* is the number of support vectors. This characterisation of a speaker, allows for a rapid scoring of a given input utterance, $\boldsymbol{y}=\{y_1,\ldots,y_n\}$, by evaluating the normalised average mapping (using Eq. (3-140)) and computing:

$$score = w_{speaker}\bar{b}_y \qquad \textbf{Eq. (3-142)}$$

- **Score-space kernels**: Derivative kernels have also been applied to speaker verification tasks. In this regard, [Wan,2005] propose the use of score-space kernels, which are based on generative models (such as GMMs) and constitutes a generalisation of the Fisher kernels [Jaakkola,1998]. In this approach, instead of using the model parameters as the extended feature-space, the partial derivatives of the utterance log-likelihood with respect to individual model parameters (weights, means and covariances of the GMM) are used. The mapping of a particular utterance, *x*, given a specific GMM, $\lambda_s = \{w_i; \mu_i; \Sigma_i\}_{i=1}^{N}$, is thus characterised as:

$$\psi_F(x) = F(\log P(x|\lambda_s)) \qquad \textbf{Eq. (3-143)}$$

where $\psi_F(x)$ is the score-vector and $F$ is typically one of the mapping functions in Table 3-1.

[Wan,2005] report results using both the Fisher kernel with and without incorporating the log-likelihood ratio in the feature expansion and the Likelihood-ratio kernel, which instead of applying the mapping function to $\log P(x|\lambda_s)$ on a specific speaker model, use the following expression as input:

$$\frac{\log P(x|\lambda_s)}{\log P(x|\lambda_{UBM})} \qquad \textbf{Eq. (3-144)}$$

| Mapping function | Expression |
|---|---|
| *First derivative* | $F=\nabla_\lambda$ |
| *First derivative and log-likelihood ratio* | $F=[\nabla_\lambda,1]^T$ |
| *First and second derivative* | $F=[\nabla_\lambda,\ \nabla^2_\lambda]^T$ |

**Table 3-1** Some examples of mapping functions

However, the features produced in the different approaches (either derivative or parametric) may have different properties and to some extent be complementary. For this reason, different works proposed the fusion of these methods, such as [Longworth,2009] where parametric and derivative approaches were fused at kernel

level and at score level or in [Li,2009] where seven classifiers (among which we can cite GMM–UBM, GLDS-SVM and GSV-SVM) were fused at the score level. [Sturim,2009] also report improvements in the recognition rates by combining SVM classifiers based upon GMM-supervector kernels and MLLR kernels.

One important issue when dealing with supervectors is the high dimensionality of the representation of the utterances in the new feature space. To deal with this problem a dimensionality reduction technique can be applied as early noted in CMLLR kernel section. One of the most used techniques which has also been used in some of the experiments conducted in this thesis is Principal Component Analysis (PCA) which is briefly presented below:

- **PCA**

  Since its introduction in 1993 [Hotelling,1933], PCA, which is one of the most powerful tools for high dimensional multivariate analysis, has been widely used in pattern recognition especially in face recognition [Sharkas,2008], as well as in speaker recognition [Zhao,2009], [Zhang,2003]. PCA is not only a mathematical method to reduce the dimensionality of the data without losing too much information (i.e., retaining most of the variations present in the original data), but a way to identify patterns in the data as it extracts orthogonal principal components with largest magnitudes, thus highlighting the similarities and differences. A deep review can be found in [Jolliffe,2002].

  The objective of this method is to find a projection $W_{PCA}$, such that when applied to the original data, R, it is projected to a new lower dimensional subspace, r, describing most of the variability in the training matrix.

  $$r=W_{PCA}xR \qquad \textbf{Eq. (3-145)}$$

  Considering that we have a training data set composed by $\{u_1,u_2,\ldots,u_N\}$ utterances, represented in an F-dimensional space, from $N$ different speakers, in order to obtain the transformation matrix, $W_{PCA}$, the following steps must be followed:

  - Normalise original data subtracting the mean from each of the data dimensions. After this process we get mean centred training data $\{w_1,w_2,\ldots,w_N\}$ which is used to build the matrix $W$.
  - Estimate the square and symmetric covariance matrix $C=W^TW$ of order $N$x$N$.
  - Perform eigenvalue decomposition of covariance matrix to get the eigenvectors $e_i$ and corresponding eigenvalues, $\lambda_i$.
  - Build matrix $E$ by re-arranging eigenvectors $e_i$, in such a way that eigenvectors associated with largest eigenvalues appear first in the matrix. In other words, eigenvectors are ordered by eigenvalue, from highest to lowest, representing the components by significance.
  - In order to reduce the dimensionality of the data, a selection of the $M$ first eigenvectors, such that $M<N$, with the largest eigenvalues is done, obtaining $W_{PCA}$, with the selected eigenvectors being the columns of $W_{PCA}$.

It must be noted that in order to work properly, the normalisation step performed on the training data must be carried out also on the test data.

One of the main advantages of this method is not only that it reduces the dimensionality of the data, but as the process of obtaining the matrix transformation only requires matrix multiplications, the complexity and time consumed by the pattern recognition system can be considerably reduced.

Another important issue when dealing with supervectors, which also appears in the case of GMM-based speaker recognition systems, is how to address inter-session variability. In this approach, each utterance is mapped to a single point in a higher dimensional space, so differences due channel mismatch, handset variability, etc. in utterances recorded from the same speakers may cause performance degradation unless session compensation methods are applied. Two of the session compensation methods widely applied in the GMM and SVM speaker recognition systems are, Joint Factor Analysis (JFA) [Kenny,2005-A], [Kenny,2008], [Vogt,2008] and Nuisance Attribute Projection (NAP) [Solomonoff,2004], respectively. [Fauve,2007] present a comparative interpretation of both approaches applied on the same dataset.

- **JFA**

  Joint factor analysis, introduced into speaker verification by [Kenny,2005-C], [Vogt,2008], aims at modelling speaker and session variability in generative-based models (for instance, GMMs) through the use of traditional statistical methods. In this approach, the way to deal with channel and speaker variability is to assume that a given speaker can be characterised by a speaker- and channel-dependent supervectors as:

  $$M = s + c \qquad \textbf{Eq. (3-146)}$$

  where $s$ represents the speaker-dependent component/supervector and $c$ is the session-dependent component/supervector, both of them statistically independent and assumed to have a standard normal distribution.

  Without taking into account the intersession variability, the speaker dependent supervector can be adapted from a speaker-independent model combining the priors for classical MAP and eigenvoice MAP approach. In this way $s$ can be decomposed in:

  $$s = m + vy + dz \qquad \textbf{Eq. (3-147)}$$

  where $m$ represents the speaker- and session-independent component. The supervector defined by a $C$ component UBM trained from $F$-dimensional acoustic feature vectors may serve as an estimate of $m$. Thus $m$ is a $CF$x1 matrix. $v$ is a rectangular low rank matrix ($CF$x$R_s$, where $R_s$ is the number of speaker factors) whose columns are also known as eigenvoices. $y$ is an independent random vector having standard normal distribution, whose components are known as speaker factors. $d$ is a $CF$x$CF$ residual diagonal matrix. Finally, $z$ is $CF$-dimensional independent random vectors having standard distribution. Taking into account these assumptions, $s$ is normally distributed with mean $m$ and covariance matrix $vv^t+d^2$.

In order to deal with channel-adaptation, it is assumed that the speaker and channel-dependent supervectors for different recordings of a given speaker have a Gaussian distribution centred on the speaker's supervector. In other words, applying eigenchannel MAP the channel factors and the corresponding eigenchannels can be estimated so that *c* can be characterised as:

$$c = ux \qquad \textbf{Eq. (3-148)}$$

where *u* is a rectangular low rank matrix (*CF*x$R_C$, where $R_C$ is the number of channel factors) whose columns are also known as eigenchannels and the components of normally distributed random vectors, *x*, are known as channel factors.

Replacing Eq. (3-147) and Eq. (3-148) in Eq. (3-146), the speaker- and channel-dependent supervector can be expressed as:

$$M = m + vy + dz + ux \qquad \textbf{Eq. (3-149)}$$

The factor analysis model can be specified by the quintuple $\Lambda=\{m,v,d,u,\Sigma\}$, where an estimate of $\Sigma$ can be the UBM covariance matrices conveniently arranged to form a diagonal *CF*x*CF* matrix. Thus, the underlying task in FA is to estimate the hyperparameters *v*, *d* and *u* from a suitable database in which sufficient recordings from multiple speakers are available from multiple sessions. Additionally the speaker and session factors for each target speaker must be estimated from enrolment data. Although a deep review and practical implementation for hyperparameter estimation can be found in [Kenny,2008], [Matrouf,2007], [Yin,2007], a hint on the main steps is given below:

- 1st: Train the eigenvoice matrix, *v*, assuming that matrices *u* and *d* are zero.
- 2nd: Train the eigenchannel matrix *u*, given the estimate of *v*, and assuming that *d* is zero.
- 3rd: Train the residual matrix *d*, given the estimates of *v* and *u*.
- Using the matrices, the next step consists in computing the speaker-, channel- and residual-factors.
- The final score is computed using the matrices and factors. Different scoring approaches have been tested in this context, having been compared in [Glembek,2009].

Although the application of this method has been successfully tested in SR systems, it shows some performance degradation when there is a mismatch between the training and test utterance lengths as pointed out by [Kinnunen,2010].

- **NAP**

The aim of NAP is to remove e the dimensions which are related with inter-session or channel variability (i.e. nuisance effects) from the original expanded space, therefore irrelevant for speaker recognition. This is actually done by

projecting out a subspace using an appropriate projection matrix, *P*, which is equivalent to develop a modified kernel which projects out channel effects.

Although the interested reader may refer to [Solomonoff,2004], [Solomonoff,2005] for theory foundations of this compensation method, a description from a practical point of view of the method is presented in [Brummer,2007], [Fauve,2007]. As already said, the aim of the method is to find a transformation from the original expanded space into the NAP-subspace in which the unwanted variability is removed. For each utterance, $O_i$, on the training set (no matter whether it belongs to a target or background speaker), the following transformation is applied:

$$\begin{aligned}\hat{\phi}(O_i,\lambda) &= P\big(\phi(O_i,\lambda)\big) = (I - SS^t)\phi(O_i,\lambda) \\ &= \phi(O_i,\lambda) - S\big(S^t\phi(O_i,\lambda)\big)\end{aligned} \qquad \textbf{Eq. (3-150)}$$

where $\phi(O_i,\lambda)$, is the function that maps the utterance $O_i$ into its corresponding supervector of dimension *D*, $\hat{\phi}(O_i,\lambda)$ representing the NAP-transformed supervector, and S (also referred to as the eigenchannel adaptation matrix) defines the orthonormal NAP-subspace.

In order to estimate the eigenchannel adaptation matrix, we may proceed in the following manner. Assuming the existence of a background set of utterances from a number, *R*, of different speakers, recorded in different sessions, *H*, we define the expansion space data matrix as:

$$A = \left[\phi(O_1^1,\lambda) \quad \cdots \quad \phi\big(O_1^{H_1},\lambda\big)\cdots\phi(O_R^1,\lambda) \quad \cdots \quad \phi\big(O_R^{H_R},\lambda\big)\right] \qquad \textbf{Eq. (3-151)}$$

where $O_i^j$, represents the $j^{\text{th}}$ utterance/session for speaker *I*, and $N=H_1+\ldots+H_R$. For each speaker, a session average supervector, $\bar{\phi}(O_i,\lambda)$, is computed and then removed from each supervector:

$$\tilde{\phi}\big(O_i^j,\lambda\big) = \phi\big(O_i^j,\lambda\big) - \bar{\phi}(O_i,\lambda), \forall j \qquad \textbf{Eq. (3-152)}$$

leading to the average normalised expansion space data matrix:

$$\tilde{A} = \left[\tilde{\phi}(O_1^1,\lambda) \quad \cdots \quad \tilde{\phi}\big(O_1^{H_1},\lambda\big)\cdots\tilde{\phi}(O_R^1,\lambda) \quad \cdots \quad \tilde{\phi}\big(O_R^{H_R},\lambda\big)\right] \qquad \textbf{Eq. (3-153)}$$

In this process, matrix $\tilde{A}$, with dimension *D*x*N*, is supposed to retain the nuisance variability whereas most of the speaker variability has been removed. The next step is to empirically define the dimension, *K*, of the NAP-transform, which defines the subspace of high variability. Once the dimension is established, the matrix which defines a base to the subspace can be estimated solving an eigenvalue problem on the covariance matrix, $C=\tilde{A}\tilde{A}^t$. For this purpose, Principal Component Analysis (PCA) can be used to calculate the *K* eigenvectors with highest eigenvalues, when linear kernels are used. In the case of using nonlinear kernels in the SVM, kernel PCA [Solomonoff,2004] can be used instead. At the end of this process and after applying some normalisation steps, the matrix S (of size *D*x*K*) is obtained.

It is important to note that NAP also removes speaker-specific information, so alternative discriminative methods which aim to introduce between class scatter information in the NAP-transform, such as Scatter Difference Analysis [Vogt,2006] have been proposed to overcome this problem.

### 3.5.2 *i*-vectors

In the latest NIST SRE evaluation, many sites have used sub-systems based on total variability spaces, also known as *i*-vectors. This approach has its starting point in two methods already presented, namely JFA and SVMs. In [Dehak,2009-C] the use of speaker factors as input to an SVM classifier was proposed. Channel factors were originally neglected as they are supposed to provide information about channel effects. However, it was found [Dehak,2009-A] that channel factors estimated using JFA also contain information about speakers.

In this new scenario, [Dehak,2009-B] proposed the used of factor analysis as a feature extractor, which instead of independently model speaker and channel variability in a high dimension space of supervectors, defines a new low-dimensional space which aims at representing speaker and channel variability simultaneously. This new space, known as total variability space, is characterised by the total variability matrix that contains the eigenvectors corresponding to the largest eigenvalues of the total variability covariance matrix. As no distinction between the speaker effects and the channel effects are made in GMM supervector space, a new utterance can be defined by the supervector:

$$M = m + Tw \qquad \textbf{Eq. (3-154)}$$

Like in JFA (previously presented), $m$ represents the speaker- and session-independent component. The supervector defined by a $C$ component UBM trained from $F$-dimensional acoustic feature vectors may serve as an estimate of $m$. $T$, which is a low rank rectangular matrix, represents the total variability matrix. $w$ represents the coordinates of the utterance/speaker in the total variability space. The components of $w$ are known as total factors, while the whole vector (which has a standard normal distribution) is also known as the *i*-vector. Similar to JFA, $M$ is normally distributed with mean $m$ and covariance matrix $TT^t$.

The total factor vector $w$ for a given utterance $U=\{u_1,\ldots,u_K\}$, characterised by a set of $F$-dimensional vectors, can be regarded as a hidden variable whose posterior distribution can be determined using Baum-Welch statistics from the UBM characterised by $\lambda_{UBM}=\{w_c,\mu_c,\Sigma_c\}_{c=1\ldots C}$, where $C$ is the number of mixture components. Eq. (3-155) defines how the *i*-vector $w$, for the given utterance can be obtained:

$$w = \left(I + T^t\Sigma^{-1}N(u)T\right)^{-1}T^t\Sigma^{-1}\tilde{F}(u) \qquad \textbf{Eq. (3-155)}$$

where:

- $\Sigma$ is the $CF$x$CF$ diagonal covariance matrix that models the residual variability not captured by $T$,
- $N(u)$ is a $CF$x$CF$ diagonal matrix whose are $N_c(u)I$, being $N_c(u)$ is the sum over all feature vectors of $U$ of the posterior probability of generating each vector by the corresponding mixture $c$,

$$N_c(u) = \sum_{k=1}^{K} P(c|u_k, \lambda_{UBM}) \qquad \textbf{Eq. (3-156)}$$

- $\tilde{F}(u)$ is a supervector of dimension *CFx1* obtained by concatenating all the centralised first order Baum-Welch statistics $\tilde{F}_c(u)$, where:

$$\tilde{F}_c(u) = \sum_{k=1}^{K} P(c|u_k, \lambda_{UBM})\,(u_k - m_c) \qquad \textbf{Eq. (3-157)}$$

  and $m_c$ is the mean of the mixture component *c*.

- The matrix T can be trained following the same procedure as the one used to train the eigenvoice matrix *v* in the FA approach (proposition 3 in [Kenny,2005-B]). The main difference remains in the fact that instead of providing a supervector per speaker, a supervector per utterance is used instead. In other words, each utterance from a given speaker is supposed to be produced by different speakers.

The major problem to be faced during the estimation of T is the need for a large amount of application-specific training data. To overcome this problem [Senoussaoui,2010] successfully proposed the use of alternative context data as a complement to the available one. Another problem that may arise also related with the amount of training data relies in the fact that the process of training the matrix T can be very expensive both in terms of memory and speed. To overcome this problem [Glembek,2011] presented some simplifications that can be applied to the process without losing too much accuracy in terms of recognition rates. A complete description of the process to obtain the *i*-vector from a given utterance can be found in [Dehak,2010].

Besides the different systems submitted to NIST SRE 2010 having successfully used this approach, [Senoussaoui,2010], [Dehak,2009-A], [Dehak,2009-B], [Dehak,2009-C] have also presented interesting results. In these works, a fast scoring procedure based on the cosine kernel, without using the SVM approach, has been proposed. The cosine kernel distance between the *i*-vector of an enrolment utterance and the *i*-vector of a test segment is compared to a decision threshold, $\theta$, to accept or reject a given trial. As no target model is required, and *i*-vectors are smaller in size, the decision process is faster and less computationally demanding compared to other methods.

$$score(w_{target}, w_{test}) = \frac{w_{test}^t w_{target}}{\sqrt{w_{test}^t w_{test}}\sqrt{w_{target}^t w_{target}}} \begin{cases} \leq \theta \rightarrow reject \\ \geq \theta \rightarrow accept \end{cases} \qquad \textbf{Eq. (3-158)}$$

Like in the case of supervectors, standard compensation techniques can also be applied. Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalisation (WCCN) are the two compensation techniques which when applied to *i*-vectors to remove channel effects in the total variability space, provide better performance.

- **WCCN**

  Originally proposed in [Hatch,2006-B] and lately extended with Principal Component Analysis [Hatch,2006-A] to deal with the problem of large feature sets, it consist in training a Generalised Linear Kernel on the form

$$K(x_1, x_2) = x_1^t W^{-1} x_2 \qquad \textbf{Eq. (3-159)}$$

where $W$ is a symmetric positive semidefinite matrix that represents the expected within-class covariance matrix over all classes (i.e. speakers) in the training data. $W$ can be characterised as:

$$W \triangleq \sum_{i=1}^{M} p(i) C_i$$
$$C_i \triangleq \mathrm{E}(x_i - \bar{x}_i)(x_i - \bar{x}_i)^t; \ \forall i \qquad \textbf{Eq. (3-160)}$$

where $C_i$ and $p(i)$ represent the covariance matrix and the prior probability of a given speaker $i/i \in \{1 \ldots M\}$, respectively. Moreover, $x_i$ represents a feature vector from class $i$, and $\bar{x}_i$ represents the expected value of $x_i$.

The WCCN feature transformation applied on the original feature space can be defined as:

$$\Phi(x) \triangleq A^t x \qquad \textbf{Eq. (3-161)}$$

with

$$AA^t \triangleq W^{-1} \qquad \textbf{Eq. (3-162)}$$

Within-class covariance normalisation constitutes an alternative to NAP. Moreover, as noted in [Stolcke,2007-A], NAP can be regarded as a simplified version of WCCN, where the main difference relies in the weighting strategy of the dimensions in the high-dimensional space. While NAP completely removes some dimensions reducing the original expanded space to the NAP-subspace, WCCN performs a complex weighting of the eigenvectors which define the subspace. Additionally, practical application of WCCN to MLLR-SVM systems can be found in [Kajarekar,2007], [Hatch,2006-A].

- **LDA**

  Linear Discriminant Analysis is a technique widely used in the field of pattern recognition [Ichino,2006], [Muroi,2008], [Dehak,2010] both for better discrimination between different classes and also for dimensionality reduction. More specifically, the main goal of LDA is to find a transformation matrix W such that when applied to the original feature space, the between-class variance is maximised while the intra-class variance is minimised.

  When applied to the case of speaker recognition, we can assume that each class, $C_i$, is made up of all recordings, $r \in C_i$, (represented for instance as $i$-vectors [Dehak,2010]) of a single speaker. Under this initial assumption, we can define the following between-class variance matrix, $S_b$, and the within-class variance matrix, $S_w$:

$$S_b = \sum_{i=1}^{S} (\mu_i - \mu)(\mu_i - \mu)^t \quad \textbf{Eq. (3-163)}$$

$$S_w = \sum_{i=1}^{S} \frac{1}{n_r} \sum_{r \epsilon C_i} (r - \mu_i)(r - \mu_i)^t \quad \textbf{Eq. (3-164)}$$

where, *S* is the number of speakers or classes, r is the set of recordings or elements belonging to a class $C_i$ / $i$=1…k, $\mu_i$ is the mean or average vector for class *i*, and $\mu$ is the speaker population mean or more generally the overall mean vector.

Based on these definitions, the goal of LDA is to maximise the Rayleigh coefficient for space direction *W*, which is defined as:

$$J(W) = \frac{|W^t S_B W|}{|W^t S_w W|} \quad \textbf{Eq. (3-165)}$$

The LDA transformation which results from maximizing this ratio can be regarded as the set of eigenvectors, *W*, corresponding to the eigenvalue decomposition in:

$$S_B W = \lambda S_w W \quad \textbf{Eq. (3-166)}$$

The dimension of this matrix can be reduced if only the best eigenvectors (i.e. the ones with highest eigenvalues) are selected. The number of best eigenvectors is optimised for the specific recognition system.

According to [Sharkas,2008], a problem may occur if the matrix *Sw* becomes singular. In this case, a solution could be the use of an intermediate representation space of the data involved. For instance, Principal Component Analysis (PCA) can be used to project the original space into PCA space and then apply LDA.

We have presented two of the most important modelling trends which aim at improving speaker recognition system's performance: supervectors and *i*-vectors. Recent studies have focused their effort on comparing the performance of these approaches. For instance, [Dehak,2008] present a comparison between JFA and GMM-SVM with both linear and non-linear kernels, applying TNorm and ZTNorm score normalisation. Results show similar performance between SVM with Gaussian supervectors and JFA without speaker factors, while the inclusion of speaker factors makes a difference in favour of JFA. Additionally, non-linear kernels outperform linear kernels, and the fusion of JFA and GMM-SVM using logistic regression improves recognition rates. In [Kajarekar,2007] two different intersession variability compensation techniques, WCNN and NAP were compared using an MLLR-SVM speaker verification system, reporting no significant performance differences. Additionally in [Dehak,2010] three channel compensation techniques already presented (NAP, WCCN and LDA) have been applied in a system where each speaker is represented in a total variability space (i.e. using *i*-vectors). In this case, the best results were achieved for the case in which LDA is followed by WCCN.

## 3.6 FUSION

So far we have presented different classification methods widely applied in speaker recognition. However, after a deep review of the different systems submitted to different NIST SRE, it can be assured that these classification methods are rarely used alone for recognition purposes. Like in other pattern classification problems, fusion techniques have been extensively used in speaker recognition.

In Chapter 1 we have already presented different proposals for fussing generative and discriminative methods to improve speaker recognition system performance. In this section we will focus on techniques typically used for fusion at the matching score.

The idea behind fusion at the matching score, relies on the fact that each speaker is modelled using multiple speaker models. Two different approaches have been studied:

- A set of features is used to create a number of different models using different classifiers.
- Different feature sets are first extracted and then a specific classifier is used for each feature set.

In both cases, each classifier provides a score or decision that may be combined in order to improve recognition rates. The simplest way to combine these scores is by weighted sum:

$$score_f = \sum_{n=1}^{N} w_n s_n \qquad \textbf{Eq. (3-167)}$$

where $N$ is the number of classifiers, $s_n$ represents the score provided by the $n$th classifier and, $w_n$ represents the weight assigned to the $n$th classifier. There are different ways to determine the specific weights for each of the classifiers involved in the system. The simplest way is to assign the same weight to all classifiers:

$$w_n = {}^{1}/_{N}, \forall n \in [1 \dots N] \qquad \textbf{Eq. (3-168)}$$

Another approach consists in using a development set in order to optimise specific weights for each classifier. A method that has become a *de facto* standard in speaker recognition systems is the one based on logistic regression proposed in [Brummer,2006], practically presented in [van Leeuwen,2007] and applied in [Brummer,2007].

In order to achieve the best set of weights for the fusion stage, it is necessary to define a quality measure that establishes the performance of a system. In [Brummer,2006] a measure of discrimination and calibration suitable for evaluating soft (application-independent) recognition decisions in log-likelihood-ratio form is defined as:

$$C_{llr}(\Gamma) = \frac{1}{\log 2}\left(\frac{P_{tar}}{\|X_{tar}\|}\sum_{x\in X_{tar}}\log\left(1+e^{-\Gamma(x)}\right) + \frac{1-P_{tar}}{\|X_{nontar}\|}\sum_{x\in X_{nontar}}\log\left(1+e^{-\Gamma(x)}\right)\right) \qquad \textbf{Eq. (3-169)}$$

where Γ identifies the system under evaluation, Γ(x) is the attempt of the system under evaluation to evaluate the log-likelihood-ratio for trial *x*, $X_{tar}$ is the set of target trials, $X_{nontar}$ is the set of non-target trials and $P_{tar}$ is the target prior. The normalised sums represent the expectations of "*log costs*" for target trials and non-target trials respectively.

This application-independent evaluation objective, $C_{llr}$, allows different interpretations. It can be interpreted as a total error-rate across a wide range of applications, or as an expected cost; it also has an information-theoretic interpretation in terms of Shannon's entropy and is a normalised (negative) discriminative log-likelihood. Regarding this last interpretation, $C_{llr}$, can be interpreted as the objective function to be minimised using linear logistic regression. A training procedure that optimises a recognition system with $C_{llr}$ as objective is actually performing maximum-likelihood training.

In a practical scenario, [van Leeuwen,2007], [Brummer,2007], we can define a linear fusion of a set of *N* speaker recognition systems, for a specific trial *x* as:

$$score_f(x) = w_0 + \sum_{n=1}^{N} w_n s_n(x) \qquad \textbf{Eq. (3-170)}$$

where $s_n(x)$ represents the score provided by the *n*th system for trial *x* and $w_0$ is a constant included to improve calibration of the fused score.

In order to train the fusion weights, a set of training scores (which must be similar to the ones presented during the recognition phase) both for target and non-target trials, represented by $X_{tar}$ and $X_{nontar}$ respectively, are needed. Additionally, $C_{llr}$ previously defined in terms of these scores and the fusion weights, is set as the optimisation objective. As far as $C_{llr}$ is a convex function of $w=[w_0,\ldots,w_N]$, at a fixed value of $P_{tar}$, and therefore has a unique global minimum, logistic regression can be used to find this minimum. In [Brummer,2007] a conjugate gradient ascent method is applied. Once fusion weights are trained, they can be used during the recognition phase.

A more complex technique consists in considering the outputs from the different classifiers, arranged as a vector, as another random variable, and train a backend classifier for this new data. For instance, in [Tong,2006] the average value of three different SVM systems is used to provide the final score of a speaker recognition system. Other approaches, like [Reynolds,2005] prefer the use of a perceptron classifier as fusion strategy.

A new trend in fusion strategies is the use of quality measures, also known as auxiliary side information, during the recognition phase. The idea behind this trend is to carry out a test-dependent fusion; in other words, adjust the fusion system for each test case, according to some quality measure defined for it. In [Garcia-Romero,2006] a linear quality-dependent score combination of two SVM classifiers that rely on high-level and low-level information is used. Quality information, such as Signal-to-noise ratio (SNR) or F0 deviations are also incorporated during the training phase of SVM classifiers. In a similar way [Solewicz,2007] introduce side information related to channel, noise and prosody aspects from train and test utterances in an SVM post-classifier. However, in [Ferrer,2008] auxiliary information (for instance, non-nativeness of speakers) is used to adapt the weights of a linear logistic regression (LLR) combiner. The use of quality measures clearly improves recognition rates and is worth of further study.

### 3.6.1 Available software tools

Focal is a free open-source toolkit that provides a set of tools (written in Matlab code) for evaluation, fusion and calibration of statistical pattern recognisers. Although originally developed for any two-class recogniser, especially for the task of speaker recognition, FoCal toolkit currently integrates three different parts:

- FoCal Two-Class: Under this section we can find different tools for evaluating the goodness of speaker detection log-likelihood ratios, for system calibration (map the score of the system to a detection log-likelihood ratio) such as Z-Cal and S-Cal, and obviously tools for fusion of multiple detection scores. In this last case, two different algorithms for optimizing the fusion weights are provided: linear logistic regression fusion and linear minimum-MSE fusion.
- FoCal Bilinear: Constitutes a modification of the tools available in FoCal Two-Class in order to allow the incorporation of quality measures or side-information.
- FoCal MultiClass: In this section we can find fusion and calibration tools to work with multi-class statistical pattern recognition scores. Among these tools we can find not only discriminative logistic regression algorithms but generative Gaussian Backends, with Probabilistic PCA, Factor-Analysis and HLDA (Heteroscedastic Linear Discriminant Analysis) covariance regularisation as well.

FoCal toolkit can be downloaded from the author's website (URL: FoCal).

Later on, FoCal toolkit has evolved into BOSARIS toolkit (URL:BOSARIS), which again provides a logistic regression solution for fusing multiple systems, as well as improved and additional capabilities relating calibration. However, it does not provide support for multiclass statistical pattern recognition scores.

# 4 TEST BENCH

Progress in the field of biometric recognition, as in many other research fields, is closely related with test beds, in the form of databases or in the form of evaluations that provide a framework for testing and comparing the performance of different systems.

In Chapter 1, a comparison of different speaker recognition systems was presented. From that comparison it is clear that there is no common protocol and moreover, even in the case of using the same database, tests are not run or design with the same data. Thus recognition rates - and therefore systems - are not straightforward comparable. From this point of view, evaluations organised by third parties, such as the Speaker Recognition Evaluation (SRE) organised by the National Institute of Standard and Technologies (NIST), or the one organised by the Idiap Research Institute provide a common framework for assessing the performance of recognition systems.

The purpose of the present chapter is twofold. In the first place, the databases used in the different experiments carried out in the framework of this thesis are presented. Secondly the set of specifically design tests using the described databases will also be described.

## 4.1 DATABASES

Different databases have been used in the set of experiments carried out during the development of the present research:

- HESPERIA
- ALBAYZIN
- MOBIO
- NIST SRE databases.

The selection of these databases is based on the following arguments. First of all, the database HESPERIA has been recorded by the GIAPSI research group as part of the research project HESPERIA[1], according to quality standards that make it suitable for tuning the vocal tract and glottal separation algorithms, as well as for the analysis of voice quality. The database ALBAYZIN was used mainly due to the following characteristics: gender variability, age variability, Spanish phonetically balanced recordings, and lack of channel variability. The use of the MOBIO speech corpus constitutes and additional challenge as it contains text-independent recordings acquired in mobile environments. This means that the content of the database has been recorded using mobile devices under real and non-controlled scenarios with no restriction on the message. Additionally, the recognition results achieved can be used to participate in the speaker verification evaluation organised by the Idiap Research Institute. Finally, NIST databases deserve special mention. As we have already said, NIST has contributed to the advance of the automatic speaker verification area providing a common framework in form of SREs. In order to participate in these evaluations with certain guarantees, it is necessary to become familiar with the type of recordings that must be processed. The use of these databases pose a greater challenge than any of those presented so far. Specifically, its use involves dealing with high inter-session variability present in form

[1] Project HESPERIA (http.//www.proyecto-hesperia.org) from the Program CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

of multiple acoustic environments (both controlled and non-controlled), multiple communication channels, different emotional states of speakers (stress, sadness, etc.), no message restriction and even recordings acquired years apart from the same speaker. Language variability is also present until NIST SRE 2010, while NIST SRE2010 and NIST SRE 2012 are focused on English language although English is not always the native language of speakers.

Although a more detailed description of each speech corpora will be presented in the following section, Table 4-1 summarises some of the most basic characteristics shared by all them.

| Database | Language | female speakers | male speakers | Context |
|---|---|---|---|---|
| HESPERIA | Spanish | 93 | 109 | Text-dependent<br>Text-constrained<br>Text-independent |
| ALBAYZIN | Spanish | 152 | 152 | Text-independent |
| MOBIO | English | 52 | 100 | Text-independent |
| NIST SRE 04 | English/Spanish/Arabic/ Mandarin/Russian/Etc. | 184* | 125* | Text-independent |
| NIST SRE 06 | English/Spanish/Arabic/ Mandarin/Russian/Etc. | 666* | 479* | Text-independent |
| NIST SRE 08 | English/Spanish/Arabic/ Mandarin/Russian/Etc. | 801* | 472* | Text-independent |
| NIST SRE 10 | English | 260 | 239 | Text-independent |

**Table 4-1** Main characteristics of the different databases used in the experiments carried out in the course of this thesis (*: only English).

### 4.1.1 The Database HESPERIA

The Database HESPERIA (Homeland sEcurity: tecnologíaS Para la sEguridad integral en espacIos públicos e infraestructurAs) was recorded for speaker recognition purposes in high security applications (specifically in biometric access control applications). It consists of 202 speakers (109 male and 93 female) from different institutions, agencies or companies involved in the research project. All the recordings were captured in three sessions under controlled conditions but using different microphones:

- High quality Cardioid lavalier microphone (labelled as c1). Allows high-fidelity recordings with a frequency range of 30Hz to 20 kHz and a bandwidth less than 3 dB flat to the band of 12 kHz. The microphone has an open circuit sensitivity of 7mV/Pa, a maximum input sound level 138 dB for 1% THD at 1 KHz, and its A-weighted signal-to-noise ratio is 67 dB. It is powered from an external DC phantom power supply.
- Headset type microphone (labelled as c2). Allows medium-quality recording. This type of microphone without preamp is standard equipment on hands-free call centres. The frequency range of the microphone shall be 20Hz to 20kHz, its sensitivity at 1kHz 7mV/Pa; maximum input sound level for 3% THD 130 dB and its A-weighted signal-to-noise ratio > 64 dB.
- Table microphone (labelled as c3). Allows medium-quality recording. Supported by a tripod stand and powered by a set of batteries embedded in the body of the microphone.

- Low quality microphone (labelled as c4). Similar to the ones usually supplied as an external component in many computers.

For each speaker different pre-established speech is recorded in 3 different sessions:

- Three productions of the Spanish vowel /a/ with normal tone and volume and lasting from 6 to 8 seconds.
- One production of the five Spanish vowels /a/, /e/, /i/, /o/ and /u/ with 3 sec duration per vowel.
- Three productions of the intended to be phonetically balanced Spanish sentence: "*Es hábil un solo día"*, with normal tone and volume.
- Recording of a reference text lasting approximately 1 minute.
- Three productions of an assigned fictious name and surname.
- Ten productions of the same 4-digit pin.
- Three productions of the 10 digits (randomly ordered in each utterance).
- Three productions of a common name and surname.
- One production of twenty different 4-digit pin (the same for all speakers)

Additionally, some of these recordings have also been recorded through a GSM mobile network (labelled as T). In this way, we are able to model the variability introduced by the transmission channel.

| | SESSION 1 | | | | | SESSION 2 | | | | | SESSION 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Type of utterance | C1 | C2 | C3 | C4 | T | C1 | C2 | C3 | C4 | T | C1 | C2 | C3 | C4 | T |
| Vowel /a/ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5 Spanish vowels | 1 | 1 | 1 | 1 | 1 | | | | | | | | | | |
| Sentence | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Reference text | 1 | 1 | 1 | 1 | 1 | | | | | | | | | | |
| Assigned name | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Common pin | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 10 digits | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 | |
| Common name | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4-digit pin | 20 | 20 | 20 | 20 | 20 | | | | | | | | | | |

**Table 4-2** Number of utterance per speaker for each session and channel in the HESPERIA database.

### 4.1.2 The Database ALBAYZIN

ALBAYZIN consists of a Spanish language corpus set that has been collected under the project of the same name[2]. The project has been carried out by a consortium of six research groups, namely:

- Univ. de Granada – Dpto. de Electrónica y Tecnología de Computadores

[2] Project ALBAYZIN was funded by the Comisión Interministerial de Ciencia y Tecnología(CICYT) (TIC91-1488-C06)

- Univ. Politécnica de Cataluña – Dpto. de Teoría de la Señal y Comunicaciones
- Univ. Politécnica de Madrid – Dpto. de Ingeniería Electrónica
- Univ. Politécnica de Madrid – Dpto. de Señales, Sistemas y Radiocomunicaciones
- Univ. Autónoma de Barcelona – Dpto. de Filología Española
- Univ. Politécnica de Valencia – Dpto. Sistemas Informáticos y Computación.

The speech corpus is divided in 3 different corpora:

- Generic phonetic corpus (GPC): Constitutes a reference framework of the Spanish language, so neither syntactic nor semantic restrictions applied. It contains 6800 phrases uttered by 204 speakers.
- Specific corpus (SC): Application dependent utterances with semantic restrictions. More specifically, it contains 6800 phrases uttered by 136 speakers with information on geographic queries.
- Lombard corpus (LC): Contains speech produced under the Lombard effect, i.e. when speakers support a high-noise level in their ears. It contains 50 phrases uttered by 40 speakers.

The database contains recordings from 304 different speakers (equally distributed among gender and with ages ranging from 18 to 55), while the amount of utterances is 15600 lasting in average 4 seconds. Table 4-3 and Table 4-4 summarise the number of utterances per speaker regarding the age distribution and the corpus to which belong.

Design, recording and labelling of the corpus have followed the standards proposed in the European project Speech Assessment Methods (SAM) ESPRIT project 2589. In particular, regarding the recordings, they have been made in a recording studio, using a Shure SM 10A headset microphone. In a post-processing step, the signals were high pass filtered to eliminate low frequency noise induced in the recording process.

| Gender | Age | # Speakers | # Utterances GPC | # Utterances SC | # Utterances LC |
|---|---|---|---|---|---|
| Female | 18-31 | 2 | 200 | 50 | 50 |
| | | 32 | 25 | | |
| | | 6 | 50 | 50 | 50 |
| | | 4 | 50 | | 50 |
| | | 6 | 25 | 50 | |
| | | 2 | 25 | 50 | 50 |
| | | 22 | | 50 | |
| | | 3 | | 50 | 50 |
| | 31-40 | 16 | 25 | | |
| | | 3 | 25 | 50 | |
| | | 1 | 25 | 50 | 50 |
| | | 3 | 50 | | |
| | | 2 | 50 | | 50 |
| | | 12 | | 50 | |
| | | 2 | | 50 | 50 |
| | 41-55 | 16 | 25 | | |
| | | 3 | 25 | 50 | |
| | | 1 | 25 | 50 | 50 |
| | | 3 | 50 | | |
| | | 2 | 50 | | 50 |
| | | 10 | | 50 | |
| | | 1 | | 50 | 50 |

**Table 4-3** Utterance distribution for female speakers regarding age and sub-corpus

| Gender | Age | # Speakers | # Utterances GPC | # Utterances SC | # Utterances LC |
|---|---|---|---|---|---|
| Male | 18-31 | 2 | 200 | 50 | 50 |
| | | 32 | 25 | | |
| | | 6 | 50 | | |
| | | 4 | 50 | | 50 |
| | | 6 | 25 | 50 | |
| | | 2 | 25 | 50 | 50 |
| | | 22 | | 50 | |
| | | 3 | | 50 | 50 |
| | 31-40 | 16 | 25 | | |
| | | 3 | 25 | 50 | |
| | | 1 | 25 | 50 | 50 |
| | | 3 | 50 | | |
| | | 2 | 50 | | 50 |
| | | 12 | | 50 | |
| | | 2 | | 50 | 50 |
| | 41-55 | 16 | 25 | | |
| | | 3 | 25 | 50 | |
| | | 1 | 25 | 50 | 50 |
| | | 3 | 50 | | |
| | | 2 | 50 | | 50 |
| | | 10 | | 50 | |
| | | 1 | | 50 | 50 |

**Table 4-4** Utterance distribution for male speakers regarding age and sub-corpus

### 4.1.3 The Database MOBIO

MOBIO is a bimodal database consisting of face images and voice information, captured on mobile devices (actually using NOKIA 93i mobile phone, and a standard 2008 MacBook). The speech corpus contains a total of 152 speakers (100 males and 52 females) that have been recorded in two different phases, with 6 different sessions per phase, in 6 different sites:

- University of Oulu (OULU)
- Idiap Research Institute (IDIAP)
- University of Avignon (LIA)
- University of Manchester (UMAN)
- University of Surrey (UNIS)
- Brno University of Technology (BUT)

Regarding the audio data, all audio was collected in English, although English is not always the native language of subjects. Additionally, as it was captured using a mobile device (not placed in a fixed position), it contains high variability in terms of both quality and acquisition environments (which means real noise background).

As previously noted the database was captured in two different phases. In the first phase, each subject was asked to provide the following information:

- 5 Short responses to the following questions: “What is your name?”, “What is your address?”, “What is your birth date?”, “What is your license number?”, and “What is your credit card number?”
- 5 Short responses (lasting 5 seconds approximately) to a set of questions randomly extracted from a pool of 40 questions.
- 10 responses (lasting 10 seconds approximately) to a set of questions randomly extracted from a pool of 40 questions.
- 1 pre-defined text lasting at least 10 seconds.

The second phase was slightly shorter as it only contains:

- 5 Short responses to the following questions: “What is your name?”, “What is your address?”, “What is your birth date?”, “What is your license number?”, and “What is your credit card number?”
- 5 responses (lasting 10 seconds approximately) to a set of questions randomly extracted from a pool of 40 questions.
- 1 pre-defined text lasting at least 10 seconds.

Table 4-5 summarises the number of speakers per site, as well as the number sessions and number of recordings per speaker. More technical details about MOBIO database can be found in [McCool,2012].

| SITE | Phase I | | | Phase II | | |
|---|---|---|---|---|---|---|
| | #Subjects (female /male) | #Sessions | #Recordings | #Subjects (female/ male) | #Sessions | #Recordings |
| BUT | 33 (15/18) | 6 | 21 | 32 (15/17) | 6 | 11 |
| IDIAP | 28 (7/21) | 6 | 21 | 26 (5/21) | 6 | 11 |
| LIA | 27 (9/18) | 6 | 21 | 26 (8/18) | 6 | 11 |
| UMAN | 27 (12/15) | 6 | 21 | 25 (11/14) | 6 | 11 |
| UNIS | 26 (6/20) | 6 | 21 | 24 (5/19) | 6 | 11 |
| UOULU | 20 (8/12) | 6 | 21 | 17 (7/10) | 6 | 11 |

**Table 4-5** Utterance, session and speaker distribution for each of the sites participating in the MOBIO database

#### 4.1.4 NIST SRE databases

As it will be pointed out later on, the data made available for the different speaker recognition evaluations have changed over the years since its first edition in 1996. The purpose of the present section is not to provide a detail description of all databases released by NIST from then until now. However a general and brief description (mainly extracted from the different evaluation plans) of the latest ones will be presented due to their interest for the 2010 and 2012 SREs.

Regardless the speech corpora, data is stored as NIST SPHERE format. In this format, files consist in a simple, flexible and self-describing file header followed by the 8-bit μ-law sample data. The header, which is readable as plain text, includes auxiliary information about the speech data in the file: number of samples, sampling rate, number of channels, kind of sample coding, as well as the language of the conversation, whether or not the data was recorded over a telephone line, and whether or not the data is from an interview session. However, it does not contain information on the type of telephone transmission channel (cellular, cordless, land line), the type of telephone instrument involved (speaker-phone, head-mounted, ear-bud, hand-held), or the room microphone type used for the interview data (ear-bud/lapel microphone, mini-boom microphone, courtroom microphone, conference room microphone, distant microphone, near-field microphone, PC stand microphone, micro-cassette microphone).

The databases contain not only SPHERE files, but the Automatic Speech Recognition (ASR) transcription data for English recordings. As the transcription is performed by automatic means, it is not error free.

The speech data have been collected by the Linguistic Data Consortium (LDC) as part of the various phases of its Mixer project or of its earlier conversational telephone collection projects. Regarding phone records, the platform Fishboard was used to collect the data from selected pairs of speakers using different telephone instruments. Both channels of the telephone conversations are provided, although in early evaluations

these conversations were processed through echo cancelling software before database release, this processing step has been omitted in the 2008, 2010 and 2012 databases.

A further difference between the available NIST databases is the language used by the speakers involved in the recordings. For instance, in 2004, 2005, 2006 and 2008 databases, most of the data was recorded in English; however they also contain conversations recorded in Arabic, Mandarin, Russian, Spanish, etc. Additionally, conversations recorded in languages other than English involved bilingual speakers which mean that English may or may not be their native language. On the other hand, in the 2010 and 2012 databases only English data is provided, but English may not be the native language of all the speakers that appear in the recordings. We will now briefly present the different NIST databases that we will use in our experiments, starting with 2004 database to the most recent and focusing only in the data specifically selected for the experiments.

- NIST SRE 2004: The selected subset of the 2004 SRE database only contains conversational telephone speech in English. Each selected recording consists of approximately five minutes of conversation collected over telephone channels, where both the transmission channel and the microphone type used are provided and labelled as follows:
    - Transmission channel:
        - Cellular
        - Cordless
        - Land line
    - Microphone type:
        - Speaker-phone
        - Head-mounted
        - Ear-bud
        - Hand-held

  All the recordings used in the set of experiments contain just one channel with just one speaker. The total number of recordings used is 3363, following the distribution reflected in Table 4-6:

| Channel | Gender | Number of recordings | Number of different speakers |
|---|---|---|---|
| Telephone | Male | 1375 | 125 |
| | Female | 1988 | 184 |

**Table 4-6** Number of recordings and speakers in the NIST SRE 2004 database

- NIST SRE 2006: The selected set of recordings from the 2006 database contains approximately five minutes of conversation collected over telephone channels or a specific microphone. Both the transmission channel and the microphone type used when telephone recordings are involved are provided in the header of each recording. Specifically, telephone recordings are classified like in NIST SRE 2004 database, while for the case of microphone data, different microphones have been used to capture the speakers' voice:

- Ear-bud/lapel microphone
- Mini-boom microphone
- Courtroom microphone
- Conference room microphone
- Distant microphone
- Near-field microphone
- PC stand microphone
- Micro-cassette microphone

All microphone recordings contain two channels with different speakers speaking on each side. Although one might think that we are facing high quality recordings as they are captured using a microphone, nothing could be further away from truth. Actually, microphones are placed in different locations of the recording room, not necessary closed to the speakers, so recording levels can become minimal and even result in barely audible recordings. The total number of recordings used is 11968, which can be classified as follows:

| Channel | Gender | Number of recordings | Number of different speakers |
|---|---|---|---|
| Microphone | Male | 1220 | 38 |
| | Female | 1474 | 46 |
| Telephone | Male | 3993 | 479 |
| | Female | 5281 | 666 |
| TOTAL | Male | 5213 | 479 |
| | Female | 6755 | 666 |

**Table 4-7** Number of recordings and speakers in the NIST SRE 2006 database

It must be noted that the speakers in the microphone recording set are a subset of the telephone recording set.

- NIST SRE 2008: The selected recordings from this database include not only conversational telephone speech data but also conversational speech data recorded over a microphone channel involving an interview scenario, and conversational telephone speech recorded over a microphone channel. All recordings contain two channels with different speakers talking on each side except in the case of the interview scenario. In this last case, one of the channels contains approximately a three minute conversational segment involving two different speakers, i.e., the interviewer and the interviewee, while the other channel contains no information at all.

  The total number of recordings selected to be used in the experiments is 14409. Although as early mentioned this database contains recordings involving different languages, for practical reasons only those in English have been considered. Table 4-8 provides a brief description of the selected contents of the database.

| Channel | Gender | Number of recordings | Number of different speakers |
|---|---|---|---|
| Microphone | Male | 616 | 58 |
| | Female | 803 | 77 |
| Interview | Male | 1606 | 63 |
| | Female | 2196 | 87 |
| Telephone | Male | 3439 | 472 |
| | Female | 5749 | 798 |
| TOTAL | Male | 5661 | 472 |
| | Female | 8748 | 801 |

**Table 4-8** Number of recordings and speakers in the NIST SRE 2008 database

- NIST SRE 2010 database: The data contained in this database can be grouped according to the different tasks proposed on the SRE 2010 evaluation plan. Specifically, the database can be clustered in four different groups:
    - 10-second Excerpts: All recordings in this group are two-channel excerpts of a telephone conversation containing between 8 and 12 seconds of actual target speech on the channel of interest.
    - Two-channel Conversations: All recordings in this group contain two-channel telephone conversation excerpts which include approximately five minutes from a longer conversation. The excision points are chosen so as not to include partial speech turns. Actual target speech will last much less than 5 minutes.
    - Interview Segments: Recordings in this group are characterised by containing interview segments, extracted from a longer interview session, which vary in duration in the range of 3 to 8 minutes. Two channels are also provided in this case. The channel of interest will provide the speech captured by a microphone placed somewhere in the interview room. The alternative channel will contain the interviewer's speech recorded from a head-mounted, close-talking microphone.
    - Summed-channel conversations: This group will consist of summed-channel telephone conversation segments of about 5 minutes in duration. This means that the two channels of a conversation involving target and non-target speaker are summed together in a single channel.

  In all the experiments carried out, 10-second excerpts and summed-channel conversations are not used. Therefore, the number of different speakers and recordings that are going to be used are shown in Table 4-9.

| Channel | Gender | Number of recordings | Number of different speakers |
|---|---|---|---|
| Microphone | Male | 893 | 171 |
| | Female | 971 | 205 |
| Interview | Male | 4501 | 199 |
| | Female | 5358 | 231 |
| Telephone | Male | 1143 | 189 |
| | Female | 1305 | 219 |
| TOTAL | Male | 6537 | 206 |
| | Female | 7634 | 234 |

**Table 4-9** Number of recordings and speakers in the NIST SRE 2010 database

- NIST SRE 2012 database: The type of recordings contained in this database is quite similar to the recordings in the NIST SRE 2010. In particular it contains both telephone and microphone recordings, however, some segments include additive noise (noise added as a post-processing step after recording) or have been recorded in an intentionally noisy environment or both. Table 4-10 shows the number of different speakers and recordings, grouped by type, that are used in the experiments. It must be noted that the speaker id is not provided for all the recordings.

| Channel | Gender | Number of recordings | Number of different speakers |
|---|---|---|---|
| Microphone | Male | 1934 | 43 |
| | Female | 2357 | 72 |
| Interview | Male | 15396 | 53 |
| | Female | 20658 | 94 |
| Telephone | Male | 8909 | 148 |
| | Female | 14258 | 231 |
| TOTAL | Male | 26239 | 154 |
| | Female | 37273 | 240 |

**Table 4-10** Number of recordings and speakers in the NIST SRE 2012 database

## 4.2 PRACTICAL SCENARIOS

As previously discussed, not only the availability of specific databases plays an important role in the advancement of the state-of-the-art. It is also necessary to provide a common framework in order to test the systems under the same conditions using the same data, thus measuring the advances on the corresponding technology (speaker recognition in this case). In this respect, the quintessential common framework in the area of speaker recognition has been provided by the National Institute of Standards and Technology (NIST), through the Multimodal Information Group, since 1996 in the form of Speaker Recognition Evaluations (SRE). However, as it will be discussed later there are some issues on the evaluation plans that result in systems not concurring under the same conditions and therefore a straight comparison based in recognition rates seems not to be always adequate. This problem also occurs in the evaluation organised by Idiap Research Institute on mobile environments.

For this reason, without ruling out the participation in these evaluations, in this section we are going to describe the practical scenarios in which, using the databases presented

so far, we have tested the validity of the hypothesis stated at the beginning of this thesis. The proposed scenarios cover multiple real-life situations, ranging from high-quality data in a text-constrained environment (especially used in access control in security environments) to maximal variability data in terms of data quality and message content (especially used in forensic applications).

### 4.2.1 Text-Constrained Speaker Recognition

When putting into operation a speaker recognition system, especially in security environments, it is usual to find out that the data used during the training phase and the recognition phase are similar and yet different and variable, in order to avoid the use of fraudulent recordings. The use of 4-digit pins can be an interesting solution, as it requires a short training phase, does not require the user to produce long sentences during the test/operation phase and can also be changed from day-to-day easily. Under this assumption, we have defined two closed-set text-constrained speaker verification experiments without cross-gender trials using the database HESPERIA. In order to test the robustness of the algorithms to channel variations the first scenario involves training and testing phases using microphone data (typically used on facility access control) while the second scenario involves microphone data on training and telephone data in the test phase (with possible application to telephone banking).

- Scenario 1 (mic-mic):

  In this scenario, 109 male speakers and 93 female speakers with ages ranging from 18 to 65 are represented. All the recordings were captured under controlled conditions and using 4 different microphone channels. Each recording consisted of an utterance of a specific 4-digit pin. So the problem can be defined as closed-set text-constrained speaker verification (without cross-gender trials). The data is divided in 3 different subsets as follows:

  - Background set: In this subset we include 25 speakers per gender with all age ranges represented, and for each of them sixty recordings, taken over four different microphones, each of them corresponding to a different 4-digit pin. The data assigned to this set is only used to learn the background parameters of the algorithm or for normalisation purposes.
  - Development set: This set is split into two different subsets: enrolment and test, and none of them include data already in the background set. The first one, the enrolment subset, is used to create a model of each of the target speakers. Like in the background set, we include 25 speakers per gender, and 20 recordings per speaker are available for training each model (corresponding to a different 4-digit pin recorded over one specific microphone). The second subset, i.e. the test, contains a list of audio samples that must be tested against all the target speakers. For this purpose, we reserve 40 additional files per speaker, resulting in a total of 25000 trials per gender. These 40 recordings consist of 4-digit pins different from the ones used during training phase, and recorded over four different microphones. The data on this set are supposed to be used to tune meta-parameters of the algorithm, (e.g. number of Gaussians, dimension of subspaces, etc.). The recognition rate achieved, in terms of *EER*, is used to define a score threshold, $\theta_{dev,}$ which will be used to evaluate the performance of the recognition system.

- Evaluation set: The final evaluation performance is analysed using this dataset, which includes 43 female speakers and 57 male speakers, and is provided in the form of *Half Total Error Rate* (HTER). HTER can be defined as the average between the number of *False Acceptance Rate* (FAR) and *False Rejection Rate* (FRR), at the operation point defined by $\theta_{dev}$. For each speaker, 20 files are used to create the model, while 40 recordings are marked as test. These 40 recordings consist of 4-digit pins different from the ones used during the training phase, and recorded over four different microphones. A score must be provided for each trial in the form of a log-likelihood ratio, representing how accurately the test segment is classified as containing, or not, speech for the target speaker against which it is confronted. All files marked as test are confronted to each of the speakers. Taking into account that no cross-gender trials are performed, this leads to 129960 male trials and 73960 female trials.

| | Background | |
|---|---|---|
| | Speakers | #Files |
| MALE | 25 | 1500 |
| FEMALE | 25 | 1500 |
| TOTAL | 50 | 3000 |

| | Development | | | | |
|---|---|---|---|---|---|
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 25 | 500 | 25 | 1000 | 25000 |
| FEMALE | 25 | 500 | 25 | 1000 | 25000 |
| TOTAL | 50 | 1000 | 50 | 2000 | 50000 |
| | Evaluation | | | | |
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 57 | 1140 | 57 | 2280 | 129960 |
| FEMALE | 43 | 860 | 43 | 1720 | 73960 |
| TOTAL | 100 | 2000 | 100 | 4000 | 203920 |

**Table 4-11** Description of the contents of the different subsets of the HESPERIA database for Scenario 1

- Scenario 2 (mic-tel):

  In this scenario, 80 male speakers and 75 female speakers with ages ranging from 18 to 65 are represented. The amount of speakers is less than in previous conditions since not all the speakers in the database HESPERIA have been recorded over a telephone channel. Like in Scenario 1, all the recordings were captured under controlled conditions but in this case using 5 different channels (4 microphones and the GSM mobile network). Each recording consisted of an utterance of a specific 4-digit pin. So the problem can be defined as closed-set text-constrained speaker verification (without cross-gender trials).

  Again, we have divided the data in 3 different subsets as follows:

  - Background set: Same as in Scenario 1.
  - Development set: This set is split into two different subsets: enrolment and test, and none of them include data already in the background set.

The first one, the enrolment subset, is used to create a model of each of the target speakers. It consists of 25 speakers per gender and 20 recordings per speaker are available for training each model (corresponding to a different 4-digit pin recorded over one specific microphone). The second subset, i.e. the test, contains a list of audio samples that must be tested against all the target speakers. For this purpose, we reserve 10 additional files per speaker, resulting in a total of 6250 trials per gender. These 10 recordings consist of 4-digit pins different from the ones used during the training phase, and recorded over a telephone channel. The data on this set are supposed to be used to tune meta-parameters of the algorithm, (e.g. number of Gaussians, dimension of subspaces, etc.). The recognition rate achieved, in terms of *EER*, is used to define a score threshold, $\theta_{dev,}$ which will be used to evaluate the performance of the recognition system.

- Evaluation set: The final evaluation performance is analysed using this dataset, which includes 20 female speakers and 30 male speakers, and is provided in the form of *Half Total Error Rate* (*HTER*). *HTER* can be defined as the average between the number of *False Acceptance Rate* (*FAR*) and *False Rejection Rate* (*FRR*), at the operation point defined by $\theta_{dev}$. For each speaker, 20 files (corresponding to a different 4-digit pin recorded over one specific microphone) are used to create the model, while 10 recordings are marked as test. These 10 recordings consist of 4-digit pins different from the ones used during the training phase, and recorded over a telephone channel. It is important to note, that in this case we are testing an extreme case of channel variability as training is performed on a microphone channel and testing on the telephone channel. A score must be provided for each trial in the form of a log-likelihood ratio, representing how accurately the test segment is classified as containing, or not, speech for the target speaker against which it is confronted. All files marked as test are confronted to each of the speakers. Taking into account that no cross-gender trials are performed, this leads to 9000 male trials and 4000 female trials.

Table 4-12 summarises the amount of data included in each of the subsets, in terms of number of recordings, number of target speakers and the total amount of trials.

| | Background | |
|---|---|---|
| | Speakers | #Files |
| MALE | 25 | 1500 |
| FEMALE | 25 | 1500 |
| TOTAL | 50 | 3000 |

| | Development | | | | |
|---|---|---|---|---|---|
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 25 | 500 | 25 | 250 | 6250 |
| FEMALE | 25 | 500 | 25 | 250 | 6250 |
| TOTAL | 50 | 1000 | 50 | 500 | 12500 |

| | Evaluation | | | | |
|---|---|---|---|---|---|
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 30 | 600 | 30 | 300 | 9000 |
| FEMALE | 20 | 43 | 20 | 200 | 4000 |
| TOTAL | 50 | 100 | 50 | 500 | 13000 |

**Table 4-12** Description of the contents of the different subsets of the HESPERIA database for Scenario 2

The results achieved in the set of experiments previously described as well as the speaker recognition system build to face these experiments will be presented in Chapter 5.

#### 4.2.2 Text-Independent Speaker Recognition

The vulnerability of automatic speaker verification systems to imposture or spoofing is widely acknowledged. In order to make a speaker verification system robust for instance to playback attacks, a possible solution may be not to limit the message used for recognition to a specific pattern (for instance, 4-digit pins). Thus asking the speaker to provide in a limited time a random phrase not known in advance, reduce the chances of attacking the system.

Under this assumption, we have defined a closed-set text-independent speaker verification experiment without cross-gender trials using the database ALBAYZIN. As previously said, this database provides a set of recordings acquired under controlled environments meeting minimum quality standards, which allow algorithm fast tuning.

In this scenario, 304 speakers equally distributed by gender with ages ranging from 18 to 55 are used. All the recordings were captured through just one channel, so no channel variability can be tested. Like in the Text-Constrained scenario, no cross-gender trials are to be performed, and the methodology followed is similar. Specifically, we have divided the data in 3 different subsets, as follows:

- Background set: In this subset we include 25 speakers per gender, and for each of them 25 files from the generic sub-corpus. The data assigned to this set is only used to learn the background parameters of the algorithm or for normalisation purposes.
- Development set: This set is split into two different subsets: enrolment and test, and none of them include data already in the background set (i.e. different speakers are present). The first one, the enrolment subset, is used to create a model of each of the target speakers. Like in the background set, we include 25 speakers per gender, however only 3 files per speaker are available for training each model (4 seconds of information per file). The second subset, i.e. the test, contains a list of audio samples that must be tested against all the target speakers. For this purpose, we reserve 22 additional files per speaker, resulting in a total of 13750 trials per gender. The data on this set is supposed to be used to tune meta-parameters of the algorithm, (e.g. number of Gaussians, dimension of subspaces, etc.). Like in previous experiments, the recognition rate achieved, in terms of *EER*, is used to define a score threshold, $\theta_{dev,}$ which will be used to evaluate the performance of the recognition system.
- Evaluation set: The final evaluation performance is analysed using this dataset, which includes 88 speakers per gender, and is provided in form of *Half Total*

*Error Rate* (*HTER*). *HTER* can be defined as the average between the number of *False Acceptance Rate* (*FAR*) and *False Rejection Rate* (*FRR*), at the operation point defined by $\theta_{dev}$. For each speaker, 3 files are used to create the model, while 47 recordings (taken from either the generic sub-corpus or the specific sub-corpus) are marked as test. A score must be provided for each trial, for instance in the form of a log-likelihood ratio, representing how accurately the test segment is classified as containing, or not, speech for the target speaker against which it is confronted. All files marked as test are confronted to each of the speakers. Taking into account that no cross-gender trials are performed, this leads to 363968 trials per gender.

Table 4-13 summarises the amount of data included in each of the subsets, in terms of number of recordings, number of target speakers and the total amount of trials.

| | Background | |
|---|---|---|
| | Speakers | #Files |
| MALE | 25 | 625 |
| FEMALE | 25 | 625 |
| TOTAL | 50 | 1250 |

| | Development | | | | |
|---|---|---|---|---|---|
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 25 | 75 | 25 | 550 | 13750 |
| FEMALE | 25 | 75 | 25 | 550 | 13750 |
| TOTAL | 50 | 150 | 50 | 1100 | 27500 |
| | Evaluation | | | | |
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 88 | 264 | 88 | 4136 | 363968 |
| FEMALE | 88 | 264 | 88 | 4136 | 363968 |
| TOTAL | 176 | 528 | 176 | 8272 | 727936 |

**Table 4-13** Description of the contents of the different subsets of the ALBAYZIN database

The results achieved in this experiment as well as the speaker recognition system built to face it will be presented in Chapter 5.

### 4.2.3 Text-Independent Speaker Recognition in Mobile Environments

An additional problem that we have to face is the fact that we live in a mobile connected world. This means for instance that the access to a system can be performed remotely using a cell phone anywhere under different noisy environments. For this reason, it seems appropriate to verify how a speaker recognition system behaves in a text-independent mobile environment. The importance of this scenario is demonstrated by the existence of an international speaker recognition evaluation meeting this constrains. The evaluation proposed for the 2013 International Conference on Biometrics (ICB 2013), which mainly consists in close-set speaker recognition using the database MOBIO, i.e. in mobile environments, allows us not only to test the performance of our speaker recognition system in an impartial scenario, but to compare our approach with speaker recognition systems designed by other research groups. Therefore the scenario

proposed in this section follows the evaluation plan provided for the "2013 SRE in Mobile Environments", so that the results achieved can be submitted.

The aim of this competition is to determine whether a specified target speaker is present or not in a given segment of speech which has been recorded in mobile environments. The speakers (and thus the recordings) included in the MOBIO database are split into three different subsets as follows:

- Background training set: The data assigned to this set can be used only to learn the background parameters of the algorithm (UBM, subspaces, etc.) or for normalisation purposes. Use of additional information is allowed as long as it is explicitly documented in the system's description. However, allowing the use of additional data causes the speaker recognition systems not to participate under the same conditions. Therefore, it is not possible to strictly establish whether the different recognition rates achieved by different systems are influenced by better parameterisation techniques, better classification methods, better intersession compensation and normalisation techniques or just by the use of more favourable additional databases. For this reason, we are not going to add any extra data to this set.
- Development set: The data assigned to this set is split into two subsets: enrolment and test. The first one is used to create a model of each of the target speakers included. The second one contains a list of audio samples that must be tested against the target speakers. The data on this set is supposed to be used to tune meta-parameters of the algorithm (e.g. number of Gaussians, dimension of subspaces, etc.). The recognition rate, regarding *EER*, achieved with this development set is used to define a score threshold, $\theta_{dev}$, which will be used to evaluate the performance of the recognition system.
- Evaluation set: The final evaluation performance is analysed using this set, and is provided in the form of *Half Total Error Rate* (*HTER*). *HTER* can be defined as the average between the number of *False Acceptance Rate* (*FAR*) and *False Rejection Rate* (*FRR*), at the operation point defined by $\theta_{dev}$. Like in the development set, some recordings are provided in order to create a model of each of the target speakers, while another set of audio recordings are marked as test. A score must be provided for each trial, for instance in the form of a log-likelihood ratio, representing how accurately the test segment is classified as containing, or not, speech for the target speaker against which it is confronted.

Table 4-14 summarises the amount of data included in each of the subsets, in terms of number of recordings, number of target speakers and the total amount of trials.

| | Background | |
|---|---|---|
| | Speakers | #Files |
| MALE | 37 | 7104 |
| FEMALE | 13 | 2496 |
| TOTAL | 50 | 9600 |

| | Development | | | | |
|---|---|---|---|---|---|
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 24 | 120 | 24 | 2520 | 60480 |
| FEMALE | 18 | 90 | 18 | 1890 | 34020 |
| TOTAL | 42 | 210 | 42 | 4410 | 94500 |

| | Evaluation | | | | |
|---|---|---|---|---|---|
| | Enrolment | | Test | | |
| | #Targets | #Files | Speakers | #Files | #Trials |
| MALE | 38 | 190 | 38 | 3990 | 151620 |
| FEMALE | 20 | 100 | 20 | 2100 | 42000 |
| TOTAL | 58 | 290 | 58 | 6090 | 193620 |

**Table 4-14** Description of the contents of the different subsets of the MOBIO database

The results achieved in this competition as well as the speaker recognition system build to face it will be presented in Chapter 5.

#### 4.2.4 NIST SRE Evaluations

The goal of the NIST SRE series is to contribute to drive the technology of text-independent speaker recognition forward, exploring and testing promising new algorithmic approaches. Although initially conducted on an annual basis, since 2006 the evaluation takes place once every two years. It is also important to note the great interest aroused by this evaluation, which since its first edition in 1996 (in which only 12 groups took part on it), the number of participants has increased steadily, reaching the number of 58 sites in its last edition.

However, after our first participation in the NIST SRE 2010, we identified some key issues that considerably influence the recognition results achieved in the evaluation process, therefore making a direct comparison of results produced by different systems somehow misleading:

- Quality of the evaluation database → A key aspect that must be considered is the database provided by the organisers for evaluation purposes. As discussed in previous sections, the database provided by NIST for the SRE 2010 presents great intra-speaker variability due to: use of different transmission channels, use of different recording devices in different places, inclusion of recordings with diverse vocal effort, inclusion of recordings collected from the same speaker years apart (age evolution), and the fact of being text-independent. Additionally, it contains a large number of different speakers classified by gender as long as no cross-gender trials are required. Moreover, this database presents an important problem: the quality of recordings. More specifically, we have to deal with recordings for which it is very difficult or even impossible to extract speaker's relevant information, due to noise (see Figure 4-1). We also face recordings that do not meet minimum quality standards (see Figure 4-2), or simply recordings that despite its labelling do not contain the speaker's information that should be present on it (see Figure 4-3).

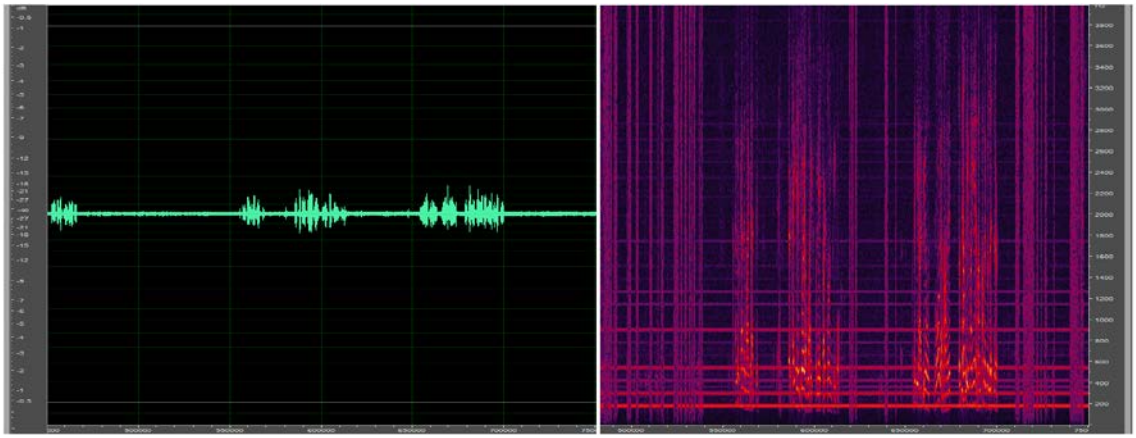


**Figure 4-1** Noisy telephonic recording (time domain – left – and frequency domain – right) in NIST SRE 2010 database

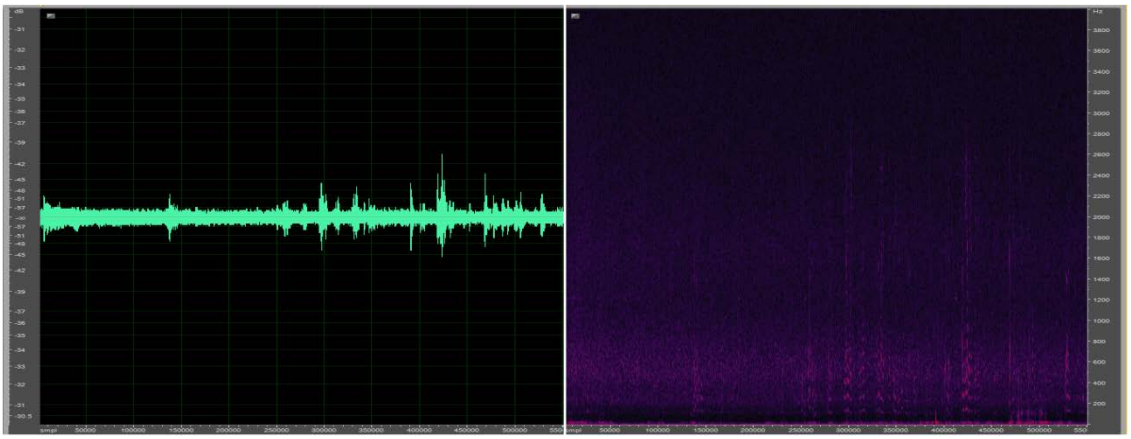


**Figure 4-2** Low quality recording (time domain – left – and frequency domain – right) in NIST SRE 2010 database

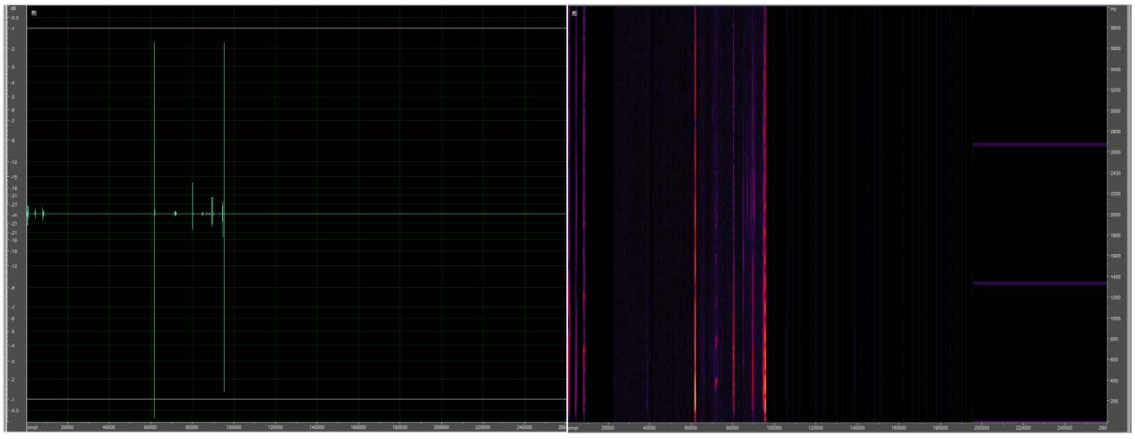


**Figure 4-3** Incorrect labelled recording (time domain – left – and frequency domain – right) in NIST SRE 2010 database. No speaker information available

- Labelling errors → It also must be noted that in the NIST SRE2012 an updated key was released, that affects multiple databases (2004, 2006 and 2008). The errors reflected on the update refer not only to incorrect assignment of the speaker id to files but also to a mismatch in the gender assigned to some speakers.

- Use of additional databases → Another important issue in these evaluations is the fact that there is no control and no limit in the number and type of additional databases that can be used by the systems in the different parts, i.e. UBM creation, normalisation, inter-session compensation, etc. This means that systems are somehow not participating under the same conditions and therefore it is not possible to strictly establish whether the different recognition rates achieved by different systems are influenced by better parameterisation techniques, better classification methods, better inter-session compensation and normalisation techniques or just by the use of more favourable additional databases.
- Good practice of participants → Last but not least, it seems necessary to emphasise that the organisation responsible of the evaluations only collects and analyses the results submitted by the different sites participating. This implies that organisers do not control in any way whether systems strictly comply with the regulation, especially regarding the use of the provided database for purposes other than those specified in the evaluation plan. In particular, the use of not allowed information (train/test recordings) in the inter-session compensation or normalisation phases is not verified, and this practice notably influences the tuning of the systems and therefore the recognition rates achieved by them. In the NIST SRE 2010 some sites where accused of this practice, although no proves where provided.

All these aspects have clearly influenced our first participation in a NIST SRE, specifically the NIST SRE 2010. For this reason, we are not going to present the results achieved by our system in that evaluation, but we are going to use the experience acquired to analyse whether the gender-dependent extended biometric parameterisation proposed in this thesis is able to provide an improvement in terms of recognition rates respect to classical parameterisation approaches in this highly adverse environment.

The procedure followed consists in using the NIST SRE 2010 plan as starting point. In other words, we use the set of tests posed in the plan as development set, and then we will verify if the achieved results can be extended to the NIST SRE 2012, taking into account that the NIST SRE 2012 presents some changes compared to the NIST SRE 2010.

Although the NIST SRE 2010 plan is quite ambitious in the number of tests covered, in what follows, we present the most relevant information that we have used, which mainly refers to the core-core task. The interested reader may refer to [NIST,2010] for further information on the whole plan. Once the core-core task is presented and briefly analysed, we will present the NIST SRE 2012 which will act as evaluation set.

- Background training set: Like in previously presented scenarios, the data assigned to this set is going to be used to learn the background parameters of the algorithm (UBM, subspaces, etc.) or for normalisation purposes. Although according to the SRE 2010 plan there is no limit in the data used for this purpose, we have selected the NIST SRE 2004, 2006 and 2008 databases presented in section 4.1.4. Thus the selected data covers a great number of different speakers, as well as different recording devices, transmission channels, and recording environments. Table 4-6, Table 4-7 and Table 4-8, provide a description of the amount of data that we are going to manage in this set, which clearly is much greater than any of the ones managed so far.

- Development set: As we have already said, we are going to use the core-core task defined in the NIST SRE 2010 plan as development set. NIST SRE focuses on speaker detection in the context of conversational speech over multiple types of channels. This means that given a target speaker and a test speech segment, the goal is to determine whether the target is speaking in the test segment, taking into account that no cross-gender detection will be required. The decision is twofold, for each trial a hard decision (true or false) must be made as well as a score in the form of log-likelihood ratio needs to be provided. Therefore the development set is divided into two different subsets: enrolment and test. Regarding the recording characteristics, it should be noted that all recordings (no matter if they are for train or test purposes) are in English. Additionally, training and test recordings include telephone speech recorded over different telephone transmission channels (cellular, cordless, land-line) and different telephone instruments (speaker-phone, head-mounted, ear-bud or hand-held) as well as speech recorded over a room microphone channel and conversational speech from an interview scenario recorded over a room microphone channel. It must also be noted that some recordings may be affected by vocal effort variability. Finally, the amount of speech available for training or test ranges from 3 to 8 minutes.

  Table 4-15 reflects the number of models that must be generated, along with the number of different speakers and the number of trials that must be performed.

| | Number of models to train | Number of different speakers | Number of trials |
|---|---|---|---|
| MALE | 2434 | 206 | 273787 |
| FEMALE | 3026 | 234 | 336961 |

**Table 4-15** Number of models, speakers and trails to be processed

It is also interesting to note that the core-core task is further divided into different trial subsets as reflected in the evaluation plan and reproduced below:

1. All trials involve interview speech with matched microphones for train and test.
2. All trials involve interview speech with unmatched microphones for train and test.
3. All trials involve interview training speech and normal vocal effort conversational telephone test speech.
4. All trials involve interview training speech and normal vocal effort conversational telephone test speech recorded over a room microphone channel.
5. All different number trials involve normal vocal effort conversational telephone speech in training and test.
6. All Telephone channel trials involve normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test.
7. All Room microphone channel trials involve normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test.

8. All Telephone channel trials involve normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test.
9. All room microphone channel trials involve normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test

The number of models that must be generated, along with the number of different speakers for each of the training sub-conditions and the number of trials that must be performed on each of the test sub-conditions are reflected in Table 4-16.

| Condition | Gender | Number of models to train | Number of different speakers | Number of trials (target/non-target) |
|---|---|---|---|---|
| 1 | Male | 990 | 196 | 989/28114 |
| | Female | 1169 | 230 | 1163/32598 |
| 2 | Male | 990 | 196 | 3463/98282 |
| | Female | 1169 | 230 | 4072/114025 |
| 3 | Male | 759 | 177 | 839/26185 |
| | Female | 877 | 211 | 801/30254 |
| 4 | Male | 731 | 171 | 1225/39166 |
| | Female | 789 | 206 | 1141/44370 |
| 5 | Male | 290 | 174 | 355/13746 |
| | Female | 290 | 202 | 357/15958 |
| 6 | Male | 181 | 147 | 178/12825 |
| | Female | 184 | 177 | 183/15486 |
| 7 | Male | 180 | 146 | 179/12786 |
| | Female | 180 | 173 | 180/15211 |
| 8 | Male | 119 | 114 | 119/10997 |
| | Female | 181 | 179 | 179/17309 |
| 9 | Male | 117 | 112 | 117/10697 |
| | Female | 176 | 174 | 173/16533 |

**Table 4-16** Number of models, speakers and trails to be process in the different conditions of core task

In the NIST SRE series, the performance of the systems is measured in two different ways: using the detection cost function and Detection Error Trade-off (DET) curves. Based on the hard decision made for each trial, the detection cost function can be defined as a weighted sum of miss and false alarm error probabilities:

$$C_{DET} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{False\ Alarm} \times P_{False\ Alarm|\ NonTarget} \times \left(1 - P_{Target}\right) \quad \textbf{Eq. (4-1)}$$

According to the evaluation plan, the parameters of the cost function are: $C_{Miss}$=10, $C_{FalseAlarm}$=1, and $P_{Target}$=0.01.

As previously stated, not only a hard decision but a score is also generated for each trial. These scores are used to produce DET curves that show how false rejection or misses may be trade off against false alarms.

However, to keep consistency with previous scenarios, we are going to work with *EER* quality measure instead of $C_{DET}$.

Before moving into the evaluation set, it is worthy to provide some notes on the systems presented to the NIST SRE 2010, and the results achieved by them.

- System description: As previously reported, up to 58 sites have participated in the NIST SRE 2010. Due to the division of the evaluation in different tasks, it is difficult to determine which of the systems presented is the one achieving the best overall results, therefore, in this section we are going to focus just on the systems that have obtained the most successful rates on different conditions of the core-core task, namely in alphabetical order: ABC, I4U, ILFY, LPT, SRI and SVIST. One thing that is common to all these systems is the fact that they result from the fusion of multiple subsystems, each using their own set of features and their own classifiers. Brief notes on the types of features and classifiers used by these systems are presented in Table 4-18, according to the system descriptions presented at NIST SRE 2010 Workshop. It should be pointed out that not all the sites provide specific information of the type of features used with each classifier, and that sometimes the same set of features have been used as a common front-end for different classifiers. This situation is highlighted in red in Table 4-18.

  Glossary of terms in Table 4-18:

  - LVCSR → Large-Vocabulary Continuous Speech Recognition
  - PLDA → Probabilistic Linear Discriminant Analysis
  - PLP → Perceptual Linear Prediction
  - SCM-SCF → Spectral Centroid Magnitude – Spectral Centroid Frequency
  - SWLP → Stabilised Weighted Linear Prediction
  - GMM-UBM-JFA → Joint Factor Analysis built on top of a the classical GMM-UBM approach
  - GMM-SVM-KL → GMM-SVM with KL(Kullback-Lieber) kernel
  - GMM-SVM-BHATT → GMM-SVM with Bhattacharyya kernel
  - GMM-SVM-FT → Feature Transform parameters vectorised and compared via SVM.

- Results: The recognition rates (in terms of *EER* or $C_{DET}$) achieved in the different tasks vary significantly among different systems submitted to the assessment, and especially between first-time attending sites and sites with long history in the evaluation editions. Lowest error rates, in terms of *EER*, set the state-of-the-art in Speaker Recognition, in the ranges shown on Table 4-17, according to the different tasks proposed in the evaluation. Since the core-core task is divided in 9 different conditions, it seems appropriate to provide *EER*'s independently for each condition. However, these values must be interpreted taking into account that no gender distinction have been procured.

| Trail Condition | *EER* range |
|---|---|
| Condition 1 | 1.0% - 2.0% |
| Condition 2 | 2.0% - 4.0% |
| Condition 3 | 1.5% - 2.0% |
| Condition 4 | 1.5% - 2.5% |
| Condition 5 | 1.5% - 2.5% |
| Condition 6 | 2.0% - 5.0% |
| Condition 7 | 4.0% - 5.0% |
| Condition 8 | 0.5% - 1.0% |
| Condition 9 | 1.0% - 2.0% |

**Table 4-17** *EER* ranges for the different tasks on NIST SRE 2010

The following sequence of DET curves shows the results achieved by all the systems in the core-core task for the 9 different conditions.

| ABC | | I4U | | IFLY | | LPT | | SRI | | SVIST | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Features** | **Classifiers** | **Features** | **Classifiers** | **Features** | **Classifiers** | **Features** | **Classifiers** | **Features** | **Classifiers** | **Features** | **Classifiers** |
| MFCC | UBM-JFA | LPCC | GMM-UBM-JFA | PLP | GMM-UBM-JFA | MFCC | GMM-UBM-JFA | Cepstrum | GMM-UBM-JFA | PLP | GMM-UBM-JFA |
| MFCC | *i*-vectors | PLP | GMM-SVM-KL | LPCC+PLP | GMM-UBM-JFA | PLP | GMM-UBM-JFA | Constrained Cepstrum | GMM-UBM-JFA | LPCC | *i*-vectors |
| MFCC | *i*-vectors | MFCC | GMM-SVM-BHATT | LPCC+PLP | GMM-SVM-NAP | MFCC | *i*-vectors | Cepstrum | GMM-UBM-JFA, class dependent | MFCC | GMM-SVM-NAP |
| Not available | *i*-vector LVCSR | SCM-SCF | GMM-SVM-FT | | | PLP | *i*-vectors | PLP-SAT Cepstrum | GMM-UBM-JFA, class dependent | | |
| Not available | PLDA *i*-vectors | SWLP | | | | | | MLLR transform from English ASR | SVM | | |
| Prosodic (temporal trajectories of pitch and energy) | UBM-JFA | | | | | | | Word Ngrams | SVM | | |
| PLP | SVM CMLLR-MLLR | | | | | | | Prosodic | GMM-JFA+ Score-level combination | | |
| MFCC | *i*-vectors | | | | | | | | | | |

**Table 4-18** Features and Classifiers used by the top 6 systems

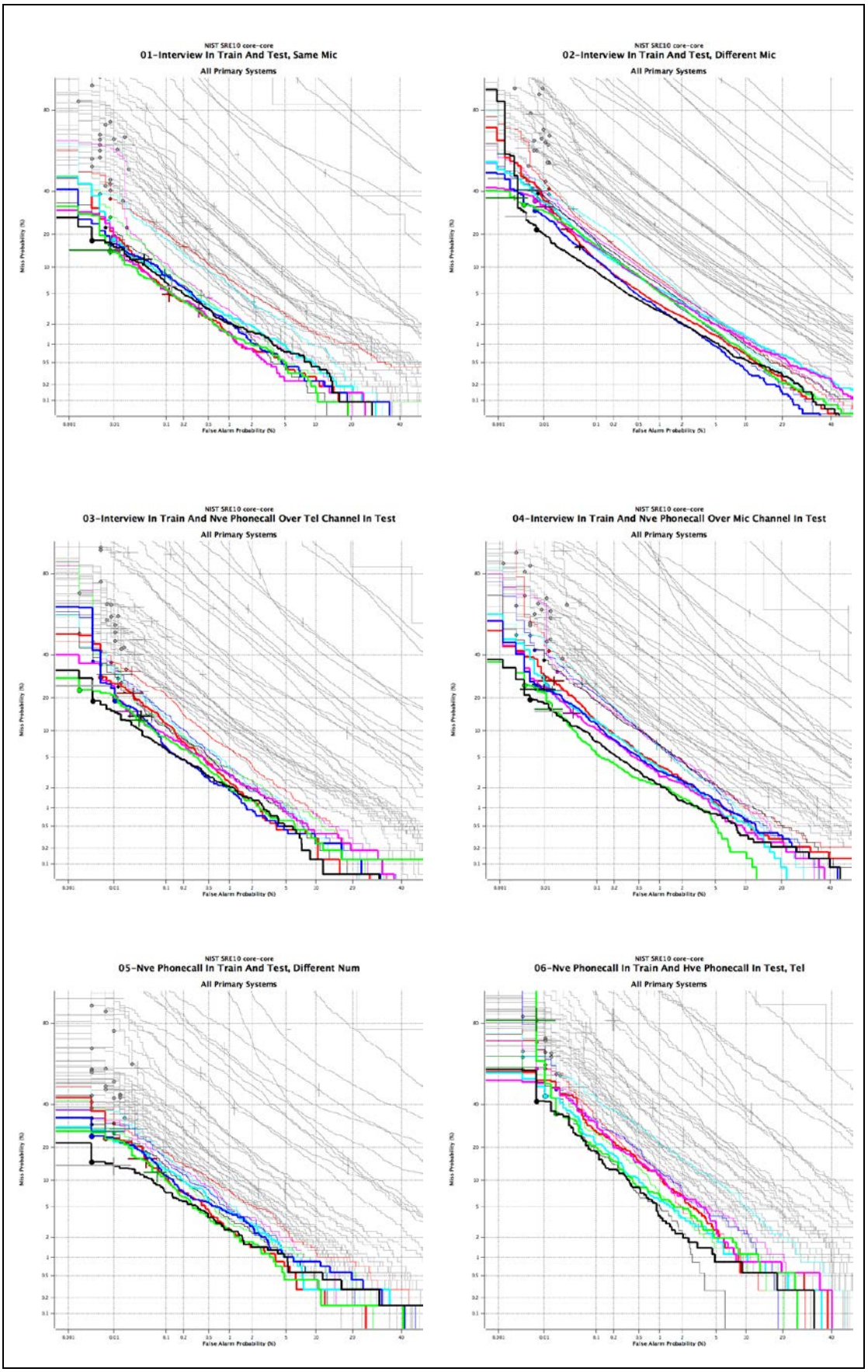



**Figure 4-4** Set of DET curves obtained for the different conditions (1 to 6) of the core-core task (provided by NIST)

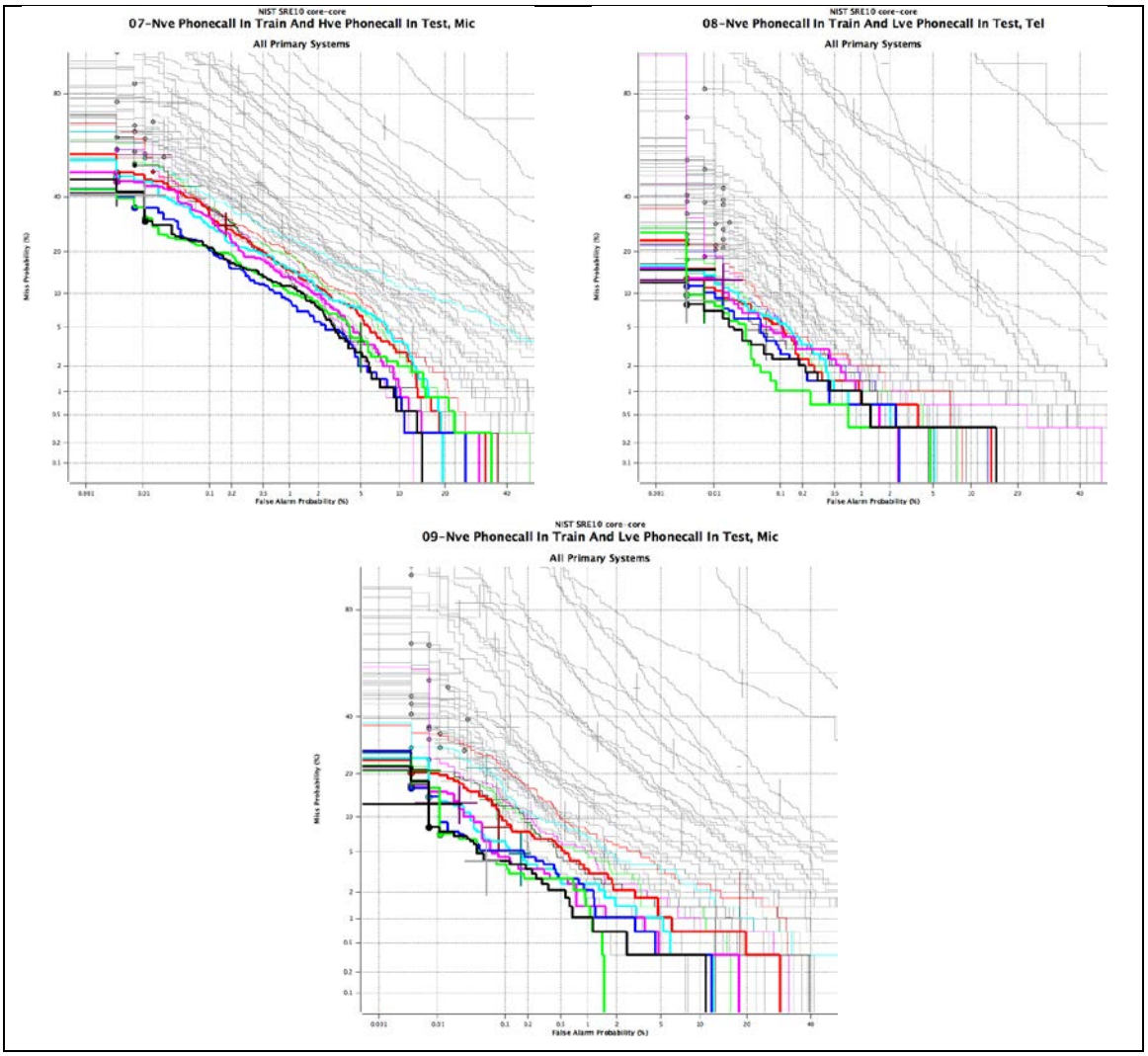



**Figure 4-5** Set of DET curves obtained for the different conditions (7 to 9) of the core-core task (provided by NIST)

- Evaluation set: As previously presented, we are going to analyse the performance of our system using the core-core task defined in the NIST SRE 2012 plan. As discussed above, there are some similarities between the 2012 and 2010 SRE plans, but also some differences that are worth noting.

  Like in previous evaluations, the task is to determine whether a specific target speaker is speaking during a given segment of speech, in the context of conversational speech over multiple types of channels. However, in this case, one or more samples of speech data, coming from multiple channels (microphone and telephone recordings), are available for training/creating the speaker's model. Similar to SRE 2010, all of the speech in SRE 2012 is expected to be in English, though English may not be the first language of some of the speakers included.

  To evaluate the performance of the system, a sequence of trials, where each trial consists of a target speaker, defined by the training data provided, and a test segment, is presented to the system. For each trial, the system must provide an output score, in form of an LLR, representing the system's confidence that the target speaker is speaking in the test segment. No detection decision needs to be provided. It must be noted that no cross-gender trials are performed.

An additional difference respect to previous SREs relies in the fact that knowledge of all targets is allowed to compute each trial's detection score.

Regarding the type of data present as test samples in the trials, the systems must deal with one two-channel excerpt from a telephone conversation or interview, containing nominally between 20 and 160 seconds of target speaker speech. Moreover, some of these test segments will have additive noise imposed. According to the NIST SRE 2012 plan, the core-core task can be divided into five different subsets:

1. All trials involving multiple segment training and interview speech in test without added noise in test
2. All trials involving multiple segment training and phone call speech in test without added noise in test.
3. All trials involving multiple segment training and interview speech with added noise in test.
4. All trials involving multiple segment training and phone call speech with added noise in test.
5. All trials involving multiple segment training and phone call speech intentionally collected in a noisy environment in test.

The subsets of core-core task differ significantly from those presented in the SRE 2010 plan. In first place, due to the fact of allowing multiple recordings, from multiple channels, under multiple environment conditions to create each target's model. Secondly due to the fact of introducing artificial noise in the test recordings (see subsets 3 and 4).

Table 4-19 summarises the most relevant information regarding the number of speakers and files involved in each trial subset.

| Trial Condition | Gender | # Different Speakers | # Train Files | | | # Test Files | # Trials | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Total | Min per Speaker | Max per Speaker | | Target | K_NT | U_NT |
| CC-0 | Male | 53 | 550 | 1 | 61 | 26511 | 8388 | 93796 | 89519 |
| | Female | 95 | 1380 | 1 | 62 | 37502 | 12231 | 245274 | 126321 |
| CC-1 | Male | 13 | 510 | 22 | 61 | 8849 | 719 | 8055 | 14691 |
| | Female | 35 | 1320 | 2 | 62 | 14013 | 2178 | 38546 | 47180 |
| CC-2 | Male | 723 | 16383 | 1 | 182 | 2588 | 1296 | 70621 | 29787 |
| | Female | 1095 | 22949 | 1 | 143 | 4045 | 2175 | 164888 | 73003 |
| CC-3 | Male | 13 | 510 | 22 | 61 | 7048 | 960 | 9136 | 7480 |
| | Female | 35 | 1320 | 2 | 62 | 10522 | 2891 | 39896 | 12568 |
| CC-4 | Male | 723 | 16383 | 1 | 182 | 3240 | 2775 | 122625 | - |
| | Female | 1095 | 22949 | 1 | 143 | 4518 | 4401 | 289218 | 4872 |
| CC-5 | Male | 723 | 16383 | 1 | 182 | 1927 | 1534 | 61311 | - |
| | Female | 1095 | 22949 | 1 | 143 | 2837 | 2349 | 148221 | 2406 |

**Table 4-19** Description of the contents of the different subsets of the NIST SRE 2012 core-core task

Where K-NT refers to a non-target trial where the test segment belongs to a known speaker, i.e. a speaker present in the target subset; and U_NT, refers to a non-target trial where no previous knowledge of the speaker is available. Additionally, it must be noted that, in the core-core task, there are 575529 trials that are not registered under one of the specified subsets, labelled in Table 4-19 as CC-0.

Finally, it must be noted that according to the evaluation plan the performance measure of the systems is the detection cost, computed slightly different if compared to 2010. However, we will keep on working with *EER* as the performance measure for consistency with previous tests.

# 5 APPLICATION TO SPEAKER RECOGNITION

In this chapter we will present both the systems we have specifically developed to face the experiments outlined in the previous chapter, as well as the most significant results that we have obtained.

Regarding the description of the recognition system, we must keep in mind two basic aspects. On one hand, the fact that the main objective of the present study is to probe that a gender-dependent extended-biometric (GDEB) characterisation of speakers can improve the performance of speaker recognition systems; on the other hand, we must take into account that, as stated before, a straight comparison of two systems results is not possible if it is not assured that both of them operate under the same conditions (which as previously established it is not the case in NIST nor ICB2013 evaluations). For these reasons, we have considered it necessary to develop not only a speaker recognition system based on GDEB characterisation but also a system that meets the state-of-the-art specifications regarding the front-end, which we call *Baseline*. Thus we can directly compare the two systems since we can guarantee that both of them are subject to the same conditions during the experiments.

In order to check whether the improvements that can be achieved through the use of the new parameterisation are consistent with any type of classifier, we found it interesting to implement different classification algorithms. Specifically, three different systems have been implemented based on the GMM-UBM paradigm, the GMM supervectors with dimensionality reduction and the *i*-vector approach. So actually, we have developed three different speaker recognition systems that can be fed with two different front-end systems. Exposing the systems to the same test bench, and by direct comparison of recognition rates, we can check whether the new speaker characterisation introduces an improvement or not.

Finally we will present the results obtained in the experiments previously described using the set of speaker recognition systems that will be presented next. The only experiment in which all systems have been used is in the case of NIST related tests, as in other cases, either because of limitations in the number of speakers or because of the absence of channel variability; we have decided that only the use of the GMM-UBM paradigm is relevant.

## 5.1 SYSTEM DESCRIPTION

### 5.1.1 Baseline front-end

We apply the *Baseline* front-end tag to the system providing classical parameterisation of voice, i.e. characterisation parameters extracted from the power spectral density (PSD) of speech as whole. As it has been shown most systems submitted to the NIST SREs, MOBIO evaluation as well as most commercial speaker recognition systems still use this kind of front-end. The PSD is estimated following standard methods such as FFT or LPC evaluated over overlapped sliding windows of the evaluated frame. Then the PSD is parameterised in order to obtain Mel-Frequency Cepstral Coefficient (MFCC) patterns and their temporal derivatives (ΔMFCC and ΔΔMFCC), which aligned in streams will serve as templates in the classification process. Additionally, different techniques can be applied in order to reduce the distortion caused by transmission channel and to make them robust to noise, such as Cepstral Mean

Subtraction (CMS), RASTA filtering, short-time Gaussianisation or/and Feature Warping. In what follows a more detailed description of the feature extraction process and the algorithms applied to reduce noise and channel distortion effects will be presented.

Figure 5-1 provides a graphical description of the process followed to extract the MFCC feature vectors from a speech signal. In the first place a pre-emphasis filter is applied to the whole signal, which actually is implemented as a high-pass filter

$$H_{pre}(z) = 1 + a_{pre}z^{-1} \qquad \textbf{Eq. (5-1)}$$

With $a_{pre}$=-0.98. Then, in step 2, the signal is broken down in short frames, typically lasting 32 ms, to which a smoothing window function is applied (e.g. a Hamming window as shown in Figure 5-1 both in temporal representation and frequency response). From the windowed frame, the Fast-Fourier Transform (FFT) is applied in order to decompose the signal into its frequency components, which constitutes step 3. For practical reasons in step 4 only the magnitude of the spectrum is retained. During step 5, mel-scale filter banks (typically 20 or 24 filters) are applied to perform energy integration over neighbouring frequency bands. Due to database specifications, the bandwidth is limited to 0-4000Hz (except for the MOBIO database which is recorded at 16 kHz). Finally, we take the log of this spectral envelope and multiply each coefficient by 20 in order to obtain the spectral envelope in dB, and the inverse FFT is applied to obtain the MFCC coefficients (typically 18).

Taking into account that this front-end is going to be applied in different scenarios, i.e. using different databases which include recordings registered under different noise conditions, it is necessary to perform a noise reduction preprocessing step. In this case a variation of the Ephraim-Malah spectral subtraction algorithm in a single channel is applied [Ephraim,1985].

As it has already been said, dynamic information extracted from MFCCs should also be computed, as well as noise and channel distortion reduction techniques may be applied. For this purpose, once the set of MFCC feature vectors have been computed for the whole speech signal, Cepstral Mean Subtraction (CMS), Feature Warping (FW) and Rasta filtering algorithms are applied. The CMS algorithm, [Furui,1981], mainly consists in computing the mean of each cepstral coefficient over the length of the current utterance, then the mean value is subtracted from the original cepstral coefficient, thus removing the channel induced effects as well as any other stationary speech component. The aim of the FW process, [Pelecanos,2001], is to transform the original cepstral coefficients so that they follow a specific target distribution, for instance a normal distribution, over a window of speech frames, typically a 3-second window. It provides a set of features that are supposed to be robust to channel mismatch, additive noise and nonlinear effects attributed to handset transducers. In the case of RASTA filtering, [Hermansky,1994], it tries to remove the spectral components that change at different rates than the one present in speech, i.e. it tries to remove convolutional and additive noise.

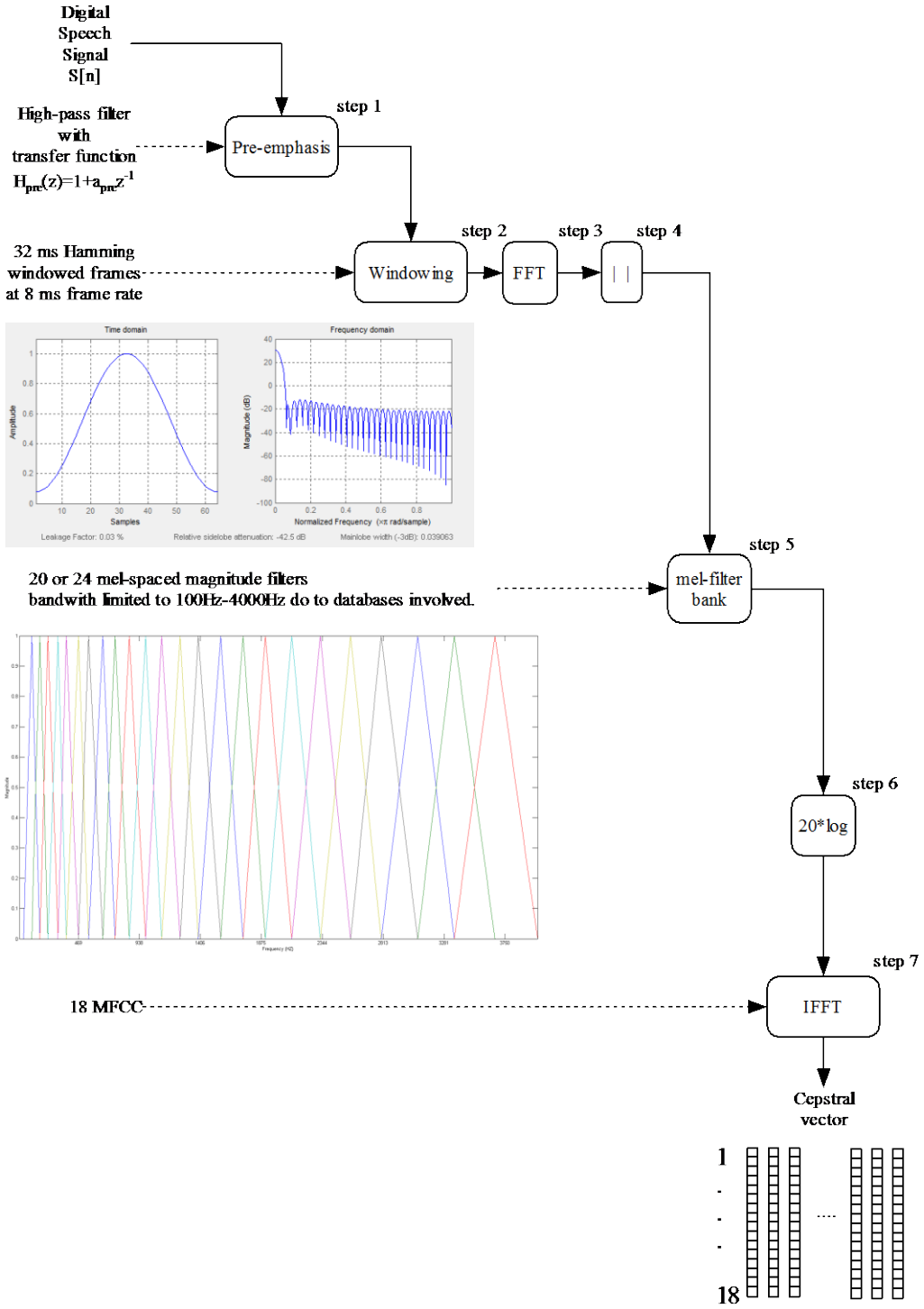



**Figure 5-1** Classical feature extraction process

Additionally, a frame selection algorithm, which is database dependent, must be applied to select only those frames containing speaker's relevant information. In the case of the experiments designed using the HESPERIA, ALBAYZIN, and MOBIO databases, where only single-channel recordings are present, an adaptive VAD algorithm based on energy detection has been implemented and computed over a 64 ms-long Blackman window with 13 ms overlap. This frame selection algorithm can also be applied in the case of NIST SRE experiments. However, in this last case a complex speaker detection algorithm can be applied based on cross-channel speaker cancelation and ASR (automatic speech recognition) availability of the given recordings.

Figure 5-2 summarises the main blocks involved in the present approach. It is worth noting that the use of this front-end does not necessarily imply that the generated feature vectors have to be gender-independent. In other words, selecting different configuration parameters in the different stages of the presented front-end depending on gender will result in a gender-dependent configuration. However, all the speaker recognition systems analysed so far use this type of front-end in a gender-independent configuration, so the same set of parameters are used regardless the gender of the speaker to model. Nevertheless, feature vectors generated by this process, contain only what we have called classical features.

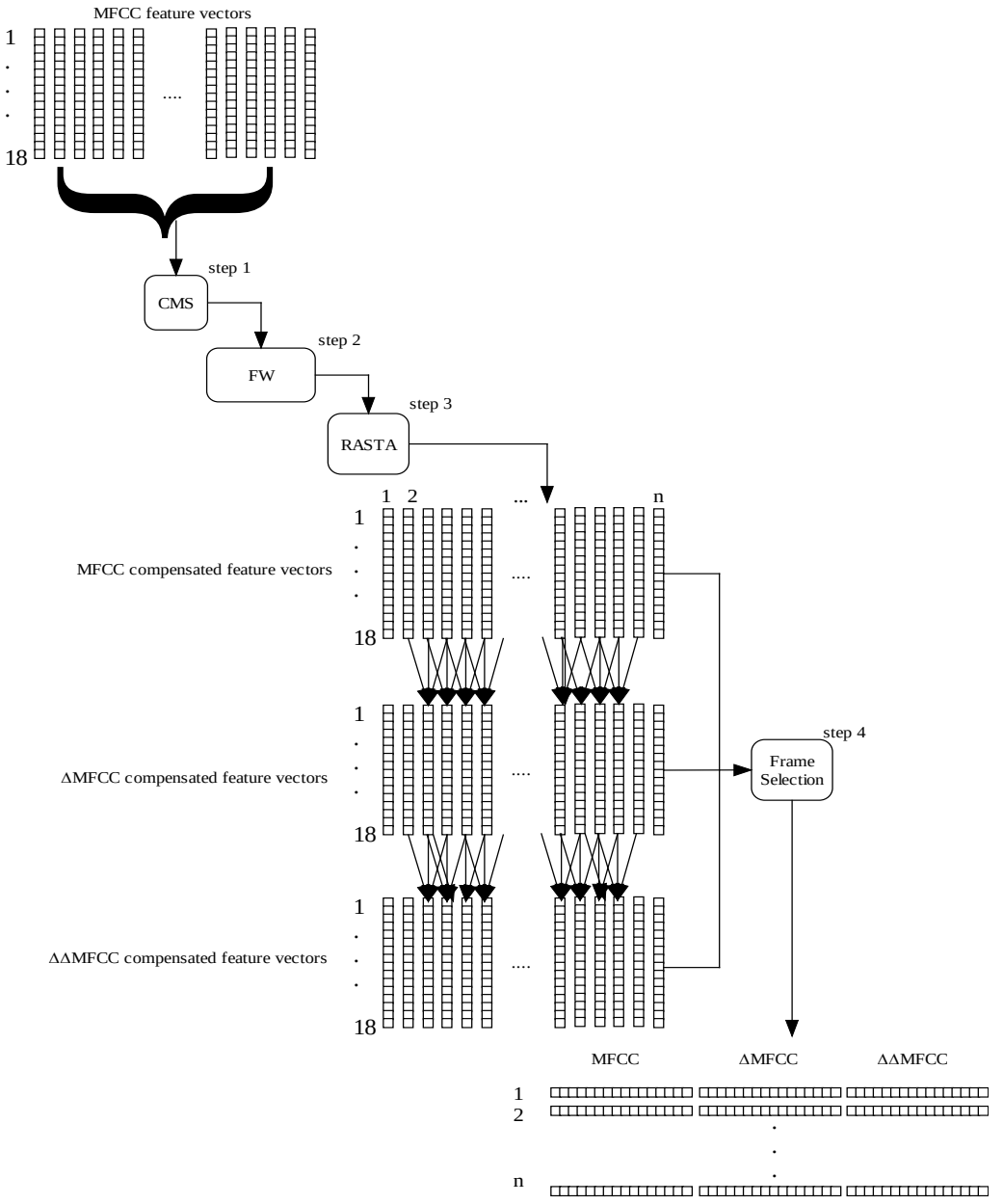



**Figure 5-2** Feature Vector Composition with compensation techniques

Additionally, alternative parameters can be added to the feature vectors such as energy of the frame (E), delta energy (ΔE), pitch (F0) or even the estimation of the formants.

Particularly, formant 3 (F3) seems to provide relevant information for speaker recognition purposes. Energy and ΔE coefficients are typically used as they constitute a heritage from the speech recognition area. The use of F0 may be justified by the fact that it takes values in different ranges depending on the gender (typically used for gender detection purposes) as well as different values depending on the speaker. Finally we propose the use of F3, which to the author's knowledge, has not been reported to be specifically used for speaker recognition. If we assume that vocal tract length is different between different speakers, and taken into account that different size vocal tracts causes the formants to move up and down on the frequency scale, and that formant F1 and F2 are classically associated to vowel detection, we can expect F3 to provide speaker related information. This concept has been previously applied for speech recognition in the form of Vocal Tract Length Normalisation [Welling,2002].

### 5.1.2 Gender-Dependent Extended-Biometric front-end

The Gender-Dependent Extended-Biometrics (GDEB) front-end can be regarded as an augmented version of the *Baseline* front-end. As previously stated the *Baseline* front-end can be configured to produce different sets of parameters depending on the gender. Taking as starting point the *Baseline* front-end, the GDEB front-end allows the user to configure different numbers of MFCCs and different numbers of filters depending on the gender, but following the same extraction process presented on section 5.1.1.

The use of the term Extended-Biometrics is based on the fact that through the use of this front-end it is possible to augment the classical features already presented, with a set of parameters obtained from the vocal tract and glottal source estimated signals extracted from the original voice signal.

The Source-Tract separation algorithm proposed in section 2.3.2 has proved its value when applied to specific high-quality text-dependent waveforms in a voice pathology detection context (especially productions of vowel *a* in which vocal tract stability is assured). In that context the algorithm provides an exact reconstruction of the glottal source from the original voice signal [Gómez,2009]. However, in the context of text-constrained or text-independent speaker recognition, the application of this algorithm is constrained by the following restrictions:

- Quality of the recordings: In voice pathology detection speech waveforms are usually recorded in controlled environments (especially hospitals) on high-quality basis. However, in speaker recognition applications this ideal situation is not always present. For instance in forensic applications, where the goal is to identify whether a questioned recording belongs to a suspected speaker, experts must face the problem of extremely different quality recordings between the questioned and the known voice. This scenario is also present in the context of NIST Evaluations, as systems must deal with recordings captured using different types of microphones, telephones (cellular, cordless, land line), different transmission channels, and even different vocal effort recordings.
- Text-independence: Although in security applications, it is usual to find text-dependent speaker recognition systems, these restrictions do not always hold, as in the case of the NIST or MOBIO evaluations. In order to apply the proposed algorithm an extra stage must be added to the system in order to detect specific voiced frames in which the glottal source should be present, thus adding more complexity to the system.

To overcome these restrictions, an alternative "*Source-Tract*" separation is proposed (see Section 5.1.2.1). The reader must note that Source and Tract have been deliberately quoted as by applying the algorithm, neither the source nor the vocal tract is extracted. Instead a glottal estimate and a vocal tract estimate are obtained that can also be used for speaker modelling purposes.

#### *5.1.2.1 Glottal Estimate – Vocal Tract Estimate Separation Algorithm*

The goal of this algorithm [Gómez,2008] is to provide a glottal estimate and a vocal tract estimate from continuous speech. It presents some common blocks with the algorithms described in Chapter 2. More precisely:

- Radiation Compensation Block: A first order prediction lattice has been implemented to compensate lip radiation effects, as explained in section 2.3.2
- Inverse Filtering Block: A k-order filtering process is applied to remove the vocal tract information from the radiation compensated speech. This process can be implemented using a k-order prediction error lattice.
- Joint-Process Estimation Block (JPE): The residual is used as the reference signal in an Adaptive Lattice-Ladder filter used for Joint-Process Estimation on the radiation-compensated speech $s_l(n)$. Through this process, a glottal estimate and a vocal tract estimate are extracted which can be considered fully uncorrelated (second-order decoupling) [Gómez,2008].

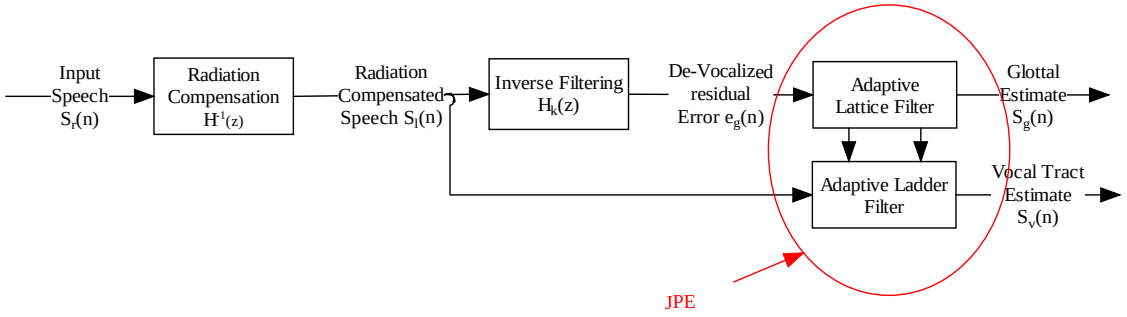



**Figure 5-3** Separation algorithm with lip radiation compensation using first order prediction lattice

#### *5.1.2.2 Algorithm's parameter tuning*

The previously described algorithm needs to be tuned according to the kind of recordings it is supposed to deal with, i.e. we need to find the best parameters in the Inverse Filtering Block that provide the best second order decoupling in the estimated signals. Different values have been tested for the two specific parameters used in the Inverse Filtering Block, namely, the order of the filter and the *forgetting factor*, which according to [Griffiths,1977] "*helps it deal better with statistical variations when operating in non-stationary environments*".

As the algorithm needs to be tuned accordingly to the kind of recordings, this means that no universal configuration can be provided, instead, when applied, it is necessary to define a development set which collects the variability of data under test and which helps in tuning the meta-parameters of the algorithms used. These meta-parameters which are usually the number of Gaussians, dimension of subspaces, etc., in our case include also the order of the filter and the *forgetting factor* of the source-tract separation algorithm as well as the number of filters in the filter bank used to compute the cepstral coefficients and the number of MFCCs for each of the 3 different signals available.

#### *5.1.2.3 Feature vector composition*

As previously stated, the main objective of the present study is to probe that a GDEB characterisation of speakers can improve the performance of speaker recognition systems. This means that the set of parameters generated by the front-end are going to be different depending on the gender. However, our proposal does not represent a complete break with the classical approach; on the contrary it can be seen as an extended version where classical MFCC parameters are augmented with MFCC parameters extracted from the vocal tract and glottal source estimates, to form the feature vector.

From this point of view Figure 5-4 shows the generic form of the feature vector generated by the GDEB front-end; which is common to both genders. Therefore, the differences between male and female feature vectors, rely on the set up of the previously described algorithms, i.e. the number of filters used to extract the MFCC for the three different signals (raw voice, glottal estimate and vocal tract estimate), the number of MFCCs and also the order of the filter and the *forgetting factor* coefficient used to get the glottal and vocal tract estimates.

Additionally, as mentioned before, we have also investigated the inclusion of other parameters frequently used in speaker recognition systems such as: frame energy, ΔE, pitch (F0), and a new one such as the third formant (F3) estimate.

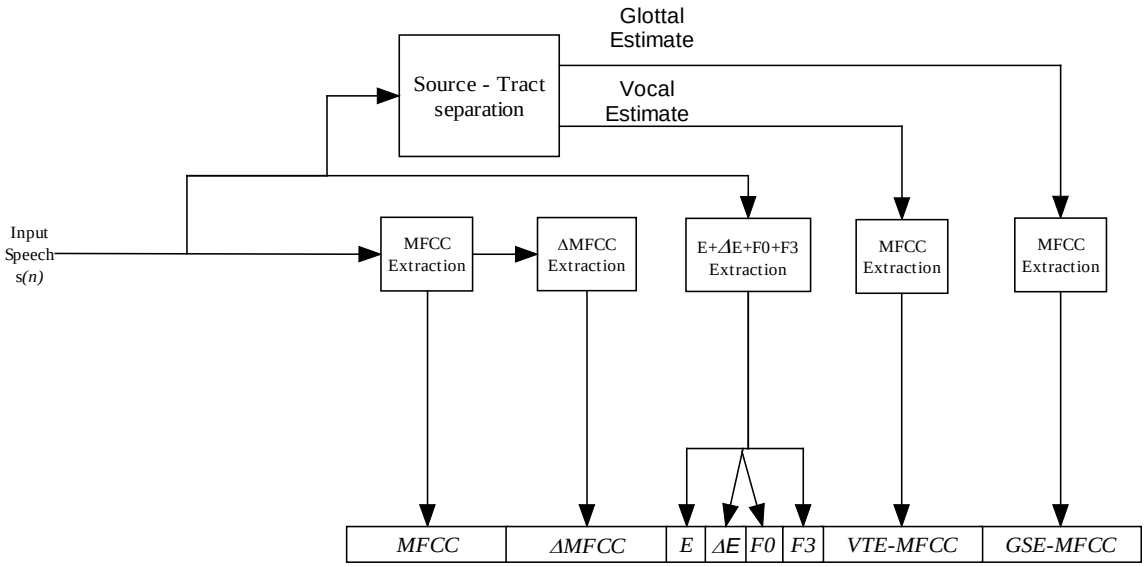



**Figure 5-4** Parameterisation scheme used for male speakers

### 5.1.3 Speaker Recognition systems

Three different classification algorithms have been implemented in, which are considered both the state of the art and new trends in speaker recognition. Specifically, these three systems are a GMM-UBM classifier, a GMM-supervector classifier with dimensionality reduction and last but not least, an *i*-vector classifier.

#### *5.1.3.1 GMM-UBM*

A block diagram of the Automatic Speaker Verification system implemented using the GMM-UBM approach is shown in Figure 5-5. In this section we do not care about the digital speech data acquisition step and the feature extraction process which have been already presented.

In the set of experiments that has been carried out, we have used a standard mixture classifier with diagonal covariance matrix. Each speaker is represented by a Gaussian Mixture Model (GMM), $\lambda_{\text{speaker-k}}$, which has been adapted from a gender-dependent Universal Background Model (UBM) using the MAP algorithm in which only the distribution means have been adapted (part *B* in Figure 5-5). The UBM is also represented as a GMM, $\lambda_{\text{UBM}}$, which has been trained from the training set via the EM-algorithm (part *A* in Figure 5-5). The number of Gaussians as well as the relevance factor used in the MAP-algorithm depend on the specific experiment carried out.

The Log-Likelihood Ratio (LLR) has been the score used to take a decision on whether a test audio-segment is likely to be spoken by a specific speaker, with claimed identity $\lambda_{\text{speaker-i}}$. In other words, the set of feature vectors extracted from a test audio segment is compared with the claimed speaker model giving a match score which measures their similarity (part *C* in Figure 5-5).

$$LLR\left(X, \lambda_{speaker-i}\right) = \log P\left(X|\lambda_{speaker-i}\right) - \log P(X|\lambda_{UBM}) \qquad \textbf{Eq. (5-2)}$$

Additionally, the decision scores can be normalised using Zero Normalisation (ZNorm) and Test Normalisation (TNorm) or by combining both of them. In the case of ZNorm, the normalised score is given by:

$$S_{ZNorm} = \frac{\log\left(P\left(X|\lambda_{speaker-i}\right)\right) - \mu_Z}{\sigma_Z} \qquad \textbf{Eq. (5-3)}$$

where $\mu_Z$ and $\sigma_Z$ are the mean and standard deviation for the impostor distribution. To estimate these values, the target speaker model is tested against utterances from impostors; this results in a set of likelihood scores from which the impostor distribution is estimated.

TNorm also uses mean and standard deviation as normalisation parameters, but in this case a set of impostors is used to estimate the log-likelihood for the test input utterance. So mean and variance is computed for these impostor scores and the normalised score is computed like in Eq. (5-3).

It must be noted that in the set of experiments defined using the databases HESPERIA, ALBAYZIN and MOBIO, the number of available speakers for normalisation purposes is quite limited. So the results obtained when applying these normalisations will be influenced by this fact.

Finally a decision is made to either accept or reject the claimant according to the match scores, and a specific threshold (part *D* in Figure 5-5).

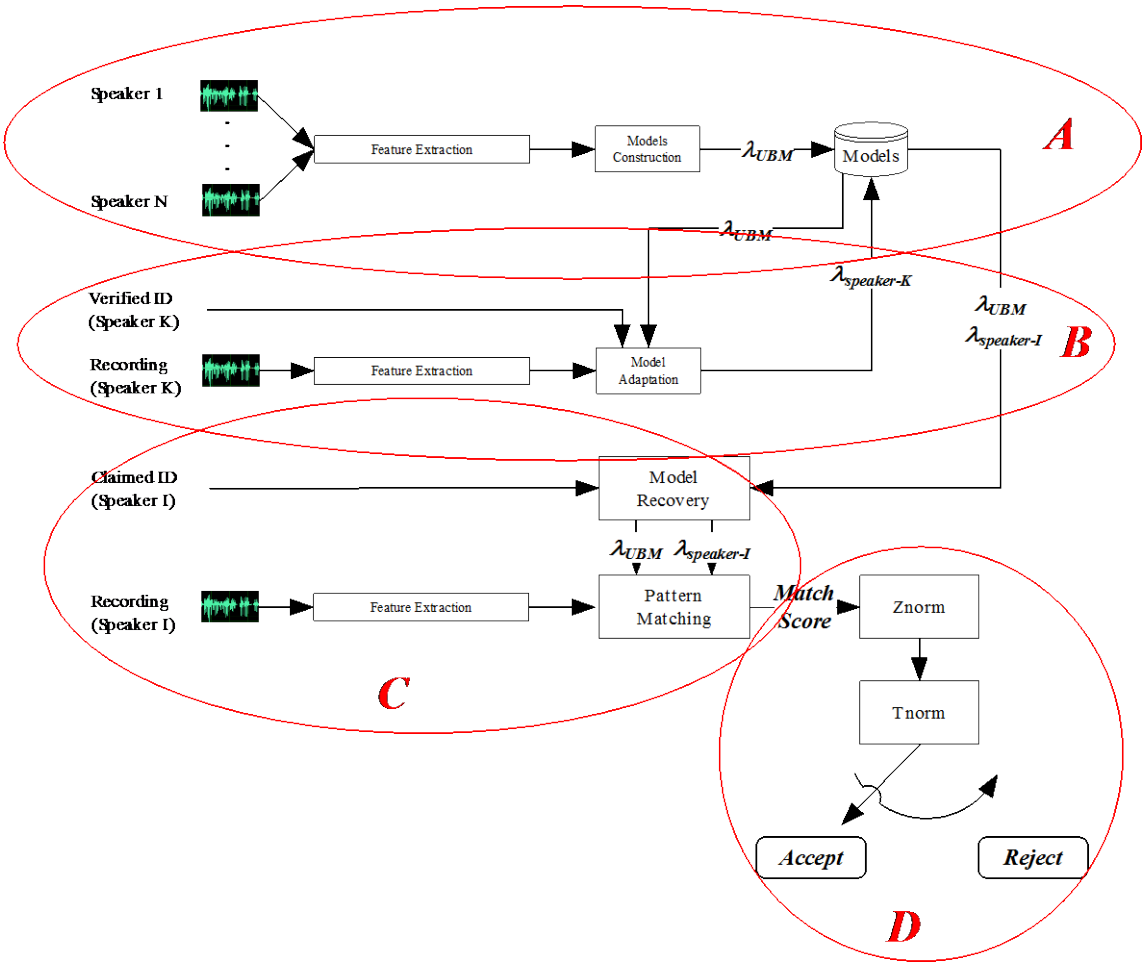



**Figure 5-5** GMM-UBM speaker verification system implemented

#### *5.1.3.2 SV-GMM*

A block diagram of the Automatic Speaker Verification system implemented using the SV-GMM approach is shown in Figure 5-6. Like in the previous section, the digital speech data acquisition step and the feature extraction process are not going to be detailed here.

In the set of experiments that has been carried out, a gender-dependent UBM has been built (part *A* in Figure 5-6) via EM-algorithm using a specific training set. From this UBM, each speaker model has been adapted (part *B* in Figure 5-6), using the MAP algorithm in which only the distribution means have been updated. The number of Gaussians as well as the relevance factor used on the MAP-algorithm depend on the specific experiment carried out. A vector to supervector mapping is then applied to transform the GMM model obtained by MAP adaptation into a supervector. In order to make the problem computationally efficient, a dimensionality reduction step has been added based on PCA (detailed in Chapter 3). For the sake of simplicity, the PCA matrix training process as well as the data used to train this matrix have been removed in Figure 5-6.

Additionally, and depending on the specific experiment to be carried out, intersession variability compensation must be applied. In our case, WCCN and LDA (both of them detailed in Chapter 3) are combined.

As the test recordings can be transformed also into a supervector by the previously explained process (part *C* in Figure 5-6), a distance between two vectors must be defined in order to make a decision on whether the test recording has been uttered by the claimed speaker (part D in Figure 5-6). Usually the Euclidean distance or the Cosine distance is used in the decision step.

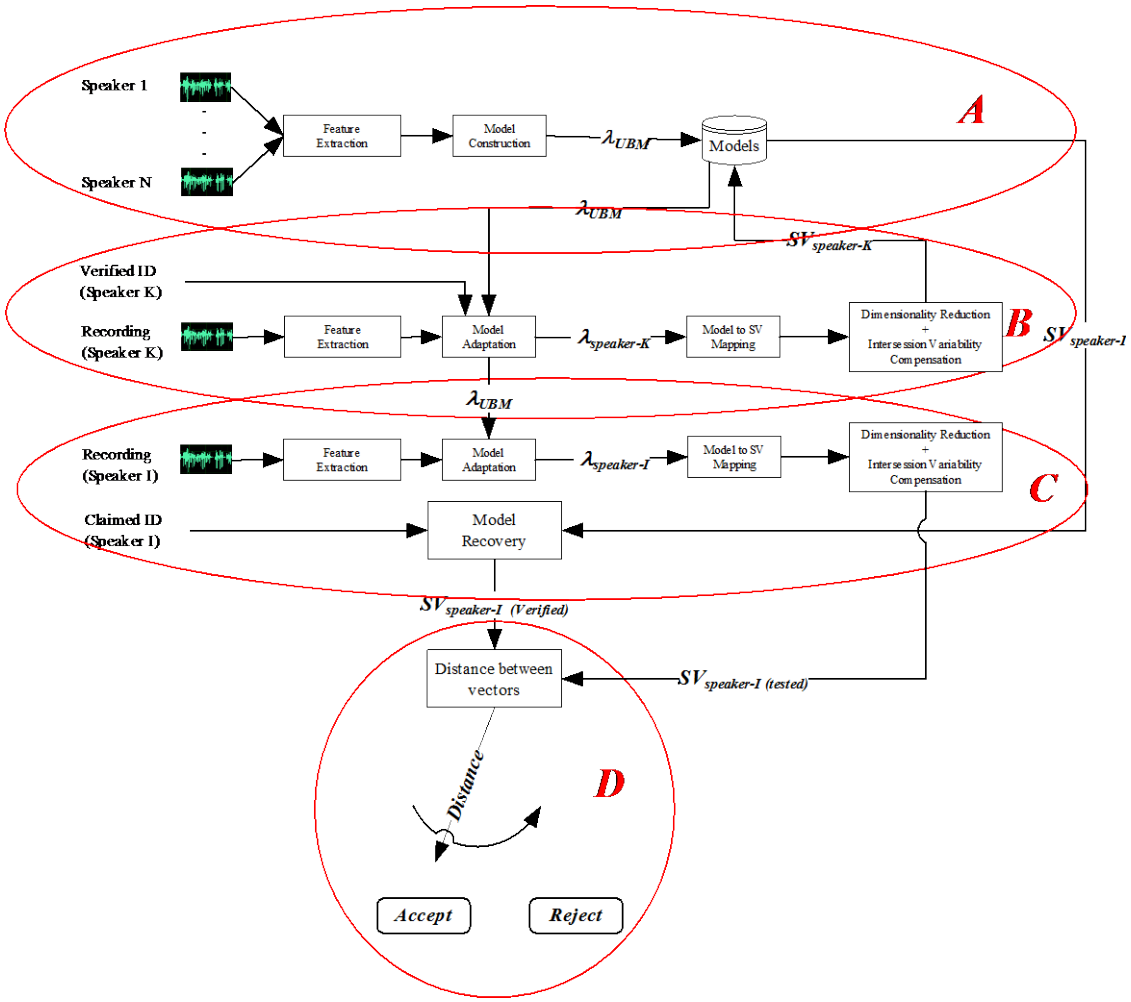



**Figure 5-6** SV-GMM speaker verification system implemented to run the NIST experiments

#### *5.1.3.3 i-vectors*

As it is shown in Figure 5-7, the Automatic Speaker Verification system implemented using the *i*-vector approach presents a strong resemblance to the SV-GMM system previously described. The main difference relies in the fact that an *i*–vector is used to represent each speaker instead of using a supervector. As in previous sections, the digital speech data acquisition step and the feature extraction process are not going to be detailed here.

Like in the previously presented systems, it is necessary to build a gender-dependent UBM, which is going to be used to build the speaker's specific model, in this case represented as an *i*-vector (part *A* in Figure 5-7). A total variability matrix is applied to transform the feature vectors extracted from the training utterance into a set of speaker factors in a low-dimensional total variability space (part *B* in Figure 5-7). For the sake of simplicity, the matrix T training process as well as the data used to train this matrix

have been removed from Figure 5-7. The same process is applied to obtain the *i*-vector from the test utterance (part *C* in Figure 5-7).

Again, WCCN and LDA are combined to perform inter-session variability compensation before scoring.

The cosine kernel distance between the *i*-vector of an enrolment utterance and the *i*-vector of a test segment is compared to a decision threshold $\theta$, to accept or reject a given trial. As no target model is required, and *i*-vectors are smaller in size, the decision process is faster and less computationally demanding compared to other methods. (Part *D* in Figure 5-7)

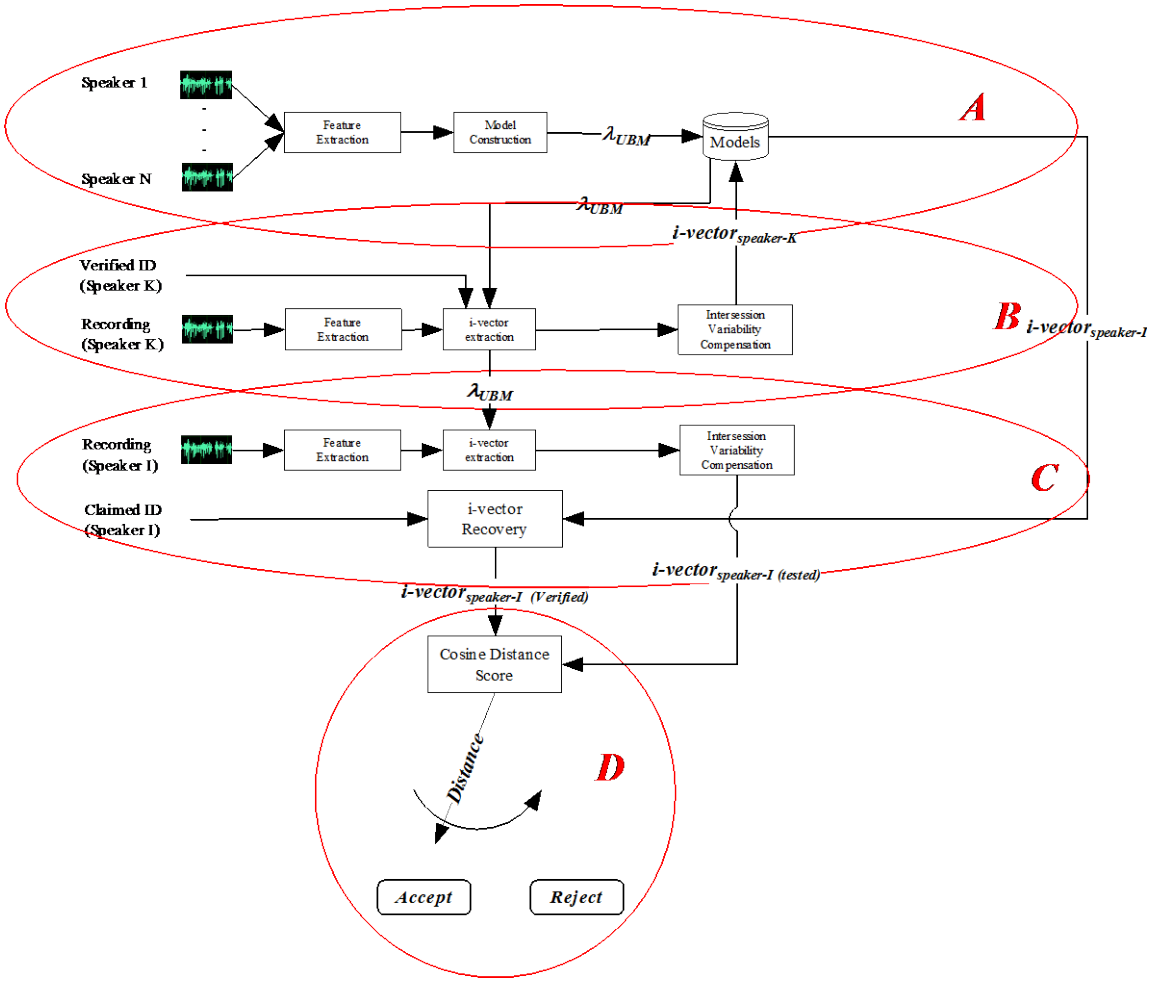



**Figure 5-7** *i*-vector speaker verification system implemented to run the NIST experiments

## 5.2 RESULTS

In this section we are going to present the results obtained for each of the experiments described in Chapter 4. It must be noted that the purpose of the experiments carried out is to show that an accurate characterisation based on GDEB characteristics will provide better recognition rates than classical approaches based on gender-independent speaker characterisation.

In almost all the experiments that we have processed, we have followed the same approach. Specifically, we begin by performing a search over classical parameters (i.e. the ones extracted using the baseline front-end), in order to get the best performance (in terms of recognition rates, mainly *EER*) taking into account that we assume a gender-independent characterisation. Simultaneously, we will try to throw light on some myths about the use of these parameters. Specifically, we will show that introducing the ΔΔ coefficients in the feature vectors used to characterise the speaker does not usually provide any additional advantage. Actually the use of these parameters is derived from the fact that many researchers currently working in speaker recognition came from the voice recognition area, in which obviously these parameters contain relevant information. In other words, ΔΔ coefficients incorporate information more related to the transmitted message than to the speaker.

Once we have found a baseline to beat, the next step consists in introducing what we have called the extended-biometric parameters into the gender-dependent feature vector in order to verify that they provide an accurate characterisation of speaker, and thus better recognition rates.

Finally, we should point out that despite having presented three different classification systems: GMM-UBM, SV-GMM and *i*-vectors; for the set of experiments designed using the databases HESPERIA, ALBAYZIN and MOBIO only the GMM-UBM approach is going to be used. The main reason for this choice is that in order to use the other two classifiers it is necessary to have a high volume of data to train the T matrix and the PCA matrix.

### 5.2.1 Text-Constrained Speaker Recognition

In this section we will present the results obtained in the two defined scenarios. The aim of this section is manifold. First of all, we need to determine whether a gender-dependent characterisation of speakers provides some improvement, in terms of recognition rates, respect to the use of a gender-independent characterisation (regarding the number of MFCC and filters used to compute them). At the same time we will use this set of experiments to analyse the usefulness of certain extra parameters frequently used in speaker recognition systems such as: frame energy, ΔE or pitch (F0), and a new one, namely, the third formant estimate (F3). We also want to verify whether the use of Δ and ΔΔ coefficients is justified in the speaker recognition area. Last but not least, we need to verify whether the extended-biometric parameters proposed are useful for speaker characterisation purposes, and thus are able to improve speaker-recognition system rates.

For each of the proposed scenarios, we will present the results obtained using both the *Baseline* and the GDEB front-ends, when applied to the GMM-UBM approach. Neither the SV-GMM nor the *i*-vector approaches will be used on these tests, due to the limited amount of data available for training the corresponding matrices. Additionally, different score normalisation techniques will be applied (i.e. TNorm, ZNorm and ZTNorm) in order to analyse the effect of such proposals on the results obtained.

In order to evaluate the performance of the systems, regardless the scenario, we will use the Equal Error Rate (*EER*) quality measure, and the Half Total Error Rate (*HTER*) which can be defined from a score threshold, $\theta_{dev}$, obtained from the development set as follows:

$$\theta_{dev} = {}^{argmin}_{\theta} |FAR_{dev}(\theta) - FRR_{dev}(\theta)| \qquad \textbf{Eq. (5-4)}$$

Where *FAR*, is the False Acceptance Rate and *FRR*, is the False Rejection Rate. This score threshold, $\theta_{dev}$, provides the *EER* operating point for the system on development:

$$EER = \frac{FAR_{dev}(\theta_{dev}) + FRR_{dev}(\theta_{dev})}{2} \qquad \textbf{Eq. (5-5)}$$

This threshold is then used on the evaluation data set to obtain the *HTER* that can be defined as:

$$HTER = \frac{FAR_{eval}(\theta_{dev}) + FRR_{eval}(\theta_{dev})}{2} \qquad \textbf{Eq. (5-6)}$$

Based on these metrics, we proposed a battery of tests using the *Baseline* front-end in order to minimise the *EER*. However, as no cross-gender trials are going to be present, we can define a new metric which we call Half Equal Error Rate, *HEER*:

$$HEER = \frac{EER_M(MFCC, F, \Delta, \Delta\Delta, G, \alpha) + EER_F(MFCC, F, \Delta, \Delta\Delta, G, \alpha)}{2} \qquad \textbf{Eq. (5-7)}$$

Where *MFCC*={12,14,16,18,19,20,21,22,23,24,25,26} denotes the number of MFCC coefficients computed, *F*={24,28,30,34,38,40,44,48,50} refers to the number of filters on the filter bank used to compute the MFCCs, Δ={true/false} and ΔΔ={true/false} refers to the whether the Δ (delta) and ΔΔ (double-delta) coefficients are present on the feature vector, *G*={256,512} refers to number of Gaussians used to build both the UBM and the speaker's models; finally $\alpha$={5,8,10,16,20} represents the relevance factor used to adapt the speaker's model from the UBM using the MAP algorithm. The admitted values for each parameter, in brackets, that have been tested are typical values used for the kind of problem we are facing. In Eq. (5-7), we distinguish between *EER* for male ($EER_M$) and female speakers ($EER_F$). Obviously, when using a gender-independent parameterisation, it is necessary to reach a compromise between both *EER* in order to minimise *HEER*. However, when using a gender-dependent parameterisation, this compromise disappears and the objective is to minimise $EER_M$ and $EER_F$ independently.

It is worth noting that the threshold used will be different depending on the gender, since no normalisation has been carried out on the scores in order to obtain a universal threshold.

Additionally, the error rates are going to be represented using the Detection Error Trade-off (DET) curves, where the *FRR* is plotted against the *FAR*. The DET curves can be used to evaluate the calibration of the verification system.

#### *5.2.1.1 Scenario 1 (mic-mic)*

In what follows, the results obtained on the development set will be presented. Throughout the use of the development set, we are going to tune the recognition system background parameters and meta-parameters. Starting with the *Baseline* front-end, in a gender-independent configuration, we are going to verify the usefulness of classical parameters (e.g. MFCCs, MFCCs+Δ, and MFCCs+Δ+ΔΔ), the effect of the number of MFCC coefficients and, the number of filters used on the filter bank to compute the MFCCs. This process can be followed as well in the case of using the GDEB front-end,

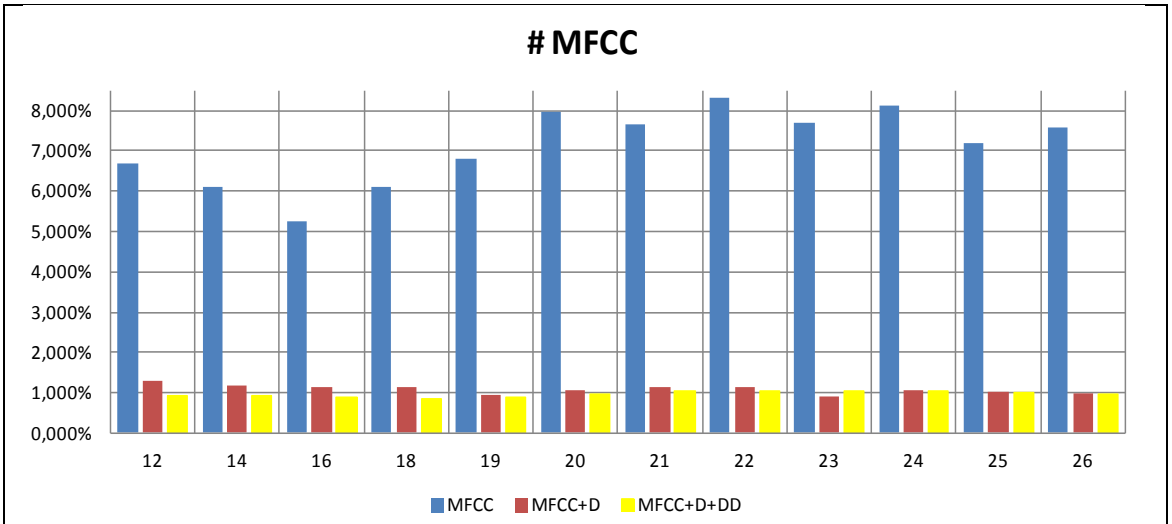
even in the case of not adding augmented parameters and just using a gender-dependent configuration. Subsequently, we will test the effect of adding what we have called extra parameters, i.e. Energy, Δ Energy, Pitch (F0) and third formant estimate (F3), both on the gender-dependent and gender-independent configurations. Last but not least, we will verify the viability of using the extended-biometric parameters extracted by the GDEB front-end, for speaker recognition purposes.

In this set of experiments, we have ruled out a detail analysis of the number of Gaussians used in each model as this number mainly depends on the number of classes to be modelled which in turn depends on the type and number of parameters used on the model. Additionally in the first approach, no score normalisation techniques have been applied, later on we will face the influence of the different score normalisation processes on the performance of the systems.

Finally, once the settings are fixed for both front-ends, and a score threshold is established on the development set, actual system performance can be checked out using the evaluation set, as we are facing the systems to meet unknown data.

The first set of figures represents the results obtained in terms of *HEER*, based on each of the configuration parameters in Eq. (5-7), assuming that we are using the *Baseline* front-end, thus a gender-independent configuration (labelled as GIC), and no score normalisation is applied.

The process followed, consists in fixing a value for a specific parameter and test, for the rest of configurable parameters which configuration provides better results in terms of *HEER*. For instance, regarding the number of MFCCs, we start by fixing its value to 12, and we test all the combinations for the remaining parameters, selecting the one providing lower *HEER*. Then we move to the next MFCC value and repeat the tests.



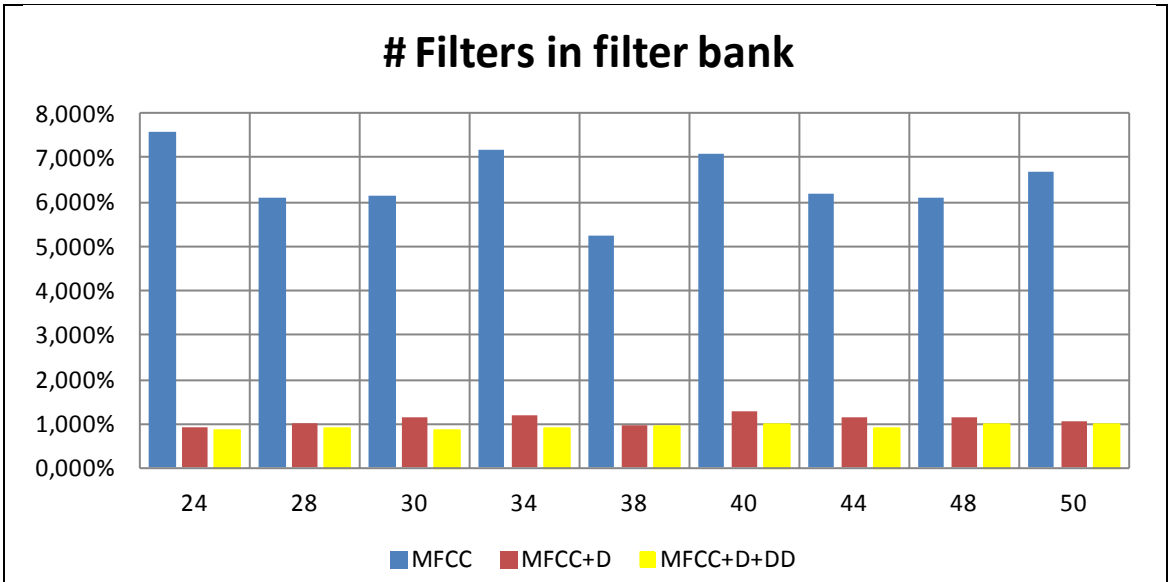

| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 6,68% | 6,11% | **5,25%** | 6,11% | 6,78% | 7,97% | 7,65% | 8,30% | 7,68% | 8,09% | 7,18% | 7,58% |
| MFCC+Δ | 1,28% | 1,17% | 1,14% | 1,15% | 0,96% | 1,05% | 1,11% | 1,13% | **0,91%** | 1,05% | 1,01% | 1,00% |
| MFCC+Δ+ΔΔ | 0,94% | 0,93% | 0,90% | **0,85%** | 0,90% | 0,96% | 1,05% | 1,06% | 1,03% | 1,06% | 1,01% | 0,95% |

**Figure 5-8** *HEER* obtained depending on the number of MFCCs and the use of Δ and ΔΔ (GIC – development set)

In Figure 5-8, we highlight the influence of the number of MFCCs, as well as the influence of the use of Δ and ΔΔ coefficients on the recognition rates. Better results

obtained for each configuration, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ are marked in bold. The first conclusion that we can draw out from it is that the use of MFCCs alone (i.e. without Δ or ΔΔ) are not accurate enough to model speakers, as recognition rates are clearly worse than the ones obtained when complementing the feature vectors with Δ and ΔΔ coefficients. In this case, the use of ΔΔ coefficients combined with Δ coefficients, systematically produces better recognition rates than just the use of MFCCs+Δ, especially for MFCC values between 12 and 20. However, the recognition rates obtained using MFCCs+Δ are also reasonably competitive, with *HEER*<1%.

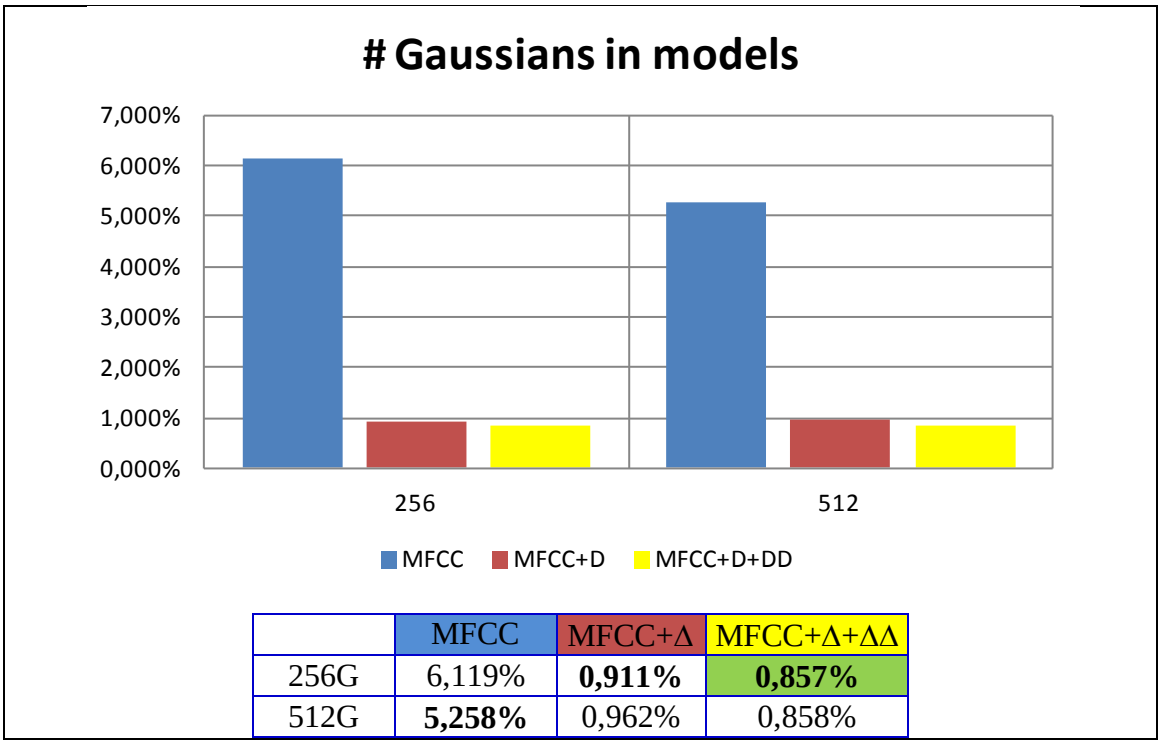
Regarding the number of filters, we operate following the same procedure already presented for the MFCC parameters. We fixed a value for *F*, and we tested all the combinations for the remaining parameters, selecting the configuration providing lower error rates in terms of *HEER*, for the specific value of *F*. Figure 5-9 provides the results obtained in terms of *HEER*, for the different values of parameter F, and for the three tested configurations, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ. In this case, the better configuration seems to be the one in which the number of filters is 24, for both MFCCs+Δ and MFCCs+Δ+ΔΔ configurations, although a slight improvement is obtained for the MFCCs+Δ+ΔΔ when 30 filters are used, but almost negligible.




| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 7,57% | 6,11% | 6,15% | 7,18% | **5,25%** | 7,08% | 6,18% | 6,11% | 6,70% |
| MFCC+Δ | **0,91%** | 0,99% | 1,14% | 1,19% | 0,96% | 1,26% | 1,15% | 1,15% | 1,05% |
| MFCC+Δ+ΔΔ | 0,85% | 0,89% | **0,85%** | 0,90% | 0,95% | 1,01% | 0,90% | 0,97% | 0,98% |

**Figure 5-9** *HEER* obtained depending on the number of filters and the use of Δ and ΔΔ (GIC – development set)

If we consider the number of Gaussians used to model the speakers in the gender-independent configuration, we can conclude (see Figure 5-10), that better results are produced when 256 Gaussians are used in the case of including Δ or ΔΔ coefficients in the feature vector, while in the case of using just MFCC alone, better results are obtained when 512 Gaussians are used.




| | MFCC | MFCC+Δ | MFCC+Δ+ΔΔ |
|---|---|---|---|
| 256G | 6,119% | **0,911%** | **0,857%** |
| 512G | **5,258%** | 0,962% | 0,858% |

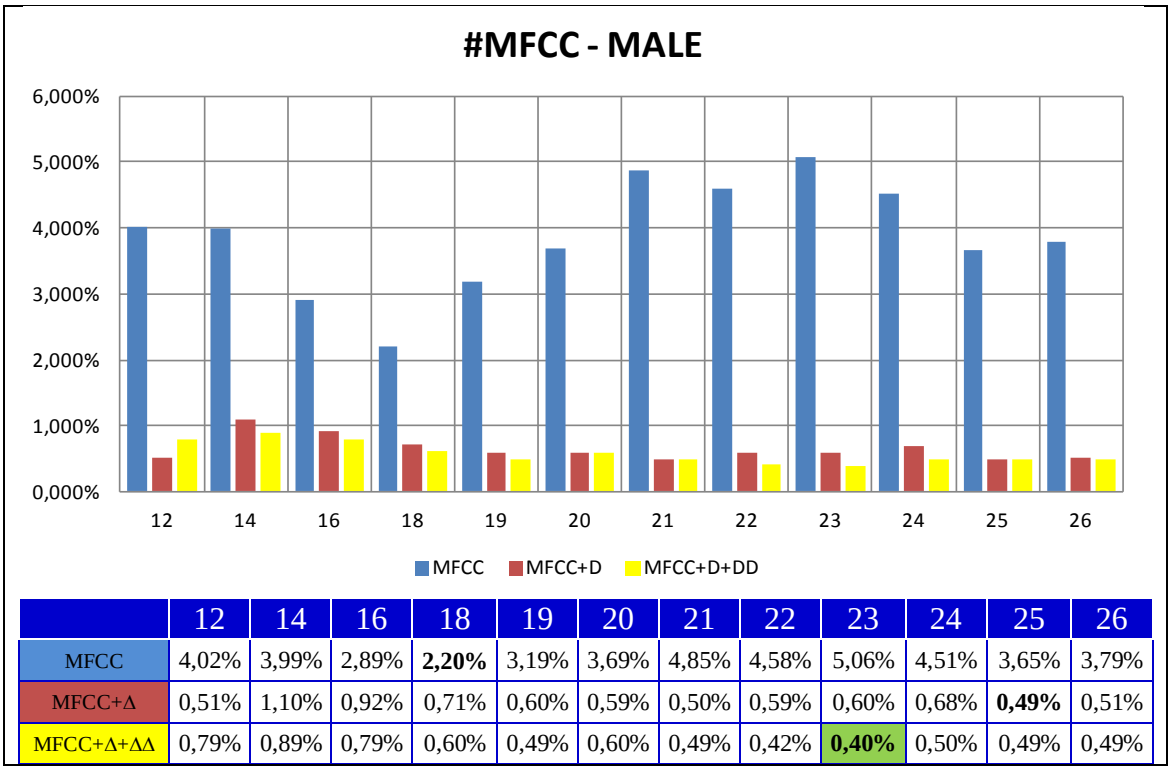
**Figure 5-10** *HEER* obtained depending on the number of Gaussians and the use of Δ and ΔΔ (GIC – development set)

The following table, presents the specific configurations providing better recognition rates in terms of *HEER*, for the *Baseline* front-end (therefore in a gender-independent configuration), for each feature vector parameter set (i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ). Moreover, it also provides the $EER_F$ and $EER_M$ obtained in each case as well as the specific score threshold.

| Parameters | *F* | *G* | *α* | $EER_M$ [$\theta_M$] | $EER_F$ [$\theta_F$] | *HEER* |
|---|---|---|---|---|---|---|
| 16MFCC **(GIC MFCC)** | 38 | 512 | 5 | 3.602% [0.254] | 6.913% [0.179] | 5.258% |
| 23MFCC+Δ **(GIC MFCC+Δ)** | 24 | 256 | 5 | 0.604% [0.324] | 1.217% [0.355] | 0.911% |
| 18MFCC+Δ+ΔΔ **(GIC MFCC+Δ+ΔΔ)** | 30 | 256 | 5 | 0.798% [0.346] | 0.916% [0.415] | 0.857% |

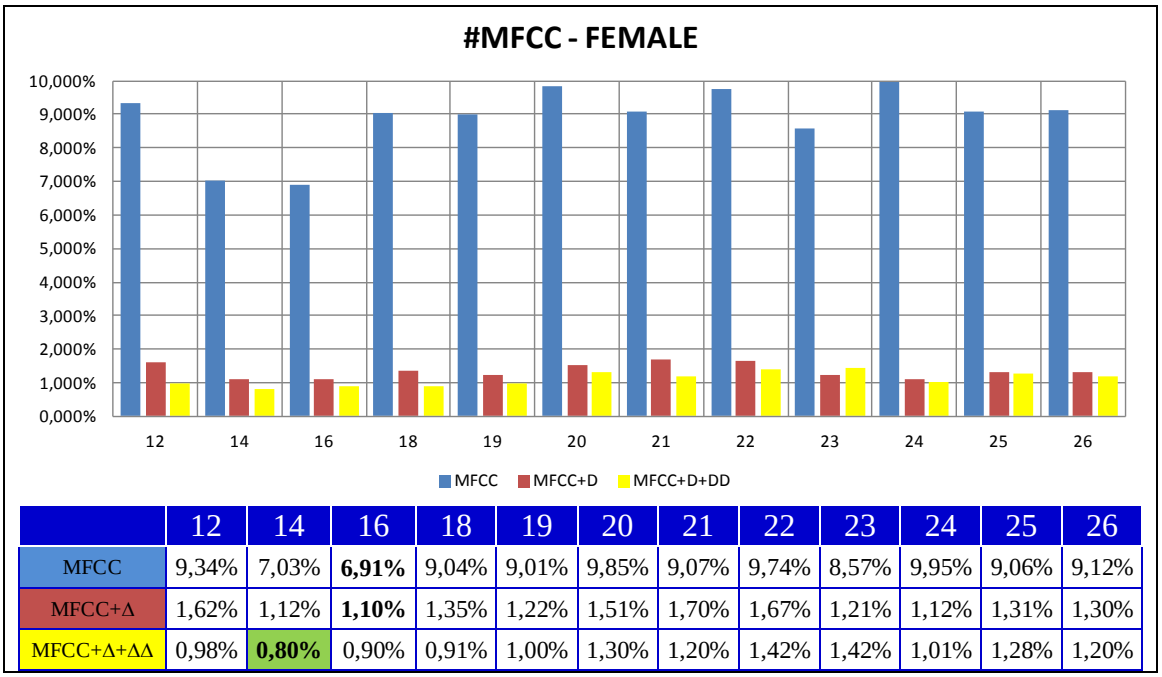
**Table 5-1** Baseline front-end based SR system providing better *HEER* in a gender-independent configuration

So far, we have used the *Baseline* front-end in a gender-independent configuration, thus the objective was to minimise *HEER*, reaching a compromise between $EER_F$ and $EER_M$. However, as we have already pointed out, a gender-dependent configuration may provide an improvement in terms of *HEER*, as $EER_F$ and $EER_M$ are independently minimised, even in the case of just using classical features (i.e. MFCC, MFCC+Δ and MFCC+Δ+ΔΔ). The same analysis performed in the case of the gender-independent configuration, has been applied for the case of the gender-dependent configuration (labelled as GDC). Figure 5-11 to Figure 5-15 present the results obtained in terms of $EER_X$, where $X = \{F,M\}$, when the effect of the number of MFCC, the number of filters and the number of Gaussians in the model is analysed.



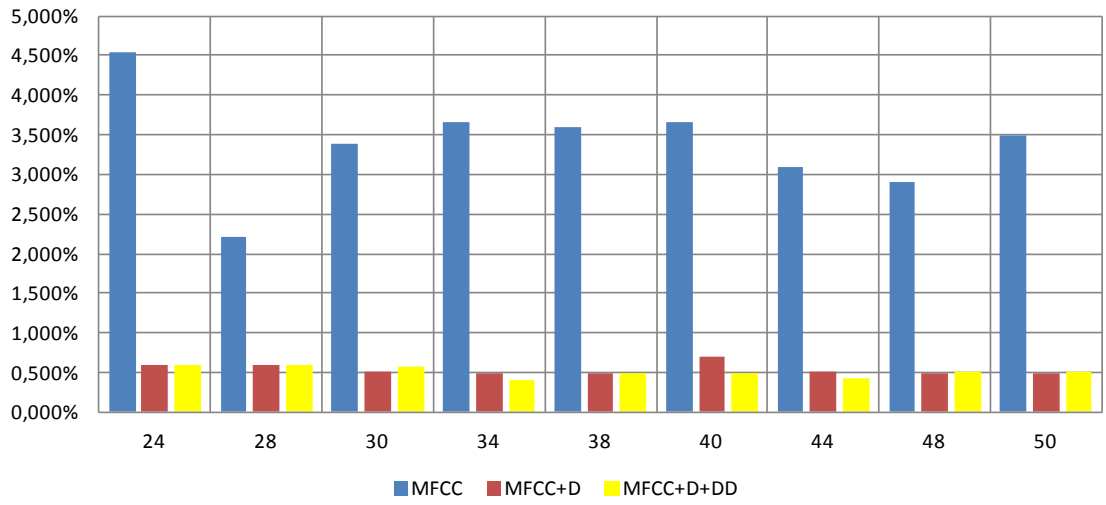

| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 4,02% | 3,99% | 2,89% | **2,20%** | 3,19% | 3,69% | 4,85% | 4,58% | 5,06% | 4,51% | 3,65% | 3,79% |
| MFCC+Δ | 0,51% | 1,10% | 0,92% | 0,71% | 0,60% | 0,59% | 0,50% | 0,59% | 0,60% | 0,68% | **0,49%** | 0,51% |
| MFCC+Δ+ΔΔ | 0,79% | 0,89% | 0,79% | 0,60% | 0,49% | 0,60% | 0,49% | 0,42% | **0,40%** | 0,50% | 0,49% | 0,49% |

**Figure 5-11** $EER_M$ obtained depending on the number of MFCCs and the use of Δ and ΔΔ (GDC – development set)



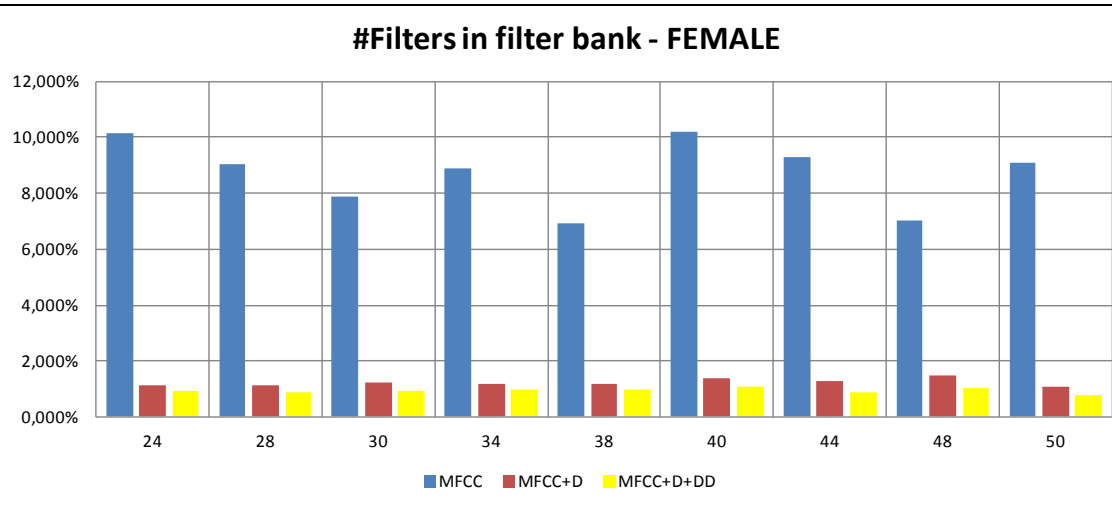

| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 9,34% | 7,03% | **6,91%** | 9,04% | 9,01% | 9,85% | 9,07% | 9,74% | 8,57% | 9,95% | 9,06% | 9,12% |
| MFCC+Δ | 1,62% | 1,12% | **1,10%** | 1,35% | 1,22% | 1,51% | 1,70% | 1,67% | 1,21% | 1,12% | 1,31% | 1,30% |
| MFCC+Δ+ΔΔ | 0,98% | **0,80%** | 0,90% | 0,91% | 1,00% | 1,30% | 1,20% | 1,42% | 1,42% | 1,01% | 1,28% | 1,20% |

**Figure 5-12** $EER_F$ obtained depending on the number of MFCCs and the use of Δ and ΔΔ (GDC – development set)



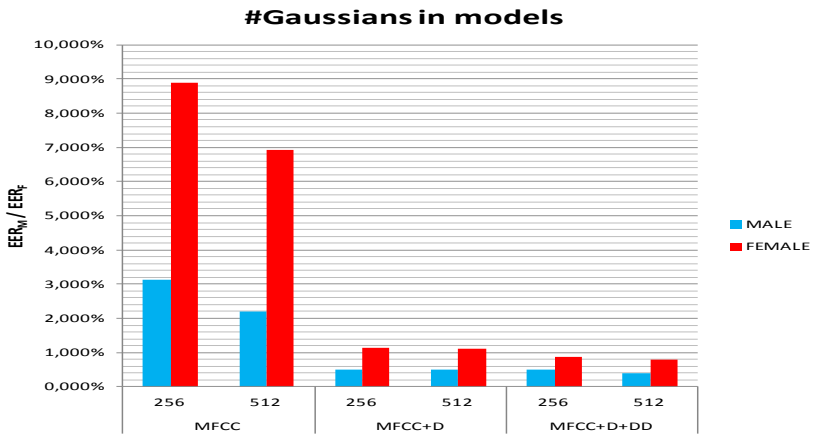

| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 4,54% | **2,20%** | 3,39% | 3,65% | 3,60% | 3,65% | 3,10% | 2,89% | 3,50% |
| MFCC+Δ | 0,60% | 0,59% | 0,51% | **0,49%** | 0,50% | 0,69% | 0,51% | 0,50% | 0,50% |
| MFCC+Δ+ΔΔ | 0,60% | 0,60% | 0,58% | **0,40%** | 0,49% | 0,49% | 0,42% | 0,50% | 0,52% |

**Figure 5-13** $EER_M$ obtained depending on the number of filters and the use of Δ and ΔΔ (GDC – development set)




| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 10,14% | 9,01% | 7,88% | 8,90% | **6,91%** | 10,18% | 9,26% | 7,03% | 9,07% |
| MFCC+Δ | 1,12% | 1,12% | 1,25% | 1,18% | 1,18% | 1,40% | 1,30% | 1,48% | **1,10%** |
| MFCC+Δ+ΔΔ | 0,90% | 0,90% | 0,91% | 0,99% | 1,00% | 1,09% | 0,88% | 1,02% | **0,80%** |

**Figure 5-14** $EER_F$ obtained depending on the number of filters and the use of Δ and ΔΔ (GDC – development set)




| | *G* | MALE - $EER_M$ | FEMALE - $EER_F$ |
|---|---|---|---|
| MFCC | 256 | 3,140% | 8,890% |
| | 512 | **2,208%** | **6,913%** |
| MFCC+Δ | 256 | 0,500% | 1,127% |
| | 512 | **0,498%** | **1,102%** |
| MFCC+Δ+ΔΔ | 256 | 0,496% | 0,883% |
| | 512 | **0,400%** | **0,803%** |

**Figure 5-15** $EER_M$ (blue) and $EER_F$ (red) obtained depending on the number of Gaussians and the use of Δ and ΔΔ (GDC – development set)

From Figure 5-11 and Figure 5-12, it is clear that different numbers of MFCCs are needed to precisely characterise speakers depending on their gender. Specifically, in the case of female speakers, no matter which configuration we are applying, better results are obtained when 16 MFCCs are used or 14 in the case of using MFCCs+Δ+ΔΔ. However, in the case of male speakers the number of MFCCs needs to be higher if we want to precisely model speakers. Particularly, 25 in the case of MFCCs+Δ, and 23 in the case of MFCCs+Δ+ΔΔ.

Regarding the number of filters used in the filter bank to extract MFCCs (see Figure 5-13 and Figure 5-14), differences arise again between genders, while remains more or less stable regardless of the configuration, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ. Specifically, for male speakers, the number of filters providing better results in terms of $EER_M$ is 34 while in the case of female speakers this same number tends to be higher, particularly, 50.

The only parameter in which no differences occur (see Figure 5-15) is the number of Gaussians used to build the UBM and the speaker models. In all the cases the better results are obtained when 512 Gaussians are used.

Therefore, we can conclude that in the set of tests carried out in this scenario, using the development set, a gender-dependent parameterisation presents a clear advantage over the gender-independent parameterisation, in terms of *HEER*, even in the case of using just classical parameters. Table 5-2 provides a comparison between the recognition rates obtained by the system, when a GDC or a GIC is used (the best results obtained so far are highlighted in light green). Additional columns have been added ($EER_X$ RR), which provide the relative reduction obtained by GDC, in terms of $EER_X$, respect to the corresponding GIC. The relative reduction in terms of *HERR* respect to GIC has been indicated in brackets in the *HEER* column.

| Parameters | Gen. | $F$ | $G$ | $\alpha$ | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| 16MFCC **(GIC MFCC)** | M/F | 38 | 512 | 5 | 3.602% [0.254] | - | 6.913% [0.179] | - | 5.258% |
| 18 MFCC **(GDC MFCC)** | M | 28 | 512 | 5 | 2.208% [0.293] | 38.69% | 6.913% [0.179] | 0.00% | 4.561% [13.25%] |
| 16 MFCC **(GDC MFCC)** | F | 38 | 512 | 5 | | | | | |
| | | | | | | | | | |
| 23MFCC+Δ **(GIC MFCC+Δ)** | M/F | 24 | 256 | 5 | 0.604% [0.324] | - | 1.217% [0.355] | - | 0.911% |
| 25MFCC+Δ **(GDC MFCC+Δ)** | M | 34 | 512 | 5 | 0.498% [0.370] | 17.57% | 1.102% [0.383] | 9.45% | 0.800% [12.14%] |
| 16 MFCC+Δ **(GDC MFCC+Δ)** | F | 50 | 512 | 5 | | | | | |
| | | | | | | | | | |
| 18MFCC+Δ+ΔΔ **(GIC MFCC+Δ+ΔΔ)** | M/F | 30 | 256 | 5 | 0.798% [0.346] | - | 0.916% [0.415] | - | 0.857% |
| 23 MFCC+Δ+ΔΔ **(GDC MFCC+Δ+ΔΔ)** | M | 34 | 512 | 5 | 0.400% [0.437] | 49.86% | 0.803% [0.432] | 12.32% | 0.602% [29.80%] |
| 14 MFCC+Δ+ΔΔ **(GDC MFCC+Δ+ΔΔ)** | F | 50 | 512 | 5 | | | | | |

**Table 5-2** GDC vs. GIC for HESPERIA's development set on scenario 1

From the results shown in Table 5-2, we can draw additional conclusions. First of all, the use of ΔΔ coefficients is clearly justified in this case, as the best recognition rates are obtained with that configuration. However, it must be noted that a GDC using MFCCs+Δ coefficients provides better results in terms of *HEER* than a GIC using MFCCs+Δ+ΔΔ coefficients. Thus the use of ΔΔ coefficients would be justified in the case of applying a GDC. Secondly, the use of a GDC systematically increases the recognition rates with respect to a GIC. In the case of using just MFCCs, a relative reduction of 13% in terms of *HEER*, is obtained thanks to the relative reduction of 39% in terms of $EER_M$. When Δ coefficients are incorporated into the feature vectors, a relative reduction of 17% is obtained in terms of *HEER*. This case constitutes a clear example that when using GIC a compromise must be reached between minimizing $EER_F$ and $EER_M$, since the GIC configuration is not the optimal in terms of *EER* for any of the genders. GDC combined with the use of MFCCs+Δ+ΔΔ provides a relative reduction in terms of *HEER* closed to 30% thanks to the relative reduction of 50% in terms of $EER_M$. Finally, the recognition rates obtained in terms of $EER_M$ and $EER_F$, are good enough to ensure that the representation of data under analysis is completely covered with the type of parameters (classical parameters) tested so far. Thus improving recognition rates by introducing new parameters can be difficult to achieve.

Figure 5-16 and Figure 5-17 show the DET curves for male and female speakers, for both configurations (GIC and GDC), and the different set of classical parameters used so far. It must be noted that it is possible for GIC and GDC to be the same, thus in the DET plots only one curve is going to be reflected (female MFCC case). Additionally, we must emphasise that the goal is the reduction of *EER* and not the *Area Under the Curve* (AUC – see Section 1.4), thus it can happen that GIC can show better results than

GDC for some of points of the curve. However, GDC will always produce better or at least equal results than GIC in terms of *EER*.

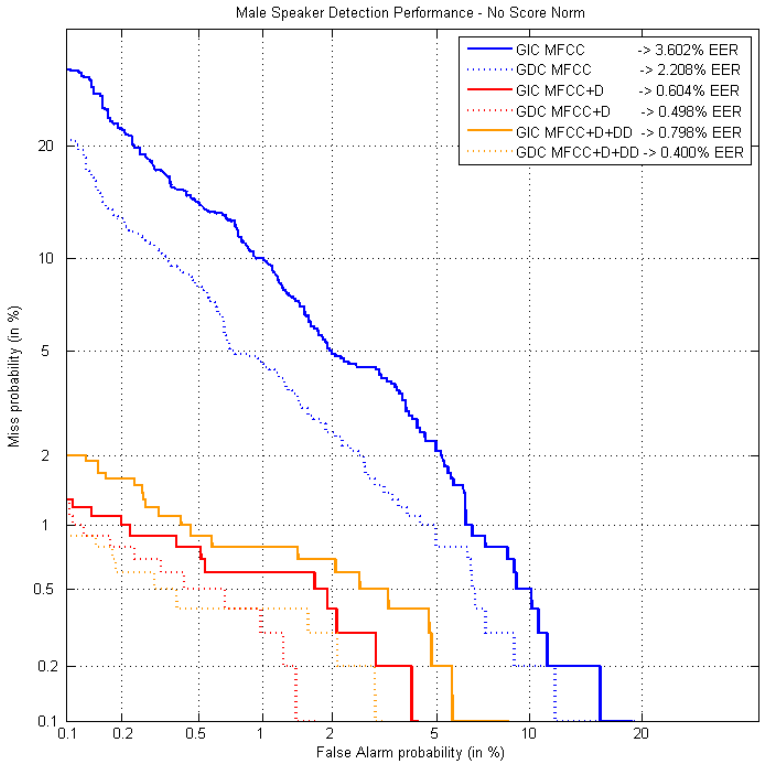



**Figure 5-16** DET curve for classic parameters on HESPERIA's male development set for GIC and GDC

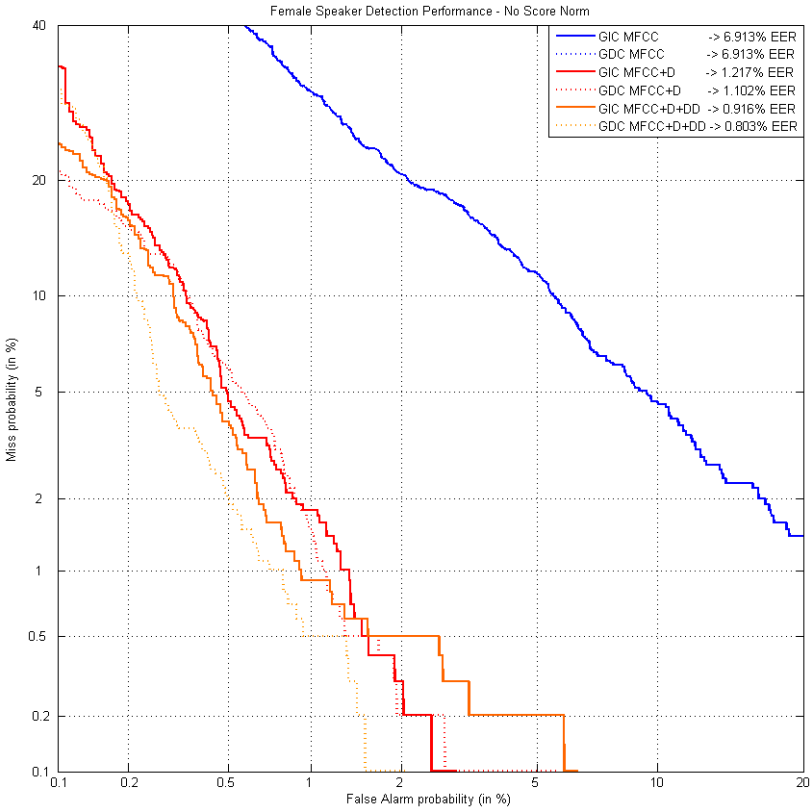



**Figure 5-17** DET curve for classic parameters on HESPERIA's female development set for GIC and GDC

In section 5.1.1, we introduced some extra parameters that can be used in conjunction with classical parameters to provide a better characterisation of speakers. We have tested the effect of these parameters by adding them to the GIC and GDC listed in Table 5-2, either alone or combined. Although all the combinations of E, ΔE, F0 and F3 have been tested on the development set, Table 5-3 includes the more successful sets of configurations, in terms of *HEER*. Although not reflected in Table 5-3, it must be noted that not all the tested combinations of what we have called extra parameters, provide a reduction of *HEER*, $EER_M$ or $EER_F$, if compared with the cases reflected in Table 5-2. For instance, in the case of female speakers when only MFCCs are used, no combination of extra parameters provides an improvement in terms of $EER_F$. Moreover, depending on the set of classical parameters used (i.e. MFCCs, MFCCs+Δ, MFCCs+Δ+ΔΔ), and on the gender of the speakers, the combination of extra parameters providing an improvement may be different.

Leaving aside the MFCC configuration, it is clear from the results shown in Table 5-3, that the inclusion of the proposed extra parameters provides an extra benefit, helping to improve recognition rates, especially in a gender-dependent configuration. Better results in terms of $EER_M$, $EER_F$, and *HEER* are obtained in all the cases using a GDC no matter whether MFCCs, MFCCs+Δ, or MFCCs+Δ+ΔΔ parameters are used. It is worth noting that E and ΔE parameters, which are part of the state-of-the-art in speech recognition (as well as in speaker recognition), also provide an advantage in this case as recordings used in this scenario are captured in controlled conditions under high quality standards. In following sections we will check their behaviour in noisy and low quality scenarios. In the set of configurations presented in Table 5-3, we can also confirm that the new parameter F3 proposed in this thesis (third formant estimate), also helps to improve recognition rates both on male and female speakers.

| Parameters | Gen. | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER* | *HEER* RR |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC** | M/F | - | 3.602% [0.254] | - | 6.913% [0.179] | - | 5.258% | - |
| **GIC MFCC** | M/F | No improvement | - | - | - | - | - | - |
| **GDC MFCC** | M | - | 2.208% [0.293] | 38.69% | 6.913% [0.179] | 0.00% | 4.561% | 13.25% |
| | F | - | | | | | | |
| **GDC MFCC** | M | E+ΔE | 2.104% [0.302] | 41.58% | - | - | - | - |
| | F | No improve. | | | | | | |
| | | | | | | | | |
| **GIC MFCC+Δ** | M/F | - | 0.604% [0.324] | - | 1.217% [0.355] | - | 0.911% | - |
| **GIC MFCC+Δ** | M/F | ΔE+F3 | 0.600% [0.308] | 0.69% | 1.205% [0.365] | 1.03% | 0.902% | 0.91% |
| **GDC MFCC+Δ** | M | - | 0.498% [0.370] | 17.57% | 1.102% [0.383] | 9.45% | 0.800% | 12.14% |
| | F | - | | | | | | |
| **GDC MFCC+Δ** | M | E+ΔE+F3 | 0.400% [0.395] | 33.79% | 1.098% [0.418] | 9.79% | 0.748% | 17.75% |
| | F | E+ΔE+F3 | | | | | | |
| | | | | | | | | |
| **GIC MFCC+Δ+ΔΔ** | M/F | - | 0.798% [0.346] | - | 0.916% [0.415] | - | 0.857% | - |
| **GIC MFCC+Δ+ΔΔ** | M/F | E+ΔE+F0+F3 | 0.400% [0.401] | 49.86% | 1.006% [0.406] | -9.81% | 0.703% | 17.96% |
| **GDC MFCC+Δ+ΔΔ** | M | - | 0.400% [0.437] | 49.86% | 0.803% [0.432] | 12.32% | 0.602% | 29.80% |
| | F | - | | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | E+ΔE+F0+F3 | 0.313% [0.440] | 60.83% | 0.577% [0.438] | 36.98% | 0.445% | 48.08% |
| | F | E+F3 | | | | | | |

**Table 5-3** $EER_M$, $EER_F$, and *HEER* obtained on the development set when extra parameters are included on the feature vectors for GIC and GDC

Finally, we must highlight three important facts about the use of MFCCs+Δ+ΔΔ. First of all, that according to the results presented in Table 5-3, its use would be justified in a gender-dependent configuration. As, for instance, GIC MFCCs+Δ provide lower $EER_M$ than GIC MFCCs+Δ+ΔΔ. Moreover, a gender-dependent configuration using MFCCs+Δ, augmented with specific extra parameters clearly provides better recognition rates in terms of *HEER*, than GIC MFCCs+Δ+ΔΔ, and almost the same as GIC MFCCs+Δ+ΔΔ augmented with extra parameters. Secondly, although not reflected in Table 5-3, we also have tested a configuration that could be considered as the state-of-the-art standard and thus the baseline to beat, namely GIC MFCCs+Δ+ΔΔ+E+ΔE. Under this configuration the results obtained are shown in Table 5-4; which are clearly worse than the ones provided by the MFCCs+Δ+ΔΔ configuration without extra parameters. Finally, it must be noted that we are working on a text-constrained scenario, thus the message conveyed by the recordings will present limited variability. In this sense, the use of parameters that are related with the message transmitted can be useful, as demonstrated by the results presented.

| Parameters | Gen. | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER* | *HEER* RR |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ+ΔΔ** | M/F | - | 0.798% [0.346] | - | 0.916% [0.415] | - | 0.857% | - |
| **GIC MFCC+Δ+ΔΔ** | M/F | E+ΔE+F0+F3 | 0.400% [0.401] | 49.86% | 1.006% [0.406] | -9.81% | 0.703% | 17.96% |
| **GIC MFCC+Δ+ΔΔ** | M/F | E+ΔE | 0.900% [0.384] | -125% | 1.010% [0.495] | -0.42% | 0.955% | -11.45% |

**Table 5-4** $EER_M$, $EER_F$, and *HEER* obtained on the development set comparing different GIC MFCC+Δ+ΔΔ configurations

Figure 5-18 and Figure 5-19 respectively provide the DET curves for male and female speakers, comparing the most relevant configurations for the present study, i.e., GIC MFCCs+Δ, GIC MFCCs+Δ+ExtParam., GDC MFCCs+Δ+ExtParam., GIC MFCCs+Δ+ΔΔ, GIC MFCCs+Δ+ΔΔ+ExtParam., GDC MFCCs+Δ+ΔΔ+ExtParam.. In other words, we compare gender-dependent with gender-independent configurations, while we check as well the influence of ΔΔ coefficients and the extra proposed parameters.

Next, we present the results obtained on the development set when using the GDEB front-end and thus not only operating on a gender-dependent configuration but also including information extracted from the vocal tract and glottal source estimates conveniently parameterised. The approach that has been followed consists in incorporating the extended biometric coefficients into the best gender-dependent configuration obtained so far without ΔΔ parameters, in two stages. First we incorporate a set of parameters extracted from the glottal source estimate (labelled as GSE), and once a specific configuration improving previous results is found, we continue by incorporating parameters extracted from the vocal tract estimate (labelled as VTE). We proceed this way because the vocal tract estimate is expected to be more related with the message carried out by voice, rather than to specific biometrical characteristics of speakers. Therefore, although we are dealing with text-constrained trials (so message variability is limited), GSE is supposed to provide more benefits than VTE in terms of recognition rates. Additionally, we rule out the use of ΔΔ in this section, as we believe that the proposed GDEB parameterisation may represent the speaker more accurately than the one including ΔΔ coefficients, which are supposed to be more related to the message transmitted.

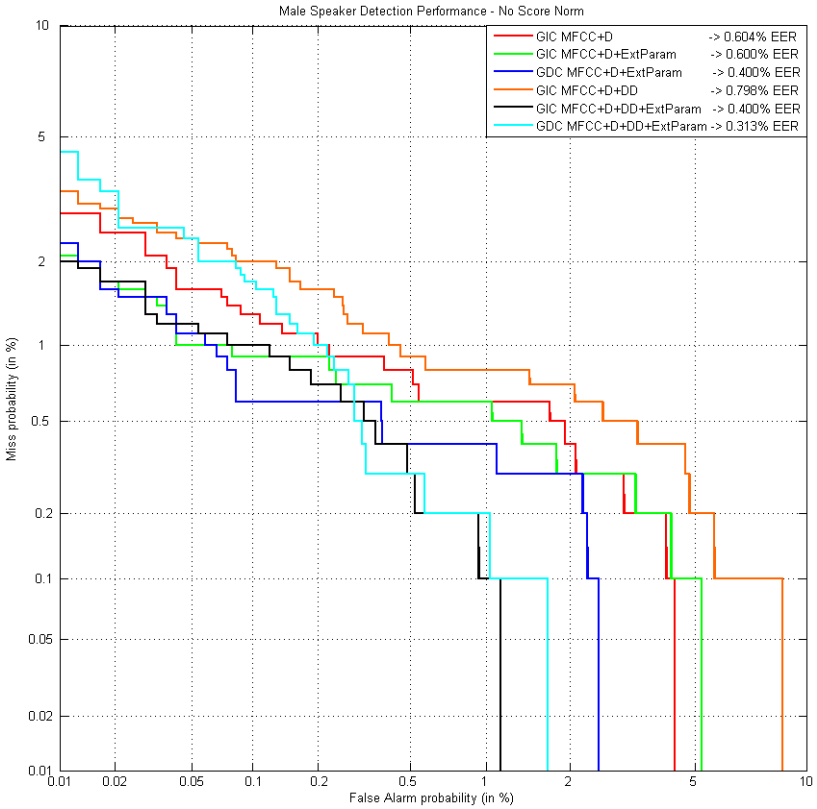



**Figure 5-18** DET curves comparing different set of parameters under GIC and GDC without extended biometrics on male's development set

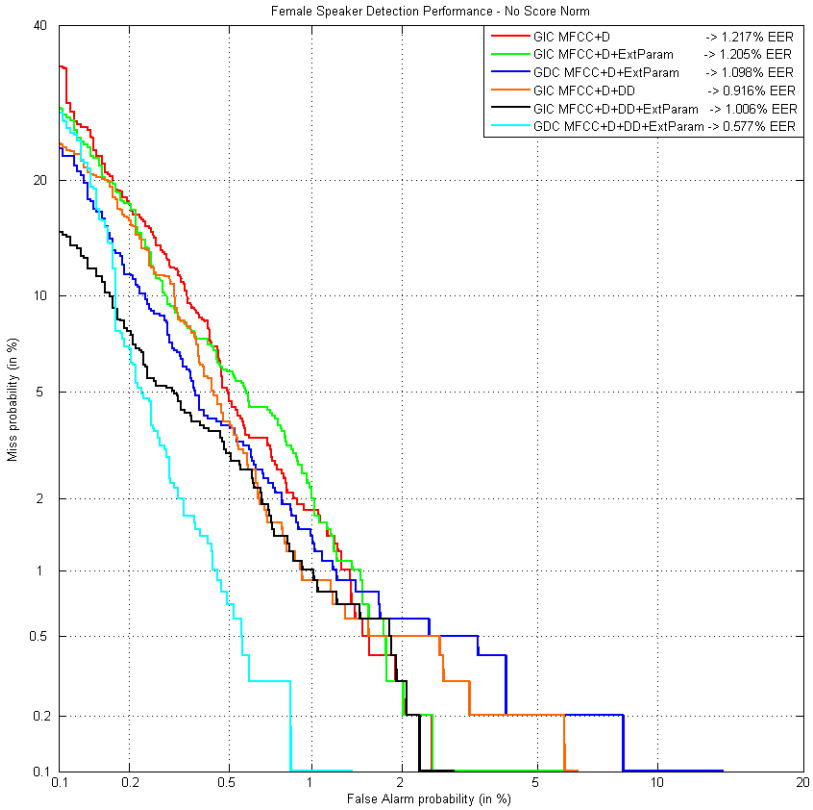



**Figure 5-19** DET curves comparing different set of parameters under GIC and GDC without extended biometrics on female's development set

| Parameters | Gen. | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | - | - | 0.604% [0.324] | | 1.217% [0.355] | - | 0.911% [-] |
| **GDC MFCC+Δ** | M | - | E+ΔE+F3 | 0.400% [0.395] | 33.79% | 1.098% [0.418] | 9.79% | 0.748% [17.75%] |
| | F | - | E+ΔE+F3 | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | - | E+ΔE+F0+F3 | 0.313% [0.440] | 48.27% | 0.577% [0.438] | 52.57% | 0.445% [51.15%] |
| | F | - | E+F3 | | | | | |
| **GSE** | M | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 11-Channel Filter bank 6 MFCC | E+ΔE+F3 | 0.213% [0.454] | 64.82% | 0.809% [0.477] | 33.50% | 0.511% [43.89%] |
| | F | **Source-Tract Sep. Alg:** Prediction Order: 20 Forgetting Factor: 0.995 **GSE:** 17-Channel Filter bank 6 MFCC | E+ΔE+F3 | | | | | |

**Table 5-5** $EER_M$, $EER_F$, and *HEER* obtained on development set (no score normalisation), comparing classical parameters with extra parameters and extended biometric parameters. (PO: Prediction Order; FF: Forgetting Factor; FB: Filter bank).

Although multiple configurations have been tested, regarding the multiple variables that can be tune in the GDEB front-end, Table 5-5 shows the ultimate configurations chosen for each gender, as well as the recognition rates obtained in each case in terms of $EER_M$, $EER_F$ and *HEER*. Additionally, the relative reduction (RR) in terms of $EER_X$ and *HEER*, compared to GIC MFCCs+Δ configuration is also provided.

DET curves corresponding to the results presented in Table 5-5 are depicted in Figure 5-20 for male speakers and Figure 5-21 for female speakers. Clearly, the parameterisation generated by the GDEB front-end, in this case just including information from the glottal source estimate in the form of mel-frequency cepstrum coefficients, is the one providing best results on the development set for male speakers. In the case of female speakers, the parameterisation generated by the GDEB front-end increases recognition rates respect to GIC and GDC when MFCCs+Δ+ExtParam. is used, but it is not able to improve recognition rates obtained by GDC MFCCs+Δ+ΔΔ+ExtParam..

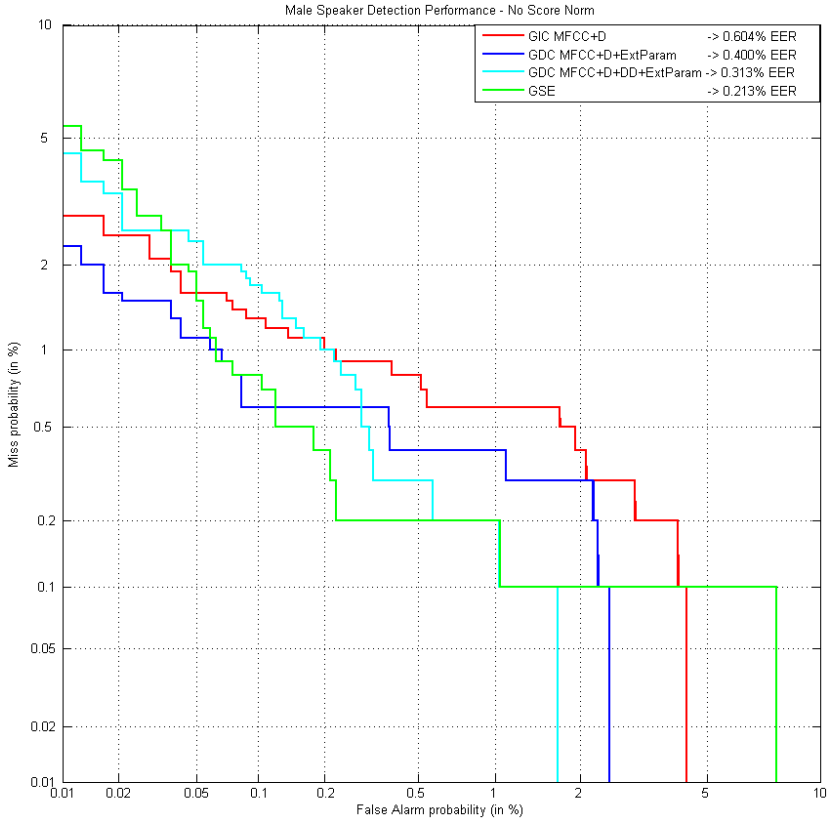



**Figure 5-20** DET curves comparing classical parameters and GDEB on HESPERIA's development set for male speakers

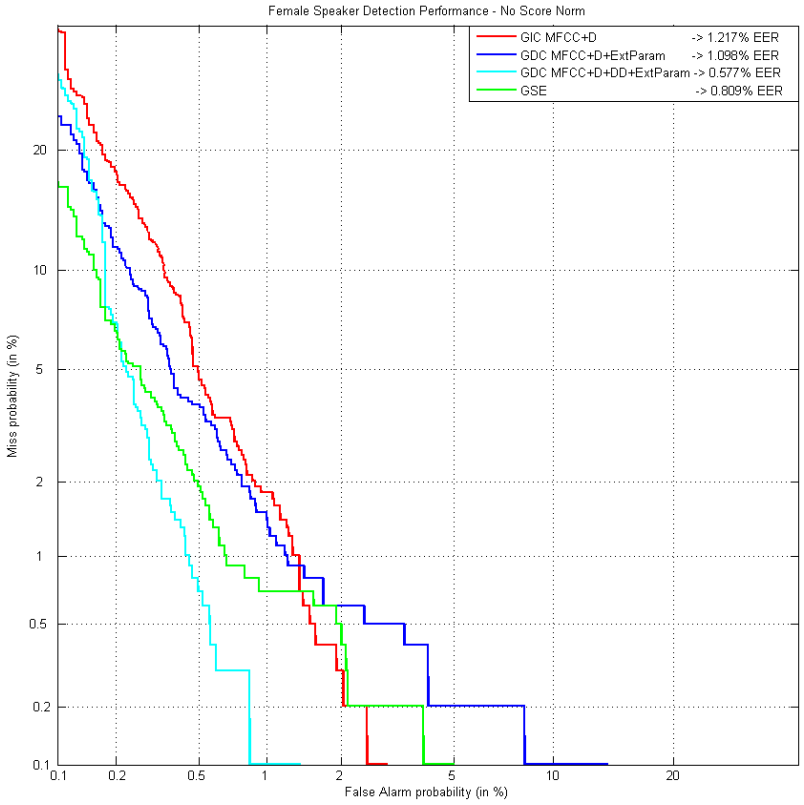



**Figure 5-21** DET curves comparing classical parameters and GDEB on HESPERIA's development set for female speakers

Before incorporating additional parameters from the vocal tract estimate into feature vectors, it seems necessary to make some observations on the use of the parameters extracted from the glottal source estimate. So far we have selected specific configurations producing the better performance in terms of $EER_X$ (i.e. no matter whether male or female case) and *HEER*. This means that in theory no better results on the development set can be obtained using the *Baseline* or the GDEB front-ends; additionally one might think that the improvements derived from incorporating information from the glottal source estimate will be highly dependent on the specific configuration chosen. However, nothing could be further away from truth, as the improvement in recognition rates appears systematically when incorporating GSE parameters under different configurations. Figure 5-22 provides a comparison of the $EER_M$ obtained under different GSE parameterisations for male speakers, while Figure 5-23 provides the same information but for female speakers. The green solid line represents the minimum $EER_X$ (y-axis) obtained when GSE is incorporated into the feature vector in form of MFCC. Different numbers of $MFCC_G$={2,4,6,8,10} have been tested, which have been computed applying a filter bank with different number of filters $F_G$=[4...38] (x-axis). Each point in the x-axis represents the minimum *EER* obtained for a specific value of $F_G$, regardless $MFCC_G$ value.

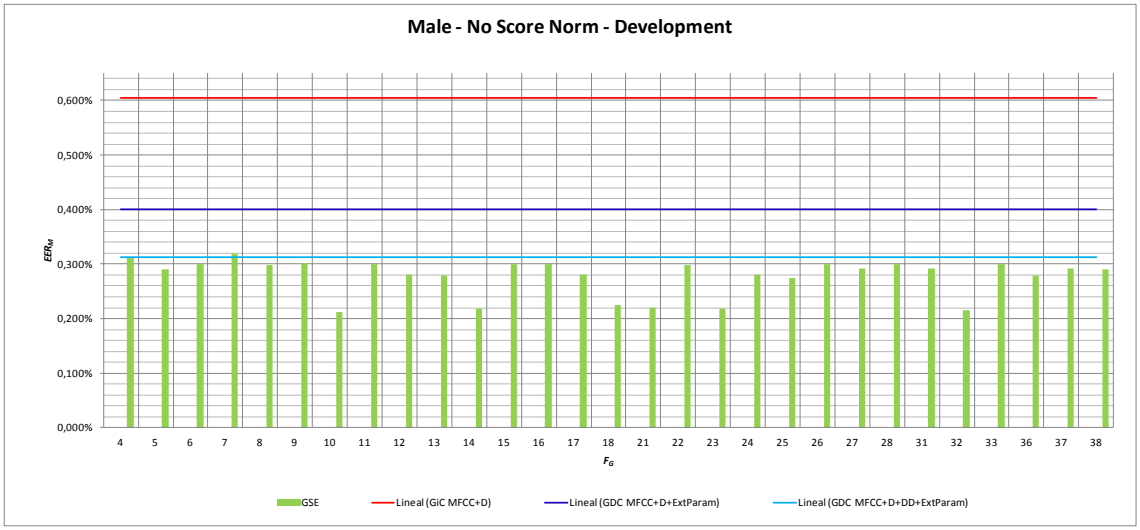



**Figure 5-22** Influence of GSE configuration on the $EER_M$ (development set)

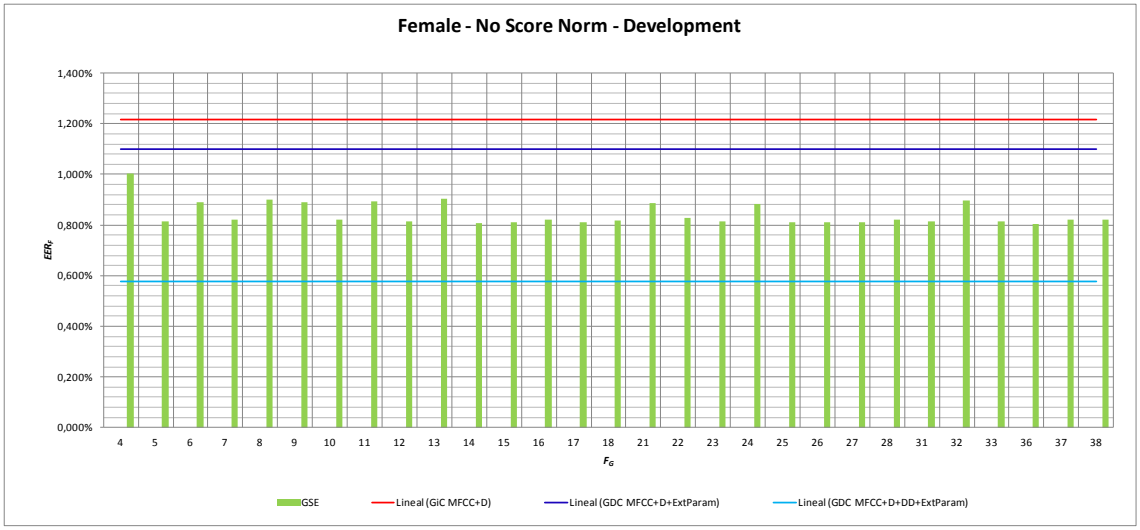



**Figure 5-23** Influence of GSE configuration on the $EER_F$ (development set)

As stated before, in the case of female speakers, the recognition rates obtained by the GDC MFCCs+Δ+ΔΔ+ExtParam. are not improved by any configuration provided by the GDEB front-end. However, as depicted in Figure 5-23, the $EER_F$ obtained when glottal information is incorporated in the feature vector is systematically lower that the $EER_F$ obtained when no GSE information is present in the GDC MFCCs+Δ.

Another factor that must be analysed is the influence of score normalisation algorithms in recognition rates. For this reason, the same experiments that have been reported so far (when no score normalisation is applied), have been conducted but applying ZNorm, TNorm and ZTNorm. Table 5-6 to Table 5-8 provide results equivalent to those that have been reported in Table 5-5 (when no score normalisation is applied), but for the previously cited score normalisations.

It must be noted that Table 5-6 to Table 5-8 includes an additional column providing the set up for classical parameters, i.e. MFCCs+Δ, MFCCs+Δ+ΔΔ, regarding the number of filters in filter bank and the number of MFCCs. This is interesting in order to show that the use of different types of score normalisation influences the values that should be assigned to the configurable parameters of the *Baseline* front-end in the search for the minimum *EER*. In other words, we cannot expect the most successful set of selected

parameters when no score normalisation is applied, to also provide the best results under different score normalisations.

From the results shown on Table 5-6, Table 5-7, and Table 5-8, it is clear that no matter the score normalisation applied, the most successful results, in terms of $EER_M$ (except for ZNorm, which equals GDC MFCCs+Δ $EER_M$), and *HEER* are always obtained when GSE is included as MFCCs in the feature vector. In the case of female speakers, the above conclusion is valid except for the specific case in which TNorm is applied. However, the difference in terms of $EER_F$ between the best result (GDC MFCC+Δ+ΔΔ $\rightarrow$ $EER_F$=0.803%) and the one obtained by GSE ($EER_F$ = 0.805%) are negligible. Anyway, regarding the results obtained in the set of tests carried out, we can conclude that a gender-dependent characterisation of speakers provides better results in terms of recognition rates than a gender-independent characterisation. Moreover, the proposed parameter, F3, is most of the times a component of the configurations providing most favourable results, thus proving its value.

Regarding the use of score normalisations, it must be noted that its use does not always provide an improvement in terms of $EER_X$. In this scenario, the best results are obtained when ZTNorm is applied in the case of female speakers, while in the case of male speakers applying ZNorm provides the best result in terms of $EER_M$.

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=38, *MFCC*=14, *G*=512, *α*=10 | - | - | 0.800% [3.953] | - | 1.305% [3.694] | - | 1.053% [-] |
| **GDC MFCC+Δ** | M | *F*=38, *MFCC*=25, *G*=512, *α*=16 | - | E+F3 | 0.200% [4.708] | 75.00% | 1.222% [3.800] | 6.41% | 0.711% [32.45%] |
| | F | *F*=38, *MFCC*=14, *G*=512, *α*=8 | - | E | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=44, *MFCC*=22, *G*=512, *α*=16 | - | E+F3 | 0.300% [4.523] | 62.50% | 0.927% [3.347] | 29.00% | 0.613% [41.73%] |
| | F | *F*=50, *MFCC*=14, *G*=512, *α*=10 | - | E+ΔE+F3 | | | | | |
| **GSE** | M | *F*=38, *MFCC*=25, *G*=512, *α*=10 | **Source-Tract Sep. Alg:** Prediction Order: 32 Forgetting Factor: 0.995 **GSE:** 13-Channel Filter bank 4 MFCC | E+F3 | 0.200% [4.148] | 75.00% | 0.904% [3.924%] | 30.76% | 0.552% [47.57%] |
| | F | *F*=38, *MFCC*=14, *G*=512, *α*=8 | **Source-Tract Sep. Alg:** Prediction Order: 23 Forgetting Factor: 0.995 **GSE:** 37-Channel Filter bank 8 MFCC | E | | | | | |

**Table 5-6** $EER_M$, $EER_F$, and *HEER* obtained for the development set (ZNorm), comparing classical parameters with extra parameters and extended biometric parameters

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=50, *MFCC*=14, *G*=512, *α* =5 | - | - | 0.900% [3.560] | - | 0.918% [3.572] | - | 0.909% [-] |
| **GDC MFCC+Δ** | M | *F*=44, *MFCC*=19, *G*=512, *α*=5 | - | F0+F3 | 0.500% [3.898] | 44.44% | 0.828% [3.607] | 9.79% | 0.664% [26.94%] |
| | F | *F*=38, *MFCC*=14, *G*=512, *α*=5 | - | F0+F3 | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=38, *MFCC*=16, *G*=512, *α*=5 | - | - | 0.600% [4.201] | 33.33% | 0.803% [3.581] | 12.52% | 0.702% [22.82%] |
| | F | *F*=40, *MFCC*=14, *G*=256, *α*=5 | - | F0+F3 | | | | | |
| **GSE** | M | *F*=44, *MFCC*=19, *G*=512, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 23-Channel Filter bank 4 MFCC | F0+F3 | 0.400% [4.032] | 55.55% | 0.805% [3.715%] | 12.30% | 0.603% [33.71%] |
| | F | *F*=38, *MFCC*=14, *G*=512, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 23 Forgetting Factor: 0.995 **GSE:** 18-Channel Filter bank 8 MFCC | F0+F3 | | | | | |

**Table 5-7** $EER_M$, $EER_F$, and *HEER* obtained for the development set (TNorm), comparing classical parameters with extra parameters and extended biometric parameters

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=38, *MFCC*=14, *G*=512, *α*=10 | - | - | 0.879% [3.151] | - | 0.910% [3.135] | - | 0.895% [-] |
| **GDC MFCC+Δ** | M | *F*=38, *MFCC*=25, *G*=256, *α*=5 | - | - | 0.300% [3.472] | 65.87% | 0.899% [2.880] | 1.149% | 0.600% [32.95%] |
| | F | *F*=38, *MFCC*=14, *G*=512, *α*=5 | - | | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=24, *MFCC*=23, *G*=256, *α*=5 | - | - | 0.400% [3.472] | 54.50% | 0.778% [2.582] | 14.48% | 0.589% [34.14%] |
| | F | *F*=50, *MFCC*=14, *G*=256, *α*=5 | - | E+F3 | | | | | |
| **GSE** | M | *F*=38, *MFCC*=25, *G*=256, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 10 Forgetting Factor:0.995 **GSE:** 9-Channel Filter bank 2 MFCC | | 0.223% [3.726] | 74.65% | 0.630% [3.038] | 30.80% | 0.426% [52.34%] |
| | F | *F*=38, *MFCC*=14, *G*=512, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 20 Forgetting Factor:0.995 **GSE:** 21-Channel Filter bank 4 MFCC | | | | | | |

**Table 5-8** $EER_M$, $EER_F$, and *HEER* obtained for the development set (ZTNorm), comparing classical parameters with extra parameters and extended biometric parameters

In order to complete previous results, Figure 5-24 and Figure 5-25 provide DET curves for male and female speakers, respectively, so the performance of the *Baseline* front end (using GIC MFFC+Δ) can be compared with the GDEB front-end when different score normalisation techniques are applied.

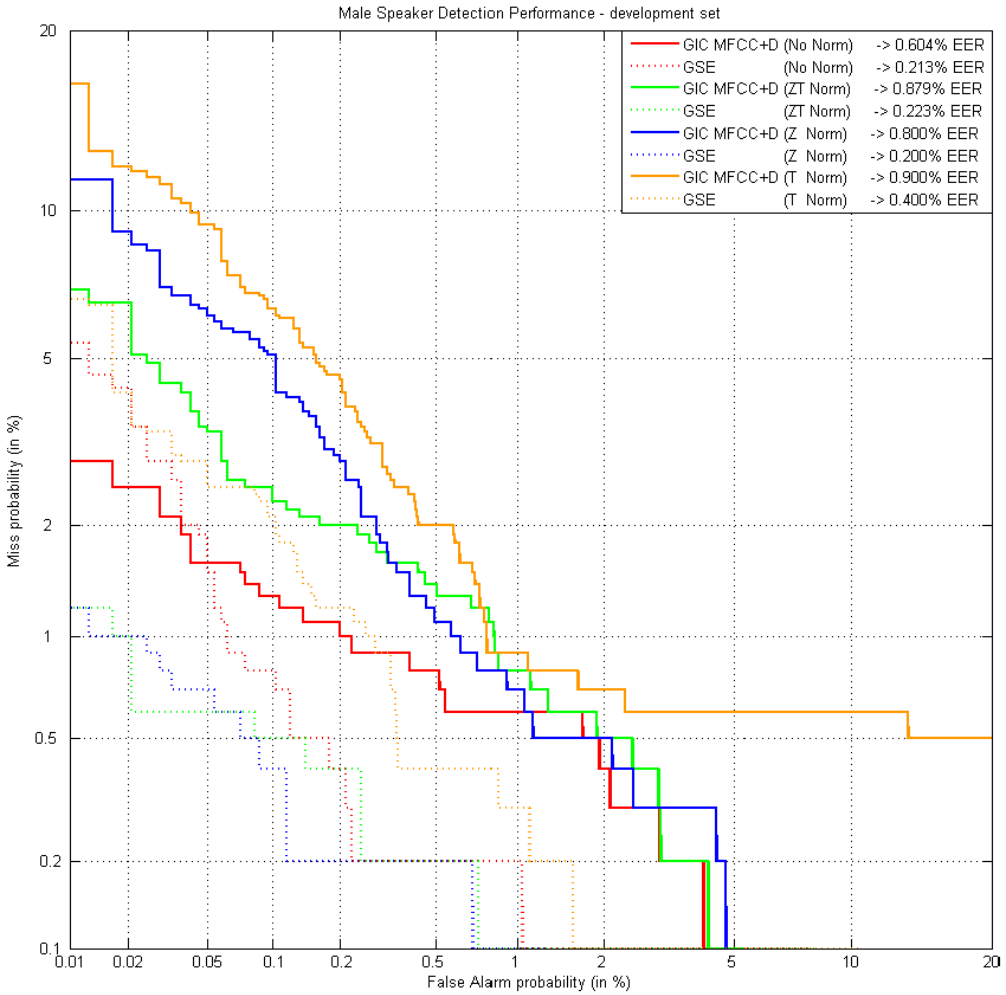



**Figure 5-24** DET curves for *Baseline* and GDEB front-end, applying different score normalisation techniques (male's development set)

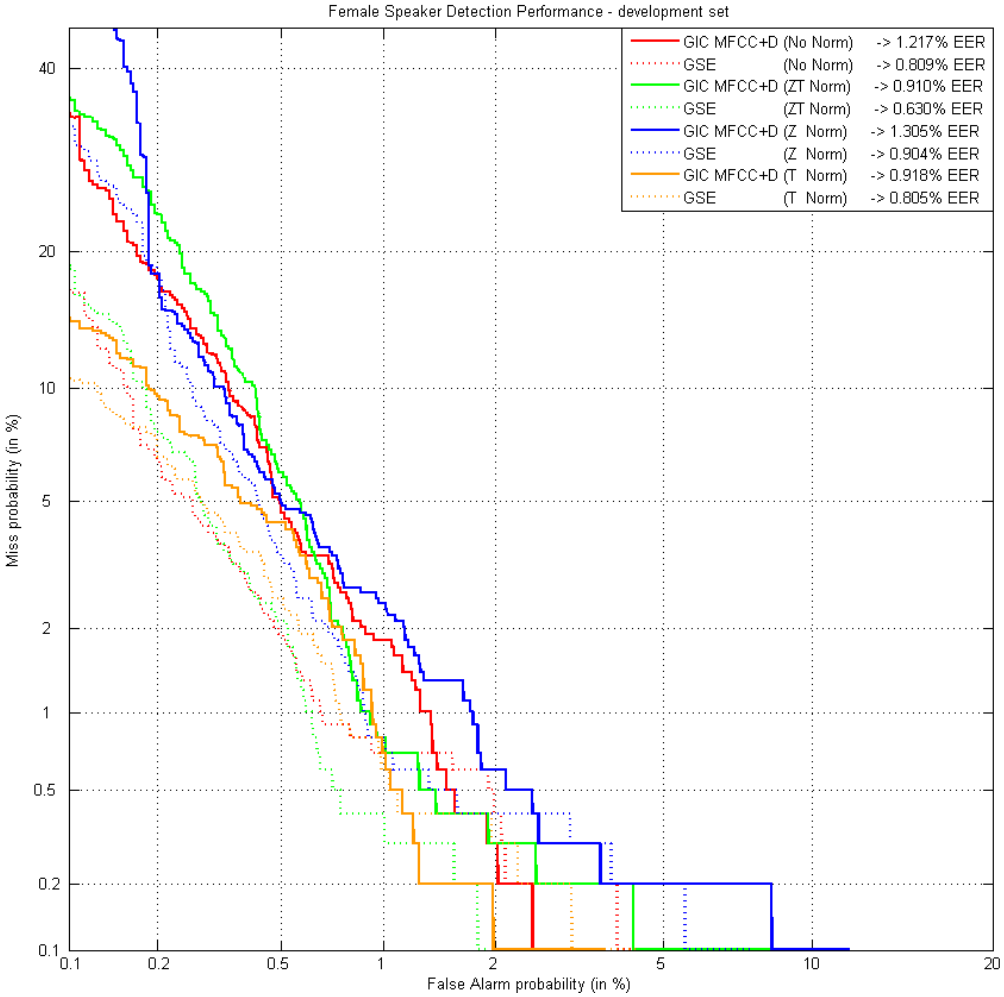



**Figure 5-25** DET curves for *Baseline* and GDEB front-end, applying different score normalisation techniques (female's development set)

There are still two additional tasks to be performed, in order to finish the study on this scenario. First of all, we have to verify the usefulness of the information provided by the vocal tract estimate (labelled as VTE), not being included yet in the experiments carried out so far. As we are working on a text-constrained scenario, with limited variability in message, there are chances on that VTE may be useful in providing an accurate description of speakers. To limit the experiments, we only run an additional test based on the configuration providing better results in terms of $EER_X$, i.e. GDC GSE ZTNorm for female speakers and GDC GSE ZNorm for male speakers.

In the case of the tests run on male speakers, the use of GSE information combined with VTE information systematically produces recognition rates that outperform the recognition rates obtained by the GIC MFCCs+Δ and GDC MFCCs+Δ+ΔΔ+ExtParam. configurations. However, it is difficult to find a configuration that outperforms the recognition rates obtained when using GDC MFCCs+Δ+ExtParam.. It must be noted that even in the case of incorporating GSE information into the feature vector we have only succeeded in matching the same $EER_M$. This situation is repeated for the case of introducing VTE information, i.e. the lower $EER_M$ obtained equals the one obtained by GDC MFCCs+Δ+ExtParam.. Therefore in this scenario VTE seems not to provide additional information. The above discussion is also valid for female speakers.

Once we have found the configurations allowing us to obtain the most successful results in terms of $EER_X$, on the development set, using both the *Baseline* and the GDEB front-ends, the final step consists in verifying if the behaviour of the speaker recognition system with the selected configurations holds for the evaluation set, i.e. for unknown data, or if the results are affected by overtraining on the development set. Table 5-9 provides the results obtained on the evaluation set applying the different configurations previously selected (see Table 5-5 to Table 5-8) for different score normalisations. Most successful results for the different score normalisation techniques applied are highlighted in green.

| Score Norm | Parameters | $EER_M$ [$\theta_M$] | $HTER_M$ | $HTER_M$ RR | $EER_F$ [$\theta_F$] | $HTER_F$ | $HTER_F$ RR |
|---|---|---|---|---|---|---|---|
| No Norm | **GIC MFCC+Δ** | 0.604% [0.324] | 0.963% | - | 1.217% [0.355] | 5.370% | - |
| | **GDC MFCC+Δ+ExtParm.** | 0.400% [0.395] | 1.169% | -21.35% | 1.098% [0.418] | 3.226% | 39.92% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 0.313% [0.440] | 0.696% | 27.69% | 0.577% [0.438] | 6.755% | -25.81% |
| | **GSE** | 0.213% [0.454] | 0.651% | 32.37% | 0.809% [0.477] | 3.326% | 38.05% |
| ZNorm | | | | | | | |
| | **GIC MFCC+Δ** | 0.800% [3.953] | 1.938% | - | 1.305% [3.694] | 4.854% | - |
| | **GDC MFCC+Δ+ ExtParm.** | 0.200% [4.708] | 1.870% | 3.48% | 1.222% [3.800] | 5.536% | -14.06% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 0.300% [4.523] | 1.204% | 37.85% | 0.927% [3.347] | 5.537% | -14.07% |
| | **GSE** | 0.200% [4.148] | 0.687% | 64.53% | 0.904% [3.924%] | 4.824% | 0.61% |
| TNorm | | | | | | | |
| | **GIC MFCC+Δ** | 0.900% [3.560] | 0.741% | - | 0.918% [3.572] | 4.048% | - |
| | **GDC MFCC+Δ+ ExtParm.** | 0.500% [3.898] | 0.401% | 45.90% | 0.828% [3.607] | 4.437% | -9.61% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 0.600% [4.201] | 0.500% | 32.52% | 0.803% [3.581] | 4.779% | -18.07% |
| | **GSE** | 0.400% [4.032] | 0.339% | 54.26% | 0.805% [3.715%] | 2.577% | 36.34% |
| ZTNorm | | | | | | | |
| | **GIC MFCC+Δ** | 0.879% [3.151] | 0.941% | - | 0.910% [3.135] | 3.958% | - |
| | **GDC MFCC+Δ+ ExtParm.** | 0.300% [3.472] | 1.252% | -33.04% | 0.899% [2.880] | 3.496% | 11.66% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 0.400% [3.472] | 0.683% | 27.42% | 0.778% [2.582] | 4.107% | -3.78% |
| | **GSE** | 0.223% [3.726] | 0.708% | 24.72% | 0.630% [3.038%] | 2.593% | 34.47% |

**Table 5-9** $HTER_X$ obtained for selected configurations on the evaluation set, applying different score normalisations

The results obtained in terms of $HTER_X$, allow us to draw two important conclusions. First of all, we can assert that the use of GSE conveniently parameterised provides an

improvement in recognition rates, which remains consistent over the development set and the evaluation set. Specifically, for the male case when applying ZNorm we obtained a relative reduction of 75% in terms of $EER_M$ (from $EER_M$ =0.800%, when classical gender-independent characterisation is used to $EER_M$ =0,200%), which is transformed into a relative reduction of 64% in terms of $HTER_M$, when moving into the evaluation set (from $HTER_M$ =1.938%, when classical gender-independent characterisation is used to $HTER_M$ =0,687%). For female speakers, when ZTNorm is applied a relative reduction of 30% in terms of $EER_F$ (from $EER_F$ =0.910%, when classical gender-independent characterisation is used to $EER_F$ =0,630%), which is transformed into a relative reduction of 34% in terms of $HTER_F$, when moving into the evaluation set (from $HTER_F$ =3.958%, when classical gender-independent characterisation is used to $HTER_M$ =2.593%)

However, there are some interesting results that deserve additional consideration. In the case of not applying any score normalisation, for female speakers, the better results on development set where obtained when using ΔΔ coefficients with a relative reduction in terms of $EER_F$ of more than 50%. However, given the results provided by this configuration in the evaluation set, it is clear that it suffers for overtraining on the development set as the results are far worse that the other results presented. This situation appears again when TNorm is applied, even though not as exaggerated. Additionally there are two situations in which the GSE configuration does not provide the most successful results on the evaluation set, namely female No Norm and male ZTNorm, even though they are better than the GIC configuration. However, in these two cases, if we consider the results globally, i.e. taking into account development and evaluation sets, the use of GSE information clearly offers better results.

Finally, Figure 5-26 to Figure 5-33, provide the DET curves that represent the results obtained in the evaluation set for male and female speakers, which are reflected in Table 5-9.

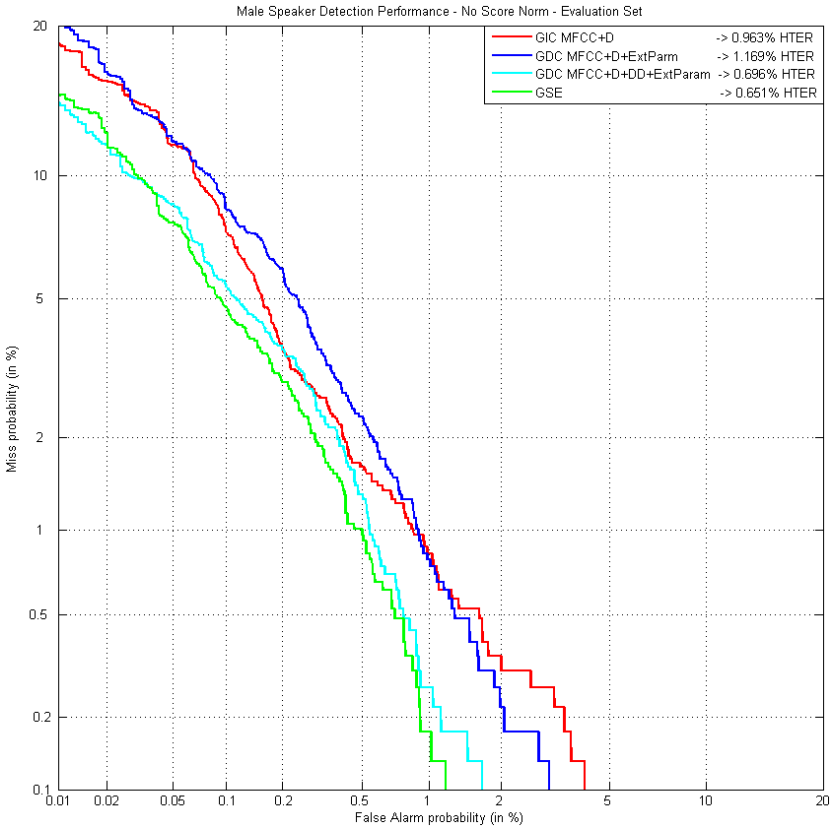



**Figure 5-26** Male's DET curves on the evaluation set from HESPERIA, without applying any score normalisation technique

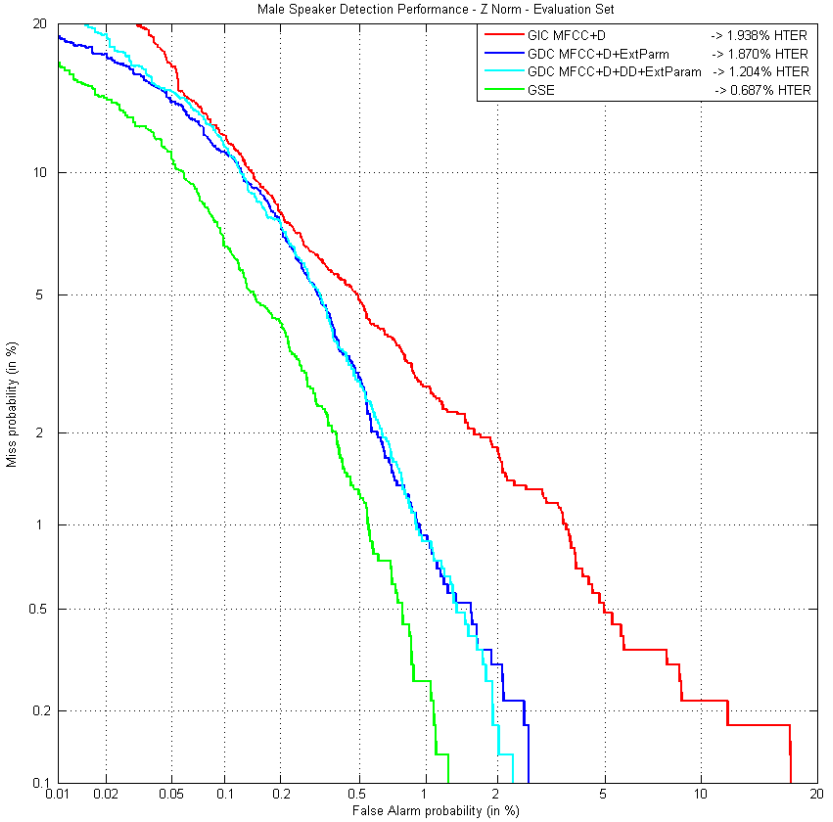



**Figure 5-27** Male's DET curves on the evaluation set from HESPERIA, applying ZNorm

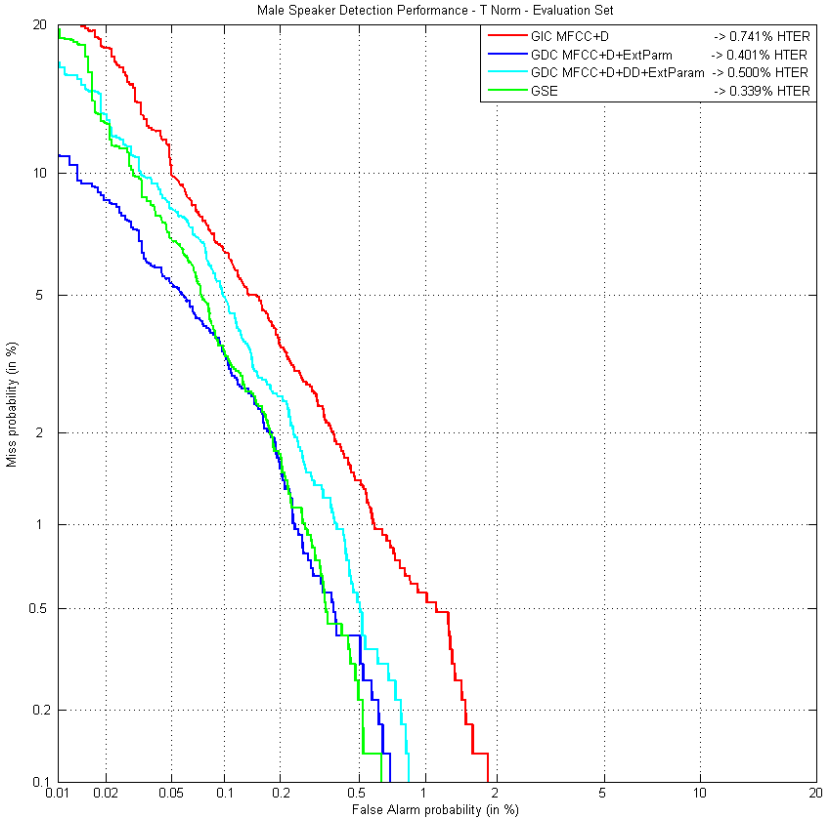



**Figure 5-28** Male's DET curves on the evaluation set from HESPERIA, applying TNorm

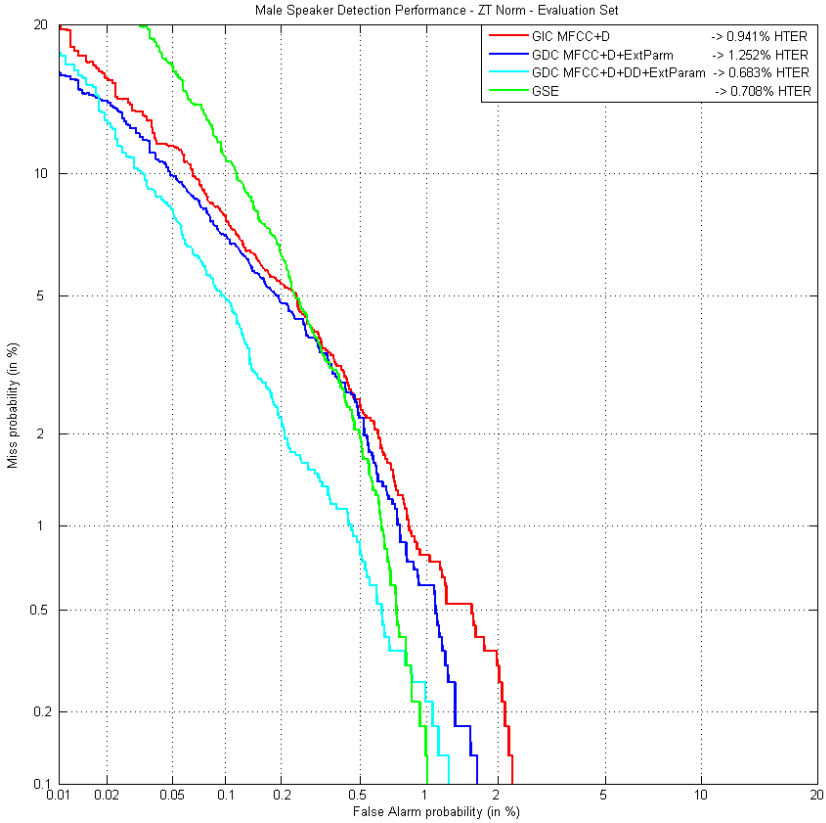



**Figure 5-29** Male's DET curves on the evaluation set from HESPERIA, applying ZTNorm

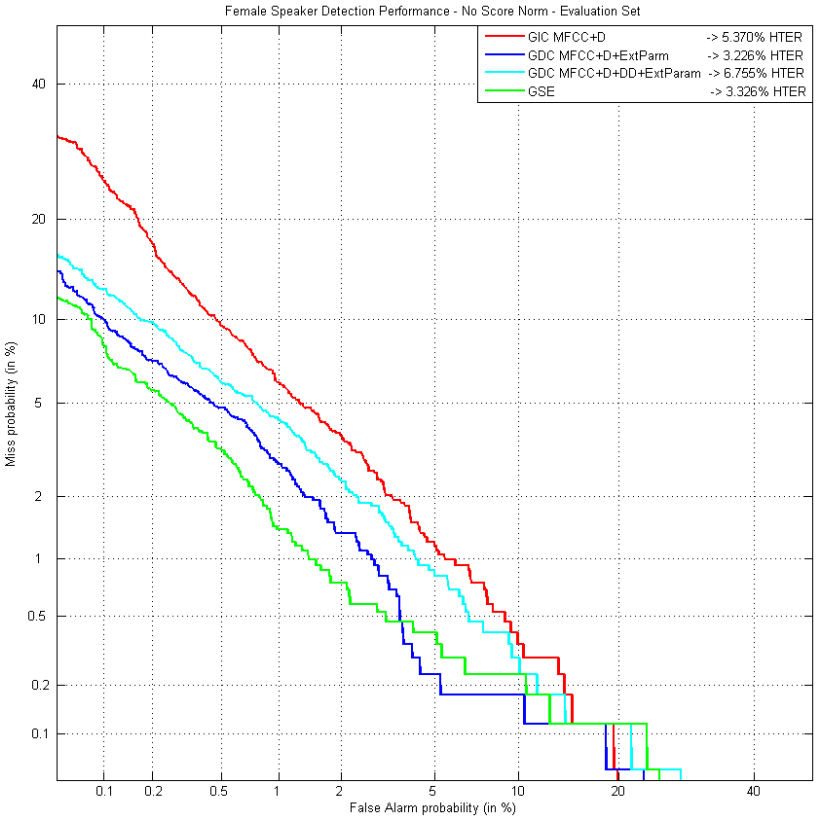



**Figure 5-30** Female's DET curves on the evaluation set from HESPERIA, without applying any score normalisation technique

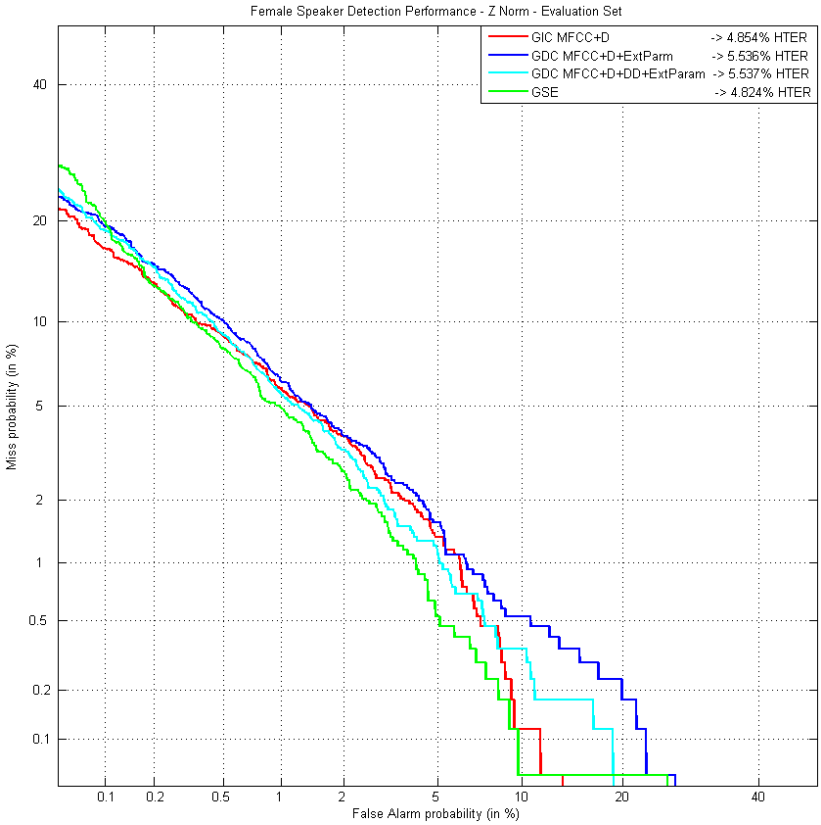



**Figure 5-31** Female's DET curves on the evaluation set from HESPERIA, applying ZNorm

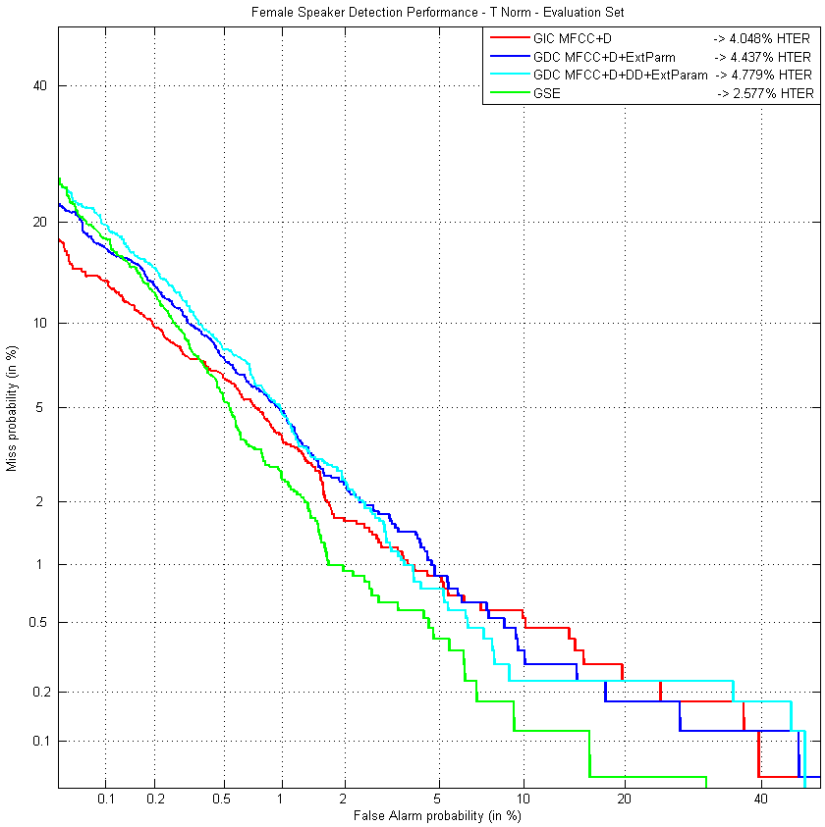



**Figure 5-32** Female's DET curves on the evaluation set from HESPERIA, applying TNorm

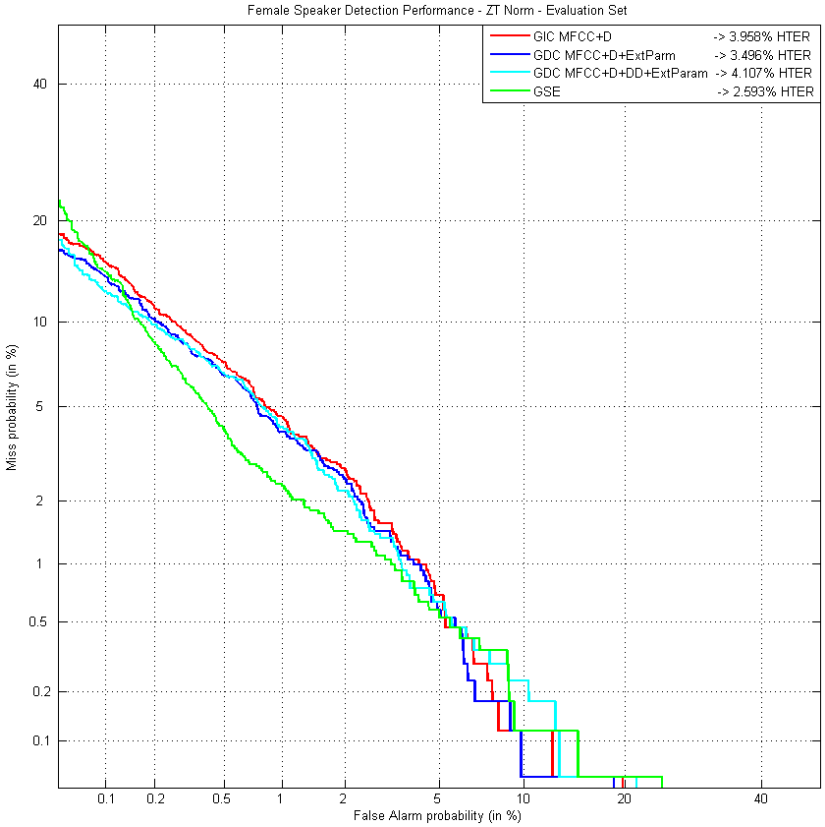



**Figure 5-33** Female's DET curves on the evaluation set from HESPERIA, applying ZTNorm

It must be noted that depending on the score normalisation applied there are some cases in which the DET curve associated with the GSE configuration is not always providing a better performance at all points of the curve. Regarding this fact is worth to note the following: When testing the system in the evaluation set, we have already defined an operation point based on results from the development set. Therefore we are not evaluating the performance of our systems at all points but at the specific one given by $\theta_{dev}$. Furthermore, as stated above, our study is not focused on minimizing *AUC* but *EER* and thus *HTER*. Anyway, it would be desirable that the DET curve provided by the gender-dependent configuration incorporating GSE information should produce better values than the rest of the configurations at all points of the curve.

#### 5.2.1.1.1 Brief Conclussions

First of all, we must point out, that due to the constrains established for this scenario (high-quality recordings in a text-constrained set-up, with channel variability limited to differences in microphones), we are able to achieve recognition rates, in terms of $EER_M$ and $EER_F$, good enough ($EER_X$ < 1%) to ensure that the representation of data under analysis is completely covered with classical parameters in a gender independent configuration. Therefore, improving recognition rates by introducing new parameters or applying scores normalization techniques are difficult to achieve. However, even in this favourable scenario for speaker recognition, the use of a gender-dependent biometric extended parameterisation in which GSE information has been incorporated provides a clear improvement in terms of recognition rates respect to the use of a gender-independent approach based on classical parameters.

A key aspect in the set of tests that have been carried out is the gender-dependent characterization. Based on the results it has been found that the optimal number of coefficients as well as the filter bank used to compute the coefficients is going to be different depending on the gender of the speaker.

Regarding the use of ΔΔ coefficients, it is worth noting that a GDC using MFCCs+Δ allows us to obtain better results in terms of *EER* than a GIC using MFCCs+Δ+ΔΔ coefficients. However, some improvements have been reported when MFCCs+Δ+ΔΔ are applied to a gender-dependent approach on the development set. On the other hand, we have also seen that the improvements obtained with the use of these parameters (ΔΔ) do not remain consistent over the development set and the evaluation set. This means that the good results achieved using MFCCs+Δ+ΔΔ suffer for overtraining on the development set.

Regarding the use of what we have called extra parameters, it must be noted that not all the tested combinations of E, ΔE, F0 and F3, provide a reduction of *HEER*, $EER_M$ or $EER_F$. Moreover, depending on the set of classical parameters used (i.e. MFCCs, MFCCs+Δ, MFCCs+Δ+ΔΔ), and on the gender of the speakers, the combination of extra parameters providing an improvement may be different. However, it has also been confirmed that the new extra parameter proposed, i.e. F3, seems to provide relevant information about speakers, thus helping to increase recognition rates.

The results obtained in terms of $EER_X$ and $HTER_X$, allow us to draw two important conclusions. First of all, we can assert that the use of GSE conveniently parameterised provides an improvement in recognition rates, which remains consistent over the development set and the evaluation set. This information extracted in form of MFCC from the glottal source estimate is used to enhance the MFCC+Δ configuration, i.e. no

ΔΔ are used. Secondly, that in this scenario VTE seems not to provide additional information, besides the already included in the other parameters used.

Finally, the use of score normalization techniques is clearly influenced by the amount of information available for normalization purposes. In this particular scenario the score normalization technique providing best results is also gender-dependent. In this regard, it is worth noting that the selected optimal configuration (in terms of number of MFCCs and channels in the filter bank) is not going to remain the same for all the score normalization techniques.

#### *5.2.1.2 Scenario 2 (mic-tel)*

The procedure that we will follow for this scenario matches the one followed for *Scenario 1*. However, it must be noted that in this case, we are evaluating the performance of the system under an extreme channel variability condition, in which the training phase is carried out using high-quality microphone recordings and test phase is conducted on telephone recordings.

We begin by presenting the results obtained on the development set. Working on the development set allows us to tune the recognition system's background parameters and meta-parameters. As we are facing different conditions than the ones presented in *Scenario 1*, the configuration parameters are likely to be different respect to the settings in *Scenario 1*. Starting again with the *Baseline* front-end in a gender-independent configuration, we are going to verify the usefulness of classical parameters (i.e. MFCCs, MFCCs+Δ, MFCCs+Δ+ΔΔ), the effect of the number of MFCC coefficients and the number of filters on the filter bank used to compute the MFCC. The same method can be applied when using the GDEB front-end, even in the case of not adding extra, GSE or VTE parameters, just applying a gender-dependent configuration on classical parameters. Next, we will test the effect of adding what we have called extra parameters, i.e. Energy, ΔEnergy, Pitch (F0) and the third formant estimate (F3), both on the gender-dependent and gender-independent configurations. Last but not least, we will verify the viability of using the extended-biometric parameters extracted by the GDEB front-end for speaker recognition purposes in this scenario of channel variability.

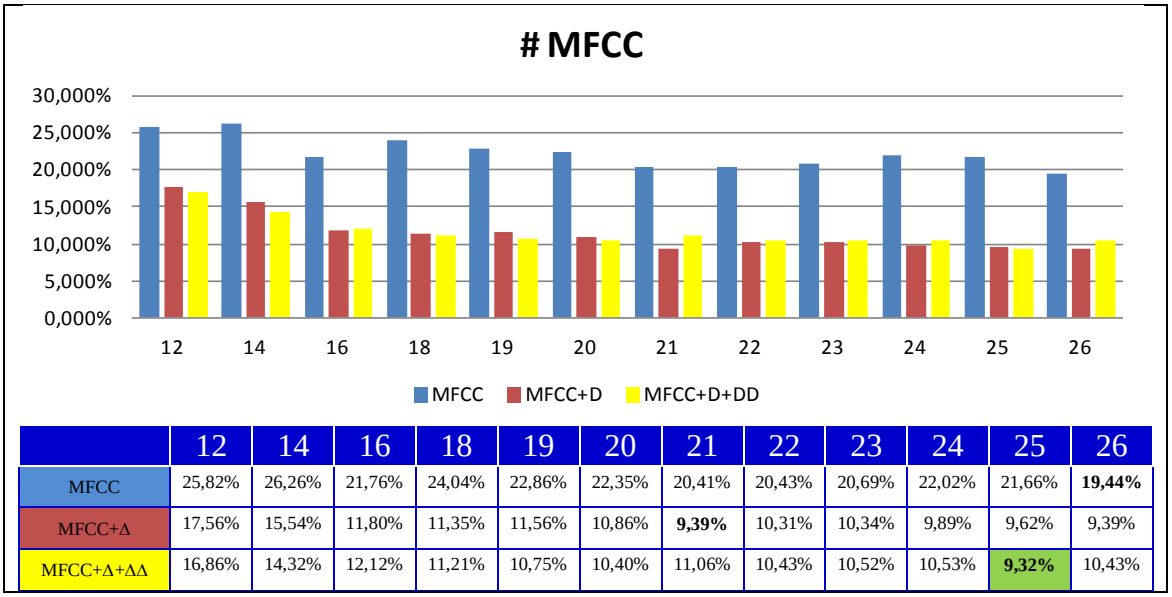
Again, we have excluded a detailed analysis of the number of Gaussians used by the systems, as this number mainly depends on the number of classes to be modelled, which in turn depends on the number and type of parameters used. In the first approach to solve the problem, no score normalisation techniques have been applied, although the influence of these algorithms on the system's performance will be analysed.

Once the configuration settings are fixed for both front-ends (i.e. *Baseline* and GDEB front-ends), and a score threshold is established on the development set, the actual performance of the system will be analysed on the evaluation set.

The first set of figures represents the results obtained in terms of *HEER*, based on each of the configuration parameters established in Eq. (5-7), assuming that we are using the *Baseline* front-end, thus a gender-independent configuration (labelled as GIC), and no score normalisation is applied.

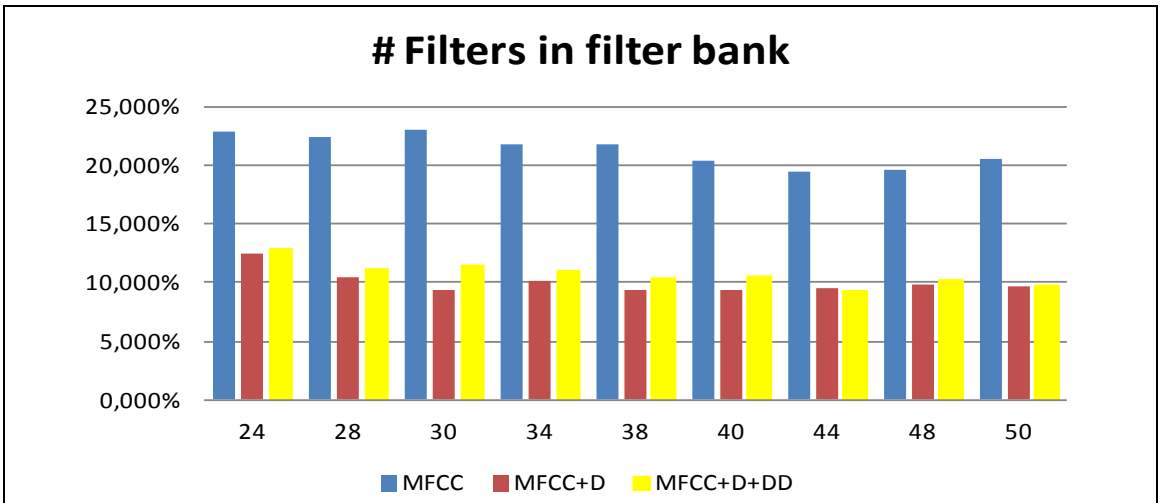
Like in the analysis of the previous scenario, the process followed consists in fixing a value for a specific parameter, run the tests for all the combinations of the remaining configurable parameters, and select the configuration providing better results in terms of *HEER*. In Figure 5-34, we highlight the influence of the number of MFCCs, as well as the influence of the use of Δ and ΔΔ coefficients on the recognition rates. The best

results for each configuration, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ, are marked in bold. The first conclusion that we can draw out from the figure is that the use of MFCCs alone (i.e. without Δ or ΔΔ) is not accurate enough to model speakers, as recognition rates are clearly worse than the ones obtained when complementing the feature vectors with Δ and ΔΔ coefficients. This result is consistent with the one obtained in the *scenario 1*, previously analysed. Although the most successful result is obtained for the configuration that includes ΔΔ coefficients combined with Δ coefficients, with *HEER*<10%, this seems to be an exception, as for the case of using just MFCC+Δ, the *HEER* takes values below 10% for multiple configurations, particularly when the number of MFCCs is 21 or higher than 24.




| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 25,82% | 26,26% | 21,76% | 24,04% | 22,86% | 22,35% | 20,41% | 20,43% | 20,69% | 22,02% | 21,66% | **19,44%** |
| MFCC+Δ | 17,56% | 15,54% | 11,80% | 11,35% | 11,56% | 10,86% | **9,39%** | 10,31% | 10,34% | 9,89% | 9,62% | 9,39% |
| MFCC+Δ+ΔΔ | 16,86% | 14,32% | 12,12% | 11,21% | 10,75% | 10,40% | 11,06% | 10,43% | 10,52% | 10,53% | **9,32%** | 10,43% |

**Figure 5-34** *HEER* obtained depending on the number of MFCCs and the use of Δ and ΔΔ (GIC – development set)



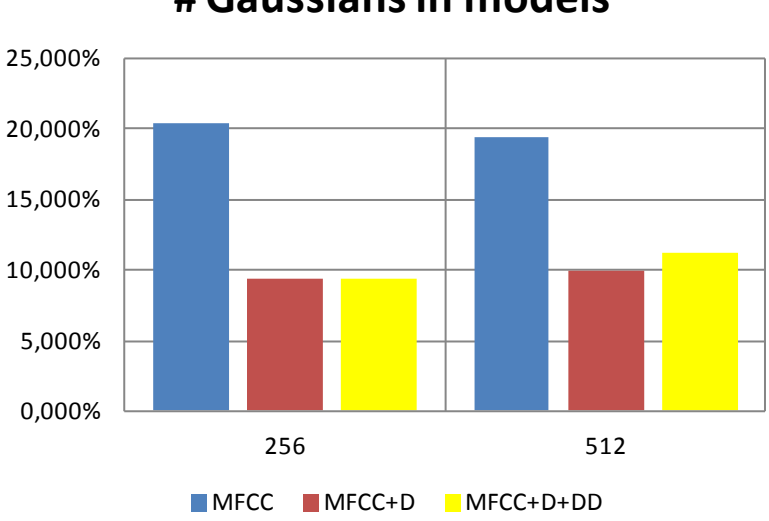

| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 22,92% | 22,41% | 22,93% | 21,83% | 21,71% | 20,41% | **19,44%** | 19,66% | 20,52% |
| MFCC+Δ | 12,42% | 10,48% | **9,39%** | 10,10% | 9,39% | 9,41% | 9,53% | 9,81% | 9,62% |
| MFCC+Δ+ΔΔ | 12,83% | 11,22% | 11,41% | 10,94% | 10,40% | 10,52% | **9,32%** | 10,15% | 9,74% |

**Figure 5-35** *HEER* obtained depending on the number of filters and the use of Δ and ΔΔ (GIC – development set)

Regarding the number of filters, we proceed like in previous analyses. In other words, we fixed a value for *F*, and we test all the combinations for the remaining parameters, selecting the configuration providing lower error rates in terms of *HEER*, for the specific value of *F*. Figure 5-35 provides the results obtained in terms of *HEER*, for the different values of parameter *F*, and for the three tested configurations, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ.

In this case, the best configuration seems to be the one in which the number of filters is 44, for both MFCCs and MFCCs+Δ+ΔΔ configurations, although the highest competitive result is obtained in the case of using MFCCs+Δ when 30 filters are used.

Considering the number of Gaussians used to model the speakers in the gender-independent set-up, we can conclude (see Figure 5-36), that the best results are obtained when 256 Gaussians are used in the case of including Δ or ΔΔ coefficients in the feature vector, while in the case of using just MFCCs alone, the best results are obtained when 512 Gaussians are used. Again, this result is consistent with the corresponding one in *Scenario 1*.




| | MFCC | MFCC+Δ | MFCC+Δ+ΔΔ |
|---|---|---|---|
| 256G | 20,415% | **9,391%** | **9,321%** |
| 512G | **19,447%** | 10,002% | 11,152% |

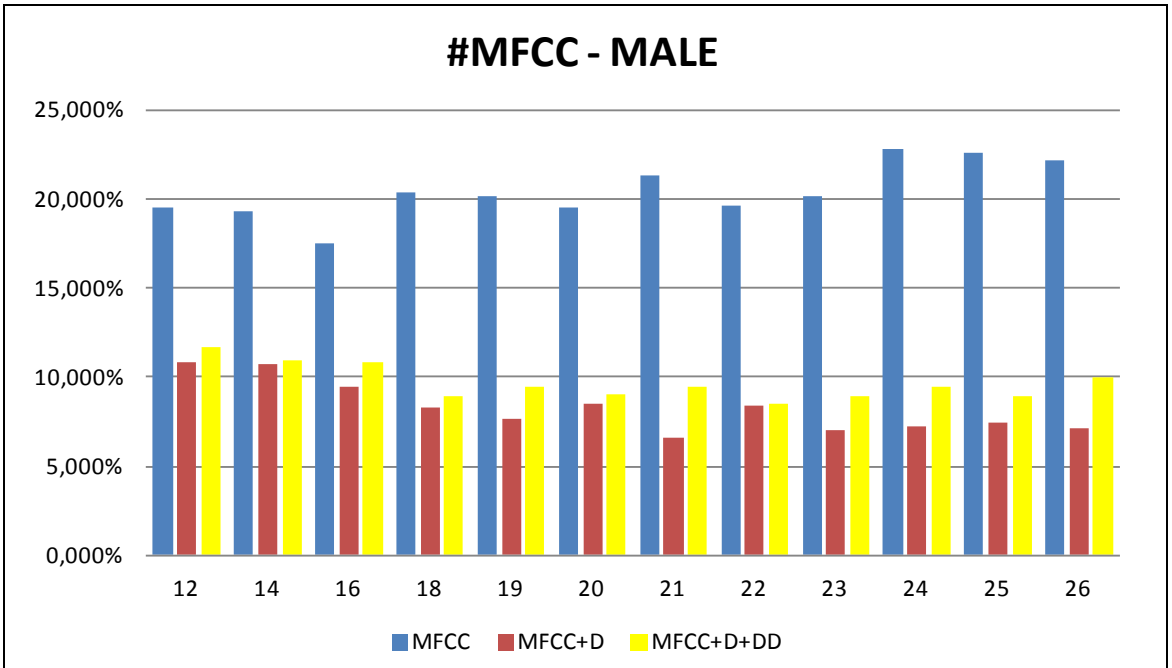
**Figure 5-36** *HEER* obtained depending on the number of Gaussians and the use of Δ and ΔΔ (GIC – development set)

The following table shows the specific configurations providing better recognition rates in terms of *HEER*, for the *Baseline* front-end (therefore in a gender-independent configuration), and for each of the set of feature vector parameters (i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ). Moreover, it also provides the $EER_F$ and $EER_M$ obtained in each case as well as the specific score threshold.

| Parameters | *F* | *G* | *α* | $EER_M$ [$\theta_M$] | $EER_F$ [$\theta_F$] | *HEER* |
|---|---|---|---|---|---|---|
| 26MFCC **(GIC MFCC)** | 44 | 512 | 5 | 22.813% [0.170] | 16.081% [0.189] | 19.447% |
| 21MFCC+Δ **(GIC MFCC+Δ)** | 30 | 256 | 5 | 7.611% [0.054] | 11.170% [0.018] | 9.391% |
| 25MFCC+Δ+ΔΔ **(GIC MFCC+Δ+ΔΔ)** | 44 | 256 | 5 | 8.962% [-0.126] | 9.680% [-0.155] | 9.321% |

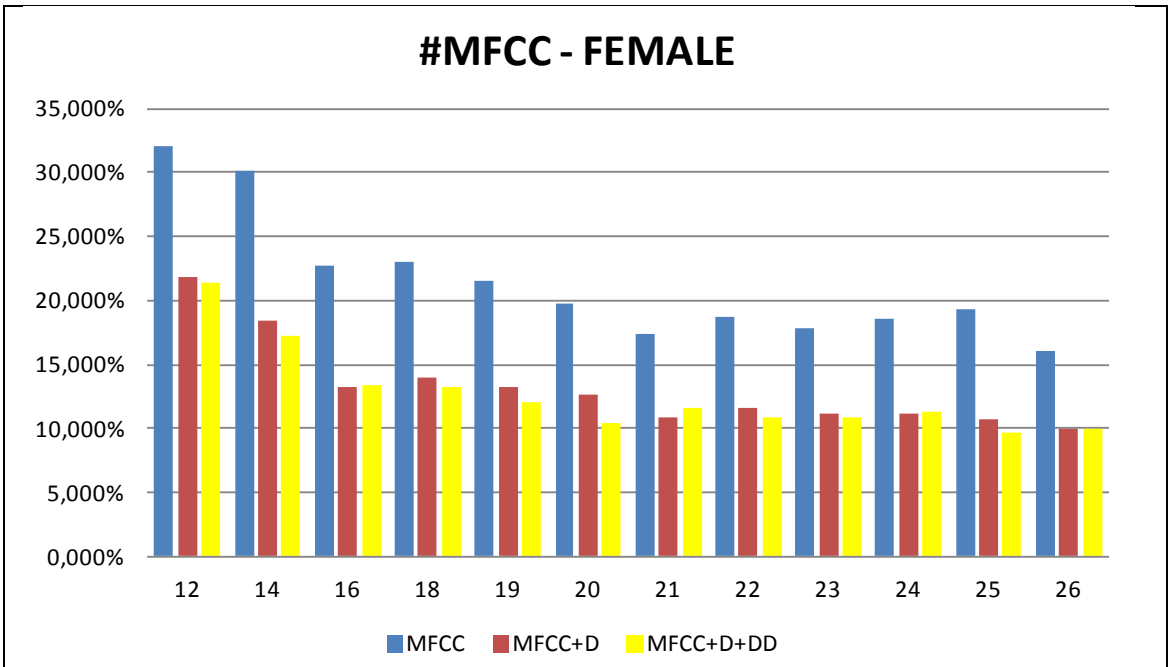
**Table 5-10** Baseline front-end based SR system providing better *HEER* in a gender-independent configuration

So far, we have used the *Baseline* front-end in a gender-independent configuration, thus the objective was to minimise *HEER*, reaching a compromise between $EER_F$ and $EER_M$. However, like in *Scenario 1*, a gender-dependent configuration may provide an improvement in terms of *HEER*, as $EER_F$ and $EER_M$ are minimised independently, even in the case of using just classical features. The same analysis performed in the case of the gender-independent configuration has been applied for the gender-dependent configuration (labelled as GDC). Figure 5-37 to Figure 5-41 provide the results obtained in terms of $EER_X$, where $X = \{F,M\}$, when the number of MFCCs, the number of filters and the number of Gaussians in the model are respectively analysed.



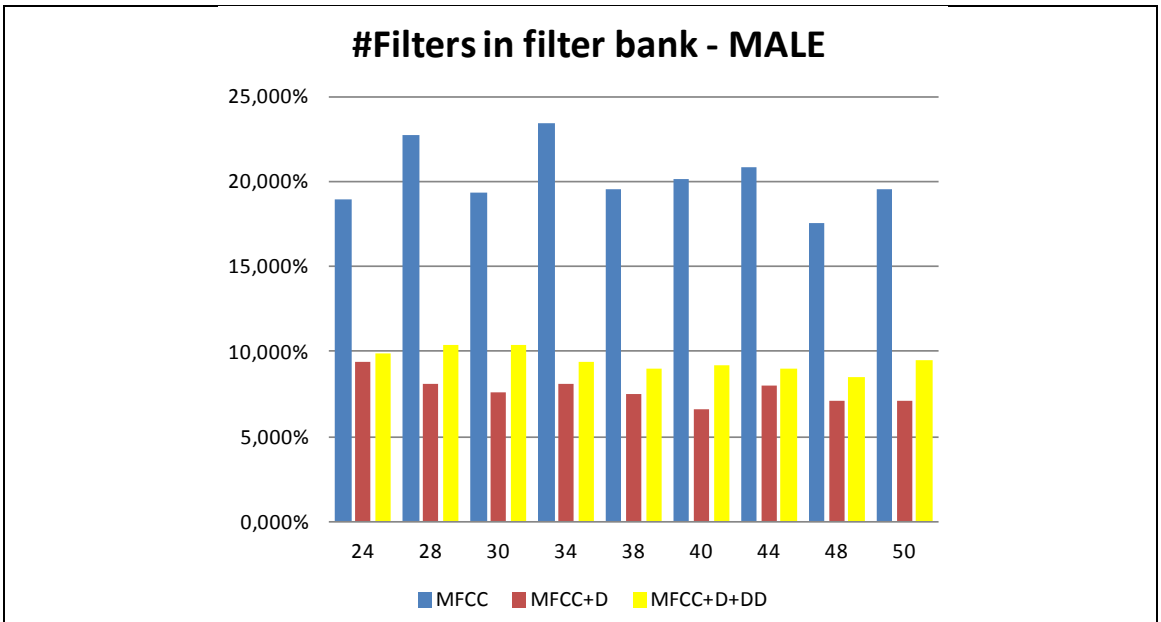

| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 19,5% | 19,3% | **17,5%** | 20,3% | 20,1% | 19,5% | 21,0% | 19,6% | 20,1% | 22,7% | 22,5% | 22,1% |
| MFCC+Δ | 10,80% | 10,77% | 9,43% | 8,25% | 7,71% | 8,48% | **6,64%** | 8,37% | 7,06% | 7,22% | 7,46% | 7,08% |
| MFCC+Δ+ΔΔ | 11,68% | 10,91% | 10,83% | 8,95% | 9,43% | 9,04% | 9,48% | **8,50%** | 8,97% | 9,42% | 8,95% | 10,03% |

**Figure 5-37** $EER_M$ obtained depending on the number of MFCCs and the use of Δ and ΔΔ (GDC – development set)



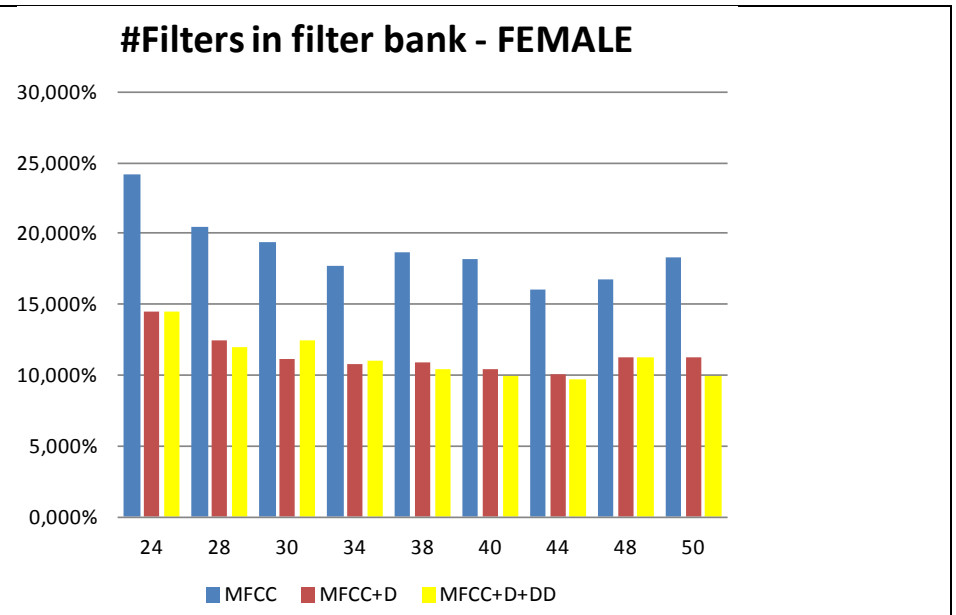

| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 32,1% | 30,1% | 22,6% | 23,0% | 21,5% | 19,8% | 17,3% | 18,8% | 17,8% | 18,5% | 19,3% | **16,0%** |
| MFCC+Δ | 21,9% | 18,4% | 13,1% | 14,0% | 13,3% | 12,6% | 10,9% | 11,6% | 11,2% | 11,2% | 10,7% | **10,0%** |
| MFCC+Δ+ΔΔ | 21,3% | 17,2% | 13,4% | 13,2% | 12,0% | 10,4% | 11,6% | 10,8% | 10,8% | 11,3% | **9,68%** | 9,94% |

**Figure 5-38** $EER_F$ obtained depending on the number of MFCCs and the use of Δ and ΔΔ (GDC – development set)



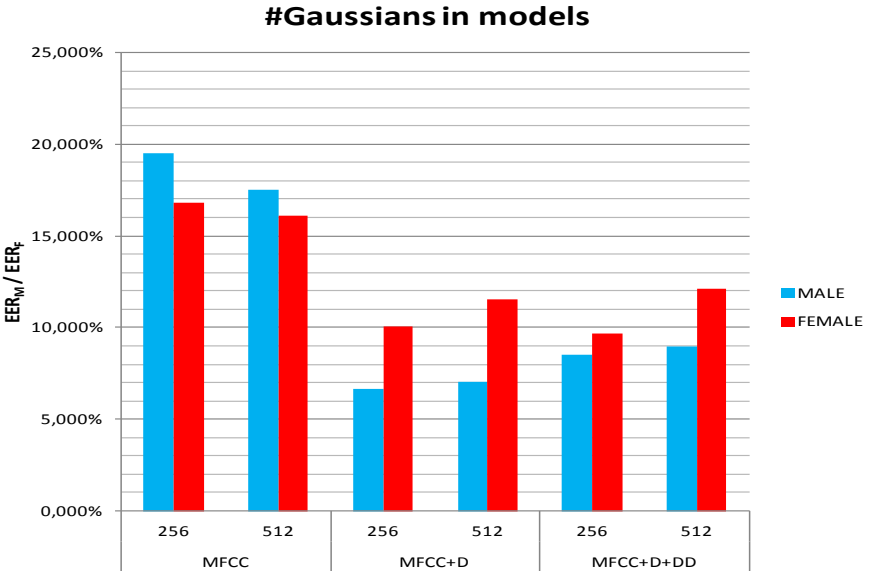

| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 18,95% | 22,75% | 19,35% | 23,44% | 19,53% | 20,14% | 20,84% | **17,51%** | 19,50% |
| MFCC+Δ | 9,40% | 8,07% | 7,56% | 8,13% | 7,49% | **6,64%** | 7,96% | 7,08% | 7,06% |
| MFCC+Δ+ΔΔ | 9,88% | 10,41% | 10,34% | 9,43% | 8,95% | 9,17% | 8,96% | **8,50%** | 9,54% |

**Figure 5-39** $EER_M$ obtained depending on the number of filters and the use of Δ and ΔΔ (GDC – development set)




| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 24,14% | 20,47% | 19,41% | 17,67% | 18,61% | 18,18% | **16,08%** | 16,81% | 18,37% |
| MFCC+Δ | 14,52% | 12,45% | 11,17% | 10,83% | 10,91% | 10,45% | **10,05%** | 11,27% | 11,22% |
| MFCC+Δ+ΔΔ | 14,47% | 12,04% | 12,43% | 11,04% | 10,43% | 9,94% | **9,68%** | 11,24% | 9,94% |

**Figure 5-40** $EER_F$ obtained depending on the number of filters and the use of Δ and ΔΔ (GDC – development set)




| | *G* | MALE - $EER_M$ | FEMALE - $EER_F$ |
|---|---|---|---|
| MFCC | 256 | 19,500% | 16,817% |
| | 512 | **17,517%** | **16,081%** |
| MFCC+Δ | 256 | **6,647%** | **10,057%** |
| | 512 | 7,065% | 11,513% |
| MFCC+Δ+ΔΔ | 256 | **8,501%** | **9,680%** |
| | 512 | 8,973% | 12,132% |

**Figure 5-41** $EER_M$ (blue) and $EER_F$ (red) obtained depending on the number of Gaussians and the use of Δ and ΔΔ (GDC – development set)

From Figure 5-37 and Figure 5-38, it is clear that different numbers of MFCCs are needed to precisely characterise speakers depending on their gender. Specifically, in the case of female speakers, no matter which configuration we are applying, better results are obtained when 25 or 26 MFCCs are used. However, in the case of male speakers the number of MFCCs needs to be lower if we want to precisely model speakers. Particularly, 21 in the case of MFCCs+Δ, and 22 in the case of MFCCs+Δ+ΔΔ. If we compare the results obtained in *Scenario 1* with the ones obtained in this scenario, differences arise on the configuration selected for this parameter, i.e. the number of MFCCs. While for male speakers, we found roughly the same values obtained for *Scenario 1*, for female speakers a radical change occurs, as we move from 16 MFCCs in *Scenario 1* to 25/26MFCCs in this new context.

Regarding the number of filters used in the filter bank to estimate the MFCCs (see Figure 5-39 and Figure 5-40), differences arise again between different genders, although not as significant as in *Scenario 1*. While for female speakers it remains stable regardless the configuration, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ, the number of filters providing better results in terms of $EER_F$ is 44; in the case of male speakers 48 filters provide better results in the case of using MFCCs or MFCCs+Δ+ΔΔ configuration, whereas 40 filters seem to be enough in the case of MFCCs+Δ configuration.

Concerning the number of Gaussians used to build the UBM and the speaker's models (see Figure 5-41), $G$=256 provides more accurate results when using MFCCs+Δ and MFCCs+Δ+ΔΔ configuration.

| Parameters | Gen. | $F$ | $G$ | $\alpha$ | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| 26MFCC **(GIC MFCC)** | M/F | 44 | 512 | 5 | 22.813% [0.170] | - | 16.081% [0.189] | - | 19.44% [-] |
| 16 MFCC **(GDC MFCC)** | M | 48 | 512 | 8 | 17.517% [0.199] | 23.21% | 16.081% [0.189] | - | 16.799% [13.61%] |
| 26 MFCC **(GDC MFCC)** | F | 44 | 512 | 5 | | | | | |
| | | | | | | | | | |
| 21MFCC+Δ **(GIC MFCC+Δ)** | M/F | 30 | 256 | 5 | 7.611% [0.054] | - | 11.170% [0.018] | - | 9.391% [-] |
| 21MFCC+Δ **(GDC MFCC+Δ)** | M | 40 | 256 | 5 | 6.647% [0.056] | 12.68% | 10.057% [-0.007] | | 8.352% [11.06%] |
| 26 MFCC+Δ **(GDC MFCC+Δ)** | F | 44 | 256 | 5 | | | | | |
| | | | | | | | | | |
| 25MFCC+Δ+ΔΔ **(GIC MFCC+Δ+ΔΔ)** | M/F | 44 | 256 | 5 | 8.962% [-0.126] | - | 9.680% [-0.155] | - | 9.321% [-] |
| 22 MFCC+Δ+ΔΔ **(GDC MFCC+Δ+ΔΔ)** | M | 48 | 256 | 5 | 8.501% [-0.102] | 5.14% | 9.680% [-0.155] | 0.0% | 9.091% [2.47%] |
| 25 MFCC+Δ+ΔΔ **(GDC MFCC+Δ+ΔΔ)** | F | 44 | 256 | 5 | | | | | |

**Table 5-11** GDC vs. GIC for HESPERIA's development set on scenario 2

Therefore, we can conclude that in the set of tests carried out in this scenario, using the development set, a gender-dependent parameterisation presents a clear advantage over the gender-independent parameterisation, in terms of *HEER*, even in the case of using just classical parameters. Table 5-11 provides a comparison between the recognition rates obtained by the system, when a GDC or a GIC is used (best results obtained so far are highlighted in light green). Additional columns have been added ($EER_X$ RR), which

provide the relative reduction obtained by GDC, in terms of $EER_X$, respect to the corresponding GIC. The relative reduction in terms of *HERR* respect to GIC has been indicated in brackets in the *HEER* column.

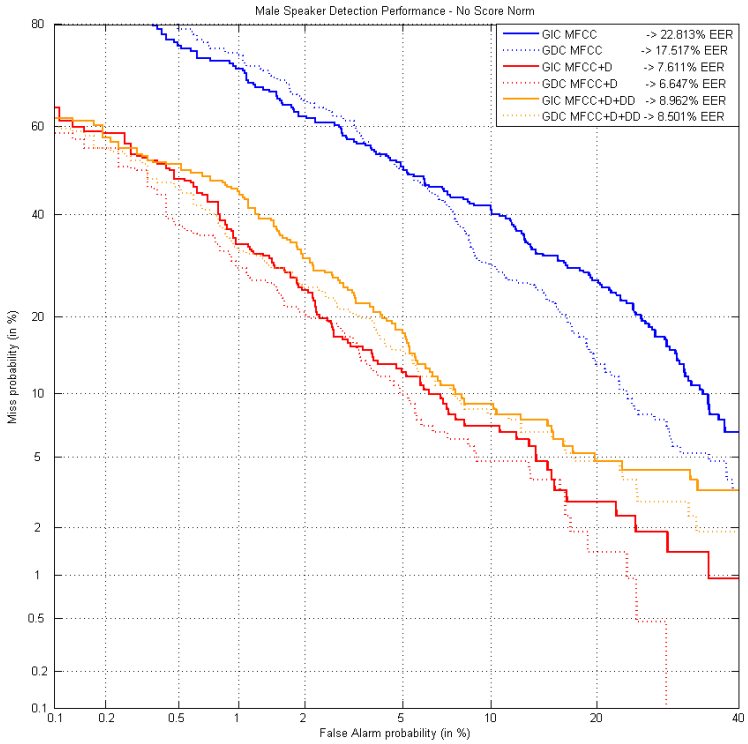



**Figure 5-42** DET curve from classic parameters on HESPERIA male development set for GIC and GDC

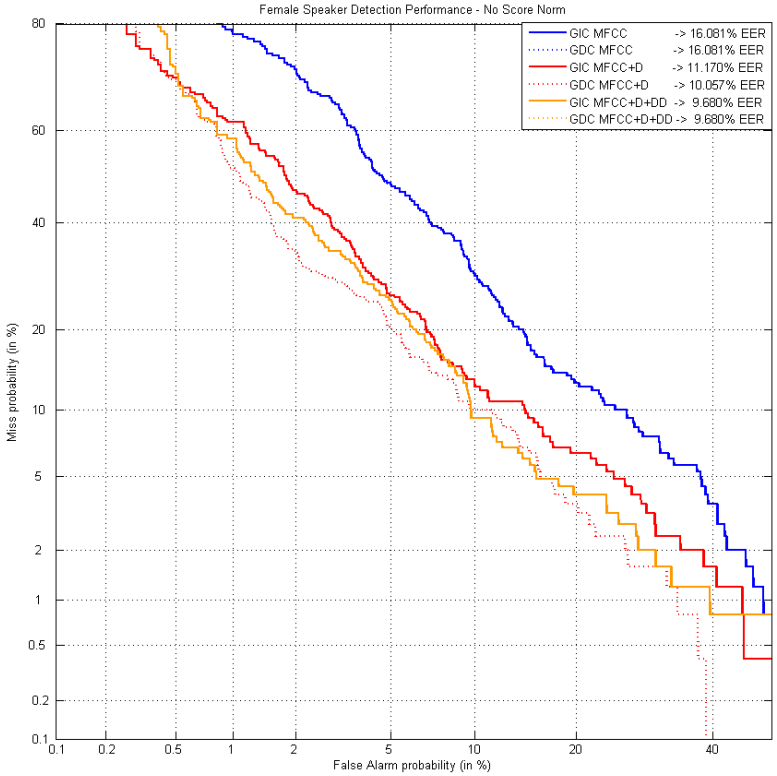



**Figure 5-43** DET curve from classic parameters on HESPERIA female development set for GIC and GDC

If we compare the results shown in Table 5-11 with the ones obtained for *Scenario 1*, it should be noted that for male speakers the best results are obtained when no ΔΔ coefficients are used, while in the case of female speakers, ΔΔ coefficients keep on being the best option.

Figure 5-42 and Figure 5-43 show the DET curves for male and female speakers for both configurations (GIC and GDC), regarding the different sets of classical parameters used so far. Again there are some cases, particularly for female speakers, in which GIC and GDC recognition rates are the same, thus in the DET plots only one curve is going to be represented. It must be reminded that the goal of the test is the reduction of *EER* and not the *Area Under the Curve* (AUC – see Section 1.4), thus it may happen that GIC shows better results than GDC for some of points of the curve. However, GDC will always produce better or at least equal results than GIC in terms of *EER*.

In this scenario, we have also tested the effect of the extra parameters defined in section 5.1.1. We have tested all the combinations of E, ΔE, F0 and F3 with just the GIC and GDC listed in Table 5-11. The most successful results, in terms of *HEER*, obtained on this set of tests are reflected in Table 5-12. According to the results shown in Table 5-12, it must be noted that not all the tested combinations of these extra parameters provide a reduction of *HEER*, $EER_M$ or $EER_F$, if compared with the cases reflected in Table 5-11. For instance, in the case of male speakers when MFCCs+Δ coefficients are used in a GIC configuration we obtained a relative reduction of 1.18% in terms of *HEER*. However, it is at the cost of increasing $EER_M$ from 7.611% to 8.008%. Like in *Scenario 1*, there is not a clear pattern about the best combination of extra parameters providing systematical improvement in recognition rates. Again the configuration of these extra parameters depends on the set of classical parameters used (i.e. MFCC, MFCC+Δ, MFCC+Δ+ΔΔ), and on the gender of the speakers.

| Parameters | Gen. | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER* | *HEER* RR |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC** | M/F | - | 22.813% [0.170] | - | 16.081% [0.189] | - | 19.44% | - |
| **GIC MFCC** | M/F | E+F3 | 18.493% [0.113] | 18.94% | 15.646% [0.107] | 2.71% | 17.069% | 12.23% |
| **GDC MFCC** | M | - | 17.517% [0.199] | 23.21% | 16.081% [0.189] | 0.00% | 16.799% | 13.62% |
| | F | - | | | | | | |
| **GDC MFCC** | M | ΔE | 14.773% [0.168] | 35.24% | 14.433% [0.016] | 10.25% | 14.603% | 24.91% |
| | F | E+ΔE+F3 | | | | | | |
| | | | | | | | | |
| **GIC MFCC+Δ** | M/F | - | 7.611% [0.054] | - | 11.170% [0.018] | - | 9.391% | - |
| **GIC MFCC+Δ** | M/F | E+ΔE+F0 | 8.008% [-0.026] | -5.21% | 10.551% [-0.057] | 5.54% | 9.279% | 1.18% |
| **GDC MFCC+Δ** | M | - | 6.647% [0.056] | 12.68% | 10.057% [-0.007] | 9.96% | 8.352% | 11.06% |
| | F | - | | | | | | |
| **GDC MFCC+Δ** | M | E+ΔE | 6.143% [0.099] | 19.30% | 8.835% [-0.006] | 20.90% | 7.489% | 20.25% |
| | F | E | | | | | | |
| | | | | | | | | |
| **GIC MFCC+Δ+ΔΔ** | M/F | - | 8.962% [-0.126] | - | 9.680% [-0.155] | - | 9.321% | - |
| **GIC MFCC+Δ+ΔΔ** | M/F | E+ΔE+F3 | 8.952% [-0.028] | 0.12% | 8.325% [-0.186] | 14.00% | 8.638% | 7.33% |
| **GDC MFCC+Δ+ΔΔ** | M | - | 8.501% [-0.102] | 5.14% | 9.680% [-0.155] | 0.00% | 9.091% | 2.47% |
| | F | - | | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | E+ΔE+F0 | 7.322% [-0.173] | 18.30% | 8.325% [-0.186] | 14.00% | 7.824% | 16.07% |
| | F | E+ΔE+F3 | | | | | | |

**Table 5-12** $EER_M$, $EER_F$, and *HEER* obtained on the development set when extra parameters are included on the feature vectors for GIC and GDC

Leaving aside the MFCC configuration, it is clear from the results shown in Table 5-12, that the inclusion of the proposed extra parameters into the feature vector produces an extra benefit, helping to improve recognition rates, especially in a gender-dependent configuration. Better results in terms of $EER_M$, $EER_F$, and *HEER* are obtained in all the cases when using a GDC no matter whether MFCC, MFCC+Δ, or MFCC+Δ+ΔΔ parameters are used. Like in *Scenario 1*, parameter F3 proposed in this thesis helps to improve recognition rates particularly in the case of female speakers. We have also tested a configuration that could be considered as state-of-the-art and thus the baseline to beat, namely GIC MFCCs+Δ+ΔΔ+E+ΔE. In terms of *HEER*, the use of GIC MFCCs+Δ+ΔΔ+E+ΔE provides a slight improvement respect to the use of GIC MFCCs+Δ+ΔΔ, and a significant improvement of almost 5% in terms of $EER_M$. However, the improvement is greater when parameter F3 is also incorporated (see Table 5-13). Additionally the configuration referred as state-of-the-art is clearly beaten by the GDC MFCCs+Δ+ExtParam., thus making our proposal more adequate than the classical one.

Regarding the use of ΔΔ coefficients, it must be noted that in terms of $EER_F$, a GDC with extra parameters (in this case E+ΔE+F3) constitutes the configuration showing the most successful results. However, in the case of male speakers, these parameters do not provide any advantage in terms of $EER_M$.

| Parameters | Gen. | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER* | *HEER* RR |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ+ΔΔ** | M/F | - | 8.962% [-0.126] | - | 9.680% [-0.155] | - | 9.321% | - |
| **GIC MFCC+Δ+ΔΔ** | M/F | E+ΔE+F3 | 8.952% [-0.028] | 0.12% | 8.325% [-0.186] | 14.00% | 8.638% | 7.33% |
| **GIC MFCC+Δ+ΔΔ** | M/F | E+ΔE | 8.555% [-0.112] | 4.55% | 10.015% [-0.128] | -3.46% | 9.285% | 0.39% |

**Table 5-13** $EER_M$, $EER_F$, and *HEER* obtained on the development set comparing different GIC MFCC+Δ+ΔΔ configurations

Figure 5-44 and Figure 5-45 respectively show the DET curves for male and female speakers, comparing the most relevant configurations for the present study, i.e., GIC MFCCs+Δ, GIC MFCCs+Δ+ExtParam., GDC MFCCs+Δ+ExtParam., GIC MFCCs+Δ+ΔΔ, GIC MFCCs+Δ+ΔΔ+ ExtParam., GDC MFCCs+Δ+ΔΔ+ExtParam.. In other words, we compare gender dependent with gender independent configurations, while we check as well the influence of ΔΔ coefficients and the extra proposed parameters.

In what follows the results obtained on the development set when the GDEB front-end is used and information extracted from the vocal tract and glottal source estimates are conveniently parameterised. The approach that has been followed, like in the previous scenario, consists in incorporating the extended biometric coefficients into the best gender dependent configuration obtained so far without ΔΔ parameters, in two stages. First we incorporate a set of parameters extracted from the glottal source estimate (labelled as GSE), and once a specific configuration improving previous results is found, we continue incorporating parameters extracted from the vocal tract estimate (labelled as VTE). We proceed this way based on results obtained on *Scenario 1*, as it has been shown that GSE provides more relevant information about speakers than VTE. Additionally, we rule out the use of ΔΔ in this section for two reasons. First of all we believe that the proposed GDEB parameterisation may represent more accurately a speaker than the one that includes ΔΔ coefficients, which are supposed to be more

related to the message transmitted. Secondly, in this particular scenario ΔΔ coefficients do not provide any improvement for male speakers and only a 5% relative reduction in terms of $EER_F$ for female speakers.

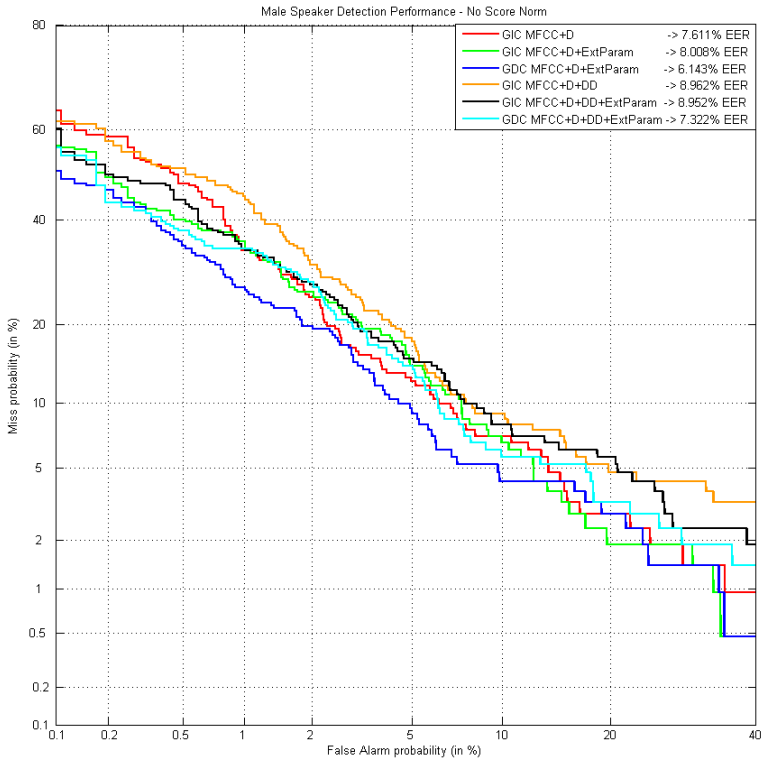



**Figure 5-44** DET curves comparing different sets of parameters under GIC and GDC without extended biometrics on male's development set

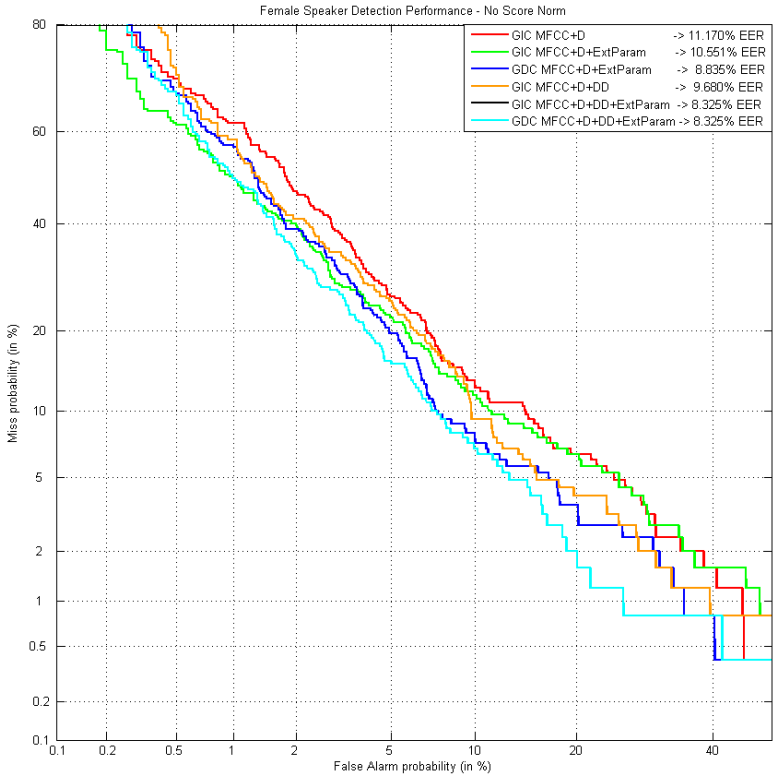



**Figure 5-45** DET curves comparing different sets of parameters under GIC and GDC without extended biometrics on female's development set

| Parameters | Gen. | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | - | - | 7.611% [0.054] | - | 11.170% [0.018] | - | 9.391% [-] |
| **GDC MFCC+Δ** | M | - | E+ΔE | 6.143% [0.099] | 19.30% | 8.835% [-0.006] | 20.90% | 7.489% [20.25%] |
| | F | - | E | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | - | E+ΔE+F0 | 7.322% [-0.173] | 3.80% | 8.325% [-0.186] | 25.47% | 7.824% [16.39%] |
| | F | - | E+ΔE+F3 | | | | | |
| **GSE** | M | **Source-Tract Sep. Alg:** Prediction Order: 20 Forgetting Factor:0.995 **GSE:** 10-Channel Filter bank 2 MFCC | E+ΔE | 5.199% [0.006] | 31.69% | 8.124% [-0.059] | 27.27% | 6.662% [29.06%] |
| | F | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor:0.995 **GSE:** 32-Channel Filter bank 6 MFCC | E | | | | | |

**Table 5-14** $EER_M$, $EER_F$, and *HEER* obtained on the development set (no score normalisation), comparing classical parameters with extra parameters and extended biometric parameters

Again multiple configurations have been tested, regarding the multiple variables that can be tuned in the GDEB front-end, Table 5-14 shows the ultimate configurations chosen for each gender, as well as the recognition rates obtained in each case in terms of $EER_M$, $EER_F$ and *HEER*. Additionally, the relative reduction (RR) in terms of $EER_X$ and *HEER*, compared to GIC MFCC+Δ configuration is also provided.

DET curves corresponding to the results presented in Table 5-14 are depicted in Figure 5-46 for male speakers and Figure 5-47 for female speakers. Clearly, the parameterisation generated by the GDEB front-end, in this case just including information from the glottal source estimate in the form of mel-frequency cepstrum coefficients, is the one providing best results on the development set for male and female speakers. However, in the case of female speakers this improvement only reaches 2.5% respect to GDC MFCCs+Δ+ΔΔ+ExtParam.. Like in *Scenario 1*, the number of filters used to compute MFCCs from the GSE component is greater in the case of female speakers than in the case of male speakers.

Although in the case of male speakers for most operating points (defined by the DET curve) the GSE configuration provides better results than any other configuration tested so far, the female case is somewhat different. As we have already pointed out, the improvement, in terms of $EER_F$, obtained by the GSE configuration respect to GDC MFCCs+Δ+ΔΔ+ExtParam. is quite small, therefore DET curves are going to be tightly packed as depicted in Figure 5-47. However, we must not forget that the objective is to minimise *EER* and not *AUC*, as previously commented in *Scenario 1*, and in this sense, the use of the GSE configuration provides better recognition rates than any other configuration tested so far for female speakers.

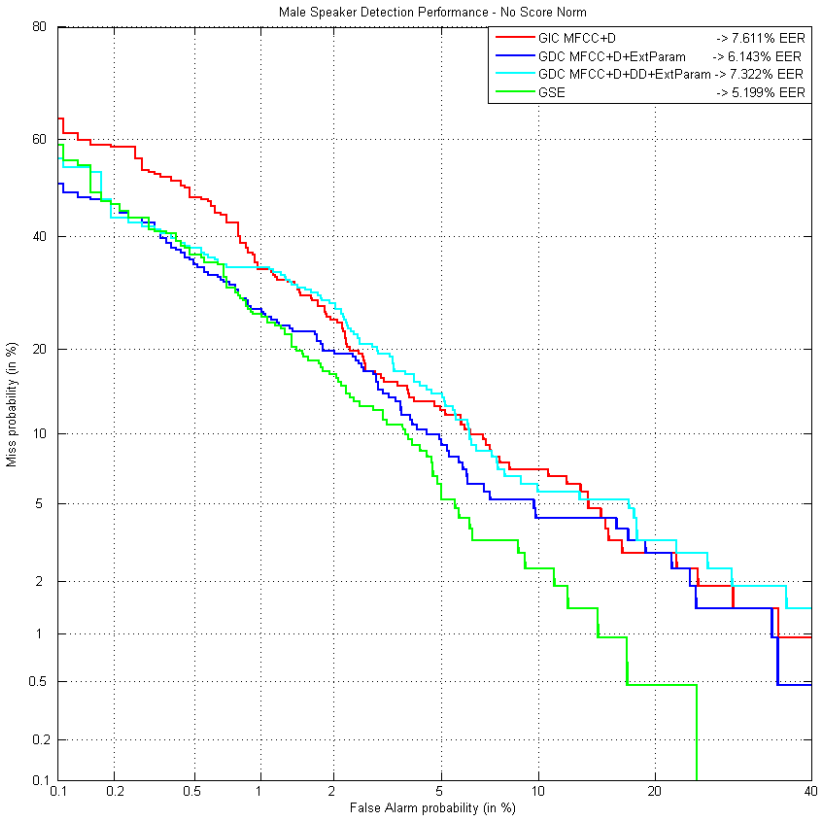



**Figure 5-46** DET curves comparing classical parameters and GDEB on HESPERIA development set for male speakers

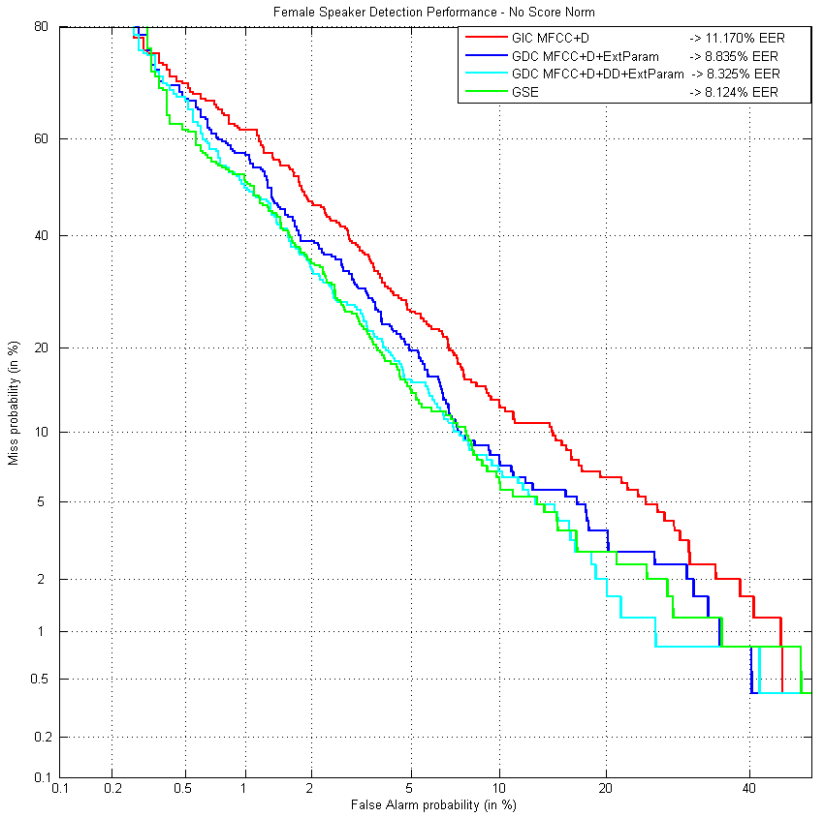



**Figure 5-47** DET curves comparing classical parameters and GDEB on HESPERIA development set for female speakers

Before incorporating additional parameters from the vocal tract estimate into the feature vectors, it seems necessary to make some observations on the use of the parameters extracted from the glottal source estimate. So far we have selected specific configurations providing the best performance in terms of $EER_X$ (i.e. no matter whether male or female case) and *HEER*. This means that in theory no better results on development set can be obtained using the *Baseline* or the GDEB front-ends. However, it is necessary to verify if the improvements derived from incorporating GSE information is highly dependent on the specific configuration or is systematically obtained.

Like in the previously presented scenario, when we deal with male speakers (see Figure 5-48) the use of GSE parameters under different configurations systematically provides a reduction in terms of $EER_M$ respect to all the configurations tested so far. The green solid line represents the minimum $EER_X$ (y-axis) obtained when GSE is incorporated into the feature vector in the form of MFCCs. Different number of $MFCCs_G$={2,4,6,8,10} have been tested, which have been computed applying a filter bank with different number of filters $F_G$=[4...38] (x-axis). Each point in the x-axis represents the minimum *EER* obtained for a specific value of $F_G$, regardless $MFCCs_G$ value.

Figure 5-49 provides the same information for female speakers. In this case, the improvement in terms of $EER_F$ is systematically obtained if compared to GIC MFCCs+Δ and GDC MFCCs+Δ+ExtParam.. However, the $EER_F$ obtained by GDC MFCCs+Δ+ΔΔ+ExtParam. is difficult to beat.

After analyzing the two proposed scenarios, we can conclude that female speakers pose a greater challenge than male speakers as far as improving recognition rates. Moreover,

the use of parameters derived from the GSE offers a clear advantage in the case of male speakers.

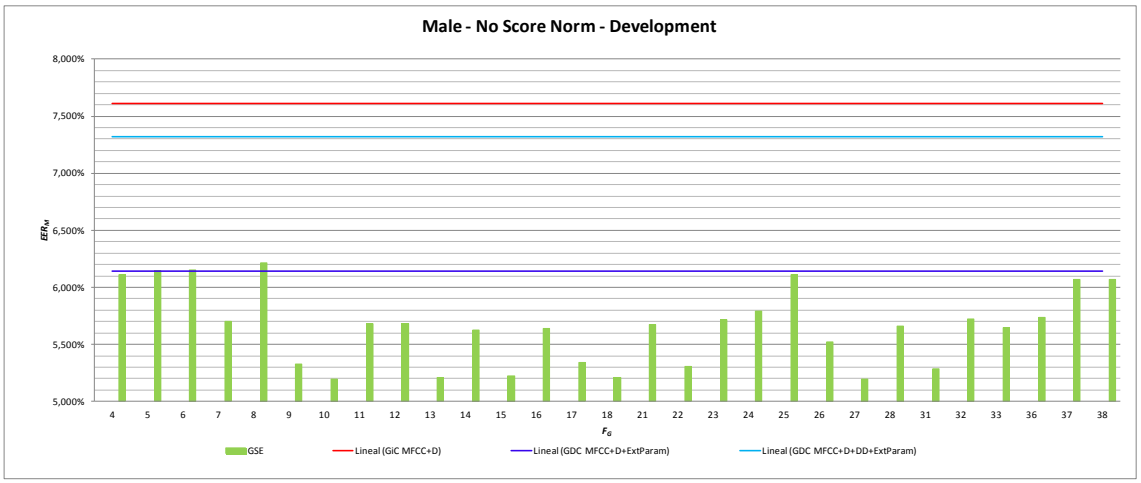



**Figure 5-48** Influence of GSE configuration on the $EER_M$ (development set)

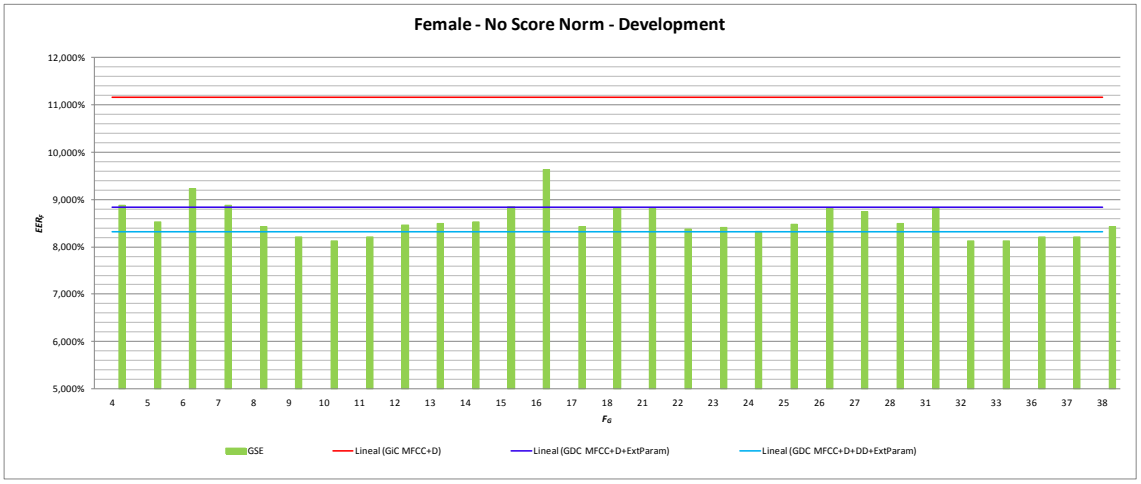



**Figure 5-49** Influence of GSE configuration on the $EER_F$ (development set)

In this scenario, we have also tested the influence of score normalisation algorithms in recognition rates. Therefore, the experiments previously reported have been conducted applying ZNorm, TNorm, and ZTNorm. Table 5-15 to Table 5-17 provide the equivalent results that have been reported in Table 5-14, when the different score normalisations are applied.

It must be noted that Table 5-15 to Table 5-17 include an additional column providing the set up for classical parameters, i.e. MFCCs+Δ., MFCCs+Δ+ΔΔ, regarding the number of filters in filter bank and the number of MFCCs. This is interesting in order to show that the use of different types of score normalisation influences the values that should be assigned to the configurable parameters of the *Baseline* front-end in the search for the minimum *EER*. In other words, we cannot expect the set of selected parameters be the most successful one when no score normalisation is applied, to provide also the best results under different score normalisations.

From the results shown on Table 5-15, Table 5-16, and Table 5-17, it is clear that no matter the score normalisation applied, the most successful results, in terms of $EER_{M,F}$ and *HEER* are always obtained when GSE is included as MFCCs in the feature vector. Additionally, a gender-dependent characterisation of speakers provides better results in

terms of recognition rates that a gender-independent characterisation. Anyway, there are two additional aspects that must be highlighted. The use of score normalisation algorithms provides a clear advantage in the case of male speakers (particularly in the case of TNorm). In the case of female speakers there is no improvement respect to the case in which no normalisation is applied. The second aspect that it is worth noting is the fact that when applying ZNorm or TNorm, classical parameters present the same setup for male and female speakers regarding F and MFCC parameters. However, the actual configuration is different as the MAP adaptation coefficient is different for each gender and extra parameters are also different.

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=48, *MFCC*=25, *G*=256, *α*=10 | - | - | 8.319% [2.223] | - | 11.370% [2.080] | - | 9.845% [-] |
| **GDC MFCC+Δ** | M | *F*=48, *MFCC*=25, *G*=256, *α*=20 | - | E+ΔE | 5.199% [2.233] | 37.50% | 9.237% [2.076] | 18.76% | 7.218% [26.68%] |
| | F | *F*=48, *MFCC*=25, *G*=256, *α*=5 | - | E+ΔE | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=48, *MFCC*=22, *G*=512, *α*=16 | - | E+ΔE | 8.008% [2.294] | 3.74% | 9.664% [1.998] | 15.01% | 8.836% [10.25%] |
| | F | *F*=44, *MFCC*=25, *G*=256=, *α*=5 | - | E+ΔE+F3 | | | | | |
| **GSE** | M | *F*=48, *MFCC*=25, *G*=512, *α*=20 | **Source-Tract Sep. Alg:** Prediction Order: 29 Forgetting Factor: 0.995 **GSE:** 26-Channel Filter bank 2 MFCC | E+ΔE | 3.924% [2.496] | 52.84% | 8.919% [2.097] | 21.56% | 6.421% [34.77%] |
| | F | *F*=48, *MFCC*=25, *G*=256, *α*=10 | **Source-Tract Sep. Alg:** Prediction Order:10 Forgetting Factor: 0.995 **GSE:** 31-Channel Filter bank 4 MFCC | E+ΔE | | | | | |

**Table 5-15** $EER_M$, $EER_F$, and *HEER* obtained on the development set (ZNorm), comparing classical parameters with extra parameters and extended biometric parameters

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=38, *MFCC*=21, *G*=256, *α*=5 | - | - | 8.019% [2.413] | - | 10.450% [1.957] | - | 9.235% [-] |
| **GDC MFCC+Δ** | M | *F*=38, *MFCC*=21, *G*=256, *α*=8 | - | E+ΔE+F3 | 6.239% [2.416] | 22.19% | 9.630% [1.967] | 7.85% | 7.935% [14.08%] |
| | F | *F*=38, *MFCC*=21, *G*=256, *α*=8 | - | E | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=34, *MFCC*=19, *G*=256, *α*=5 | - | E+ΔE | 9.069% [2.133] | -13.10% | 9.287% [2.097] | 11.13% | 9.178% [0.61%] |
| | F | *F*=50, *MFCC*=25, *G*=256, *α*=5 | - | E+ΔE | | | | | |
| **GSE** | M | *F*=38, *MFCC*=21, *G*=256, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 32 Forgetting Factor: 0.995 **GSE:** 18-Channel Filter bank 8 MFCC | E+ΔE+F3 | 5.189% [2.378] | 35.29% | 8.392% [2.124] | 19.70% | 6.790% [26.47%] |
| | F | *F*=38, *MFCC*=21, *G*=256, *α*=10 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 21-Channel Filter bank 8 MFCC | E | | | | | |

**Table 5-16** $EER_M$, $EER_F$, and *HEER* obtained on the development set (TNorm), comparing classical parameters with extra parameters and extended biometric parameters

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=38, *MFCC*=19, *G*=256, *α*=8 | - | - | 7.665% [1.361] | - | 11.253% [0.915] | - | 9.459% [-] |
| **GDC MFCC+Δ** | M | *F*=48, *MFCC*=20, *G*=512, *α*=5 | - | E+ΔE | 6.132% [1.680] | 20.00% | 9.722% [1.372] | 13.61% | 7.927% [16.20%] |
| | F | *F*=38, *MFCC*=21, *G*=512, *α*=5 | - | E+F0 | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=34, *MFCC*=19, *G*=256, *α*=10 | - | ΔE | 8.491% [1.252] | -10.77% | 10.040% [0.996] | 10.78% | 9.265% [2.05%] |
| | F | *F*=50, *MFCC*=26, *G*=256, *α*=5 | - | - | | | | | |
| **GSE** | M | *F*=48, *MFCC*=20, *G*=512, *α*=20 | **Source-Tract Sep. Alg:** Prediction Order: 29 Forgetting Factor: 0.995 **GSE:** 17-Channel Filter bank 10 MFCC | E+ΔE | 4.717% [2.191] | 38.46% | 8.568% [1.039] | 23.87% | 6.642% [29.78%] |
| | F | *F*=38, *MFCC*=21, *G*=256, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 13 Forgetting Factor: 0.995 **GSE:** 23-Channel Filter bank 10 MFCC | E+F0 | | | | | |

**Table 5-17** $EER_M$, $EER_F$, and *HEER* obtained on the development set (ZTNorm), comparing classical parameters with extra parameters and extended biometric parameters

In order to complete previous results, Figure 5-50 and Figure 5-51 provide DET curves for male and female speakers, respectively, so the performance of the *Baseline* front-end (in a GIC MFCCs+Δ setup) can be compared with the GDEB front-end when different score normalisation techniques are applied.

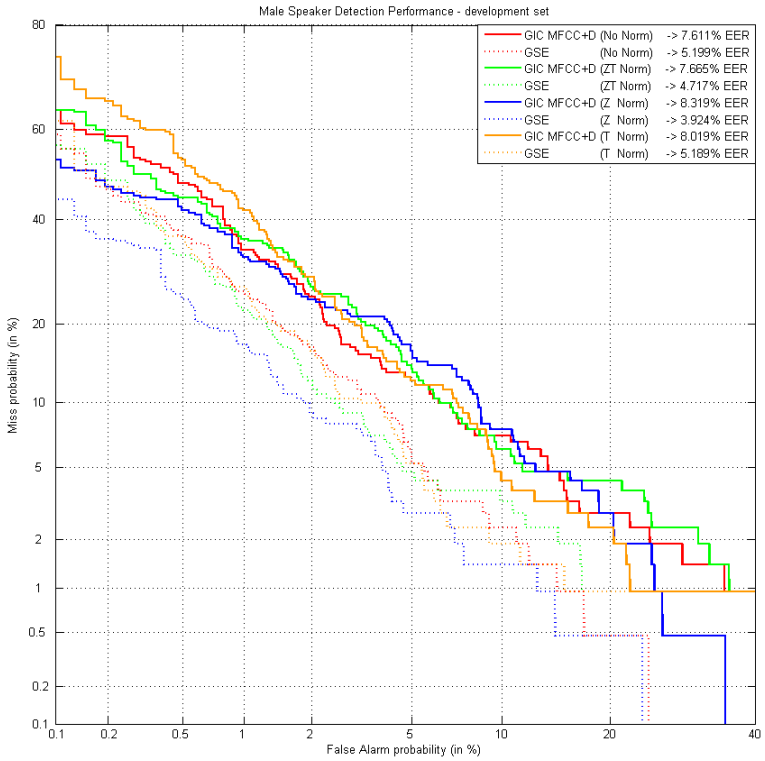



**Figure 5-50** DET curves for *Baseline* and GDEB front-end, applying different score normalisation techniques (male's development set)

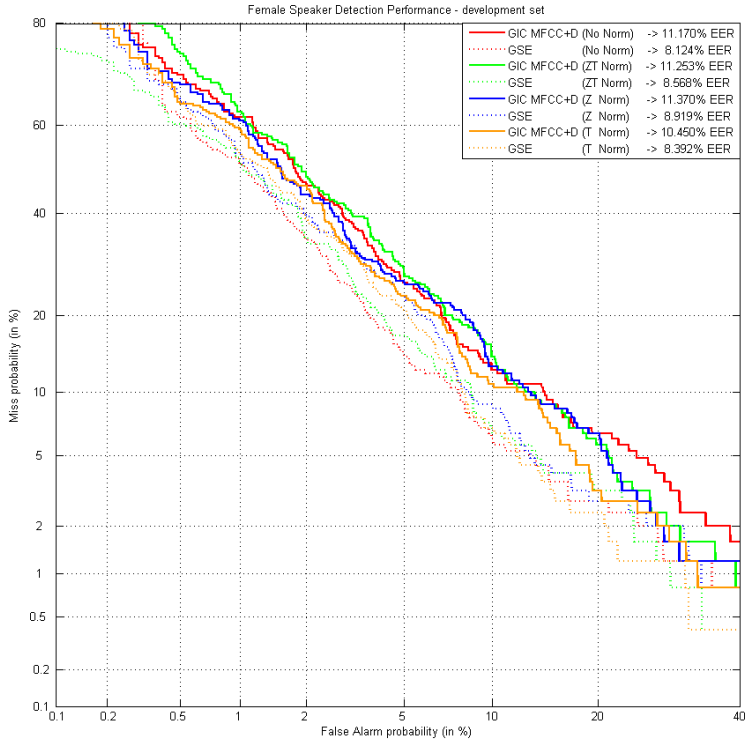



**Figure 5-51** DET curves for *Baseline* and GDEB front-end, applying different score normalisation techniques (female's development set)

Like in *Scenario 1*, there are still two additional tasks to be carried out. First of all we have to verify the usefulness of the information provided by the vocal tract estimate (labelled as VTE), which we have not yet included in the experiments carried out so far. To limit the experiments, we only run an additional test based on the configuration providing better results in terms of $EER_X$, i.e. GDC GSE No Score Norm for female speakers and GDC GSE ZNorm for male speakers.

Like in the test run on Scenario 1, the use of GSE information combined with VTE information in a gender-dependent setup, systematically produces recognition rates that outperform the ones obtained by GIC MFCCs+Δ and in the case of male speakers the ones obtained by GDC MFCCs+Δ+ΔΔ+ExtParam.. However, it is difficult to find a configuration that outperforms the recognition rates produced when using GSE information alone. Therefore, in this text-constrained scenario, VTE do not seem to provide additional information, neither for male nor for female speakers.

Once we have found the configurations that allow us to obtain the most successful results in terms of $EER_X$ on the development set, using both the *Baseline* and the GDEB front-ends, the final step consists in verifying if the behaviour of the speaker recognition system with the selected configurations holds for the evaluation set, i.e. for unknown data, or if the results are affected by overtraining on the development set. Table 5-18 provides the results obtained on the evaluation set applying the different configurations previously selected (see Table 5-14 to Table 5-17) for different score normalisations. Most successful results for the different score normalisation techniques applied are highlighted in green.

The results obtained in terms of $HTER_X$, allow us to draw two important conclusions. First of all, we can assert that the use of GSE conveniently parameterised provides an improvement in recognition rates, which remains consistent over the development set and the evaluation set. Specifically, for the male case when applying ZNorm we obtained a relative reduction of 53% in terms of $EER_M$ (from $EER_M$ =8.319%, when classical gender-independent characterisation is used, to $EER_M$ =3.924%). When moving into the evaluation set, the same configuration obtains a relative reduction of 31% in terms of $HTER_M$, (from $HTER_M$ =8.374%, when classical gender-independent characterisation is used to $HTER_M$ =5.776%). For female speakers, when no score normalisation is applied a relative reduction of 27% in terms of $EER_F$ is obtained (from $EER_F$ =11.170%, when classical gender-independent characterisation is used to $EER_F$ =8.124%). Correspondingly, a relative reduction of 35% in terms of $HTER_F$ is obtained when moving into the evaluation set (from $HTER_F$ =13.895%, when classical gender-independent characterisation is used to $HTER_M$ =8.974%). Thus the results obtained in *Scenario 1* appear to be consistent with those obtained in this scenario. Regarding the use of ΔΔ coefficients it must be noted that in any of the tested configurations, their use provides an improvement in terms of $EER_{F,M}$, with respect to the use of GSE information.

| Score Norm | Parameters | $EER_M$ [$\theta_M$] | $HTER_M$ | $HTER_M$ RR | $EER_F$ [$\theta_F$] | $HTER_F$ | $HTER_F$ RR |
|---|---|---|---|---|---|---|---|
| No Norm | **GIC MFCC+Δ** | 7.611% [0.054] | 9.920% | - | 11.170% [0.018] | 13.895% | - |
| | **GDC MFCC+Δ+ExtParm.** | 6.143% [0.099] | 5.598% | 43.57% | 8.835% [-0.006] | 12.039% | 13.35% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 7.322% [-0.173] | 7.494% | 24.45% | 8.325% [-0.186] | 9.737% | 29.92% |
| | **GSE** | 5.199% [0.006] | 4.920% | 50.41% | 8.124% [-0.059] | 8.974% | 35.42% |
| ZNorm | | | | | | | |
| | **GIC MFCC+Δ** | 8.319% [2.223] | 8.374% | - | 11.370% [2.080] | 12.842% | - |
| | **GDC MFCC+Δ+ ExtParm.** | 5.199% [2.233] | 6.098% | 27.18% | 9.237% [2.076] | 12.263% | 4.51% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 8.008% [2.294] | 9.063% | -8.24% | 9.664% [1.998] | 10.697% | 16.70% |
| | **GSE** | 3.924% [2.496] | 5.776% | 31.02% | 8.919% [2.097] | 9.961% | 22.44% |
| TNorm | | | | | | | |
| | **GIC MFCC+Δ** | 8.019% [2.413] | 9.282% | - | 10.450% [1.957] | 10.658% | - |
| | **GDC MFCC+Δ+ ExtParm.** | 6.239% [2.416] | 5.609% | 39.57% | 9.630% [1.967] | 9.829% | 7.78% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 9.069% [2.133] | 8.305% | 10.53% | 9.287% [2.097] | 9.961% | 6.54% |
| | **GSE** | 5.189% [2.378] | 4.207% | 54.67% | 8.392% [2.124] | 8.961% | 15.93% |
| ZTNorm | | | | | | | |
| | **GIC MFCC+Δ** | 7.665% [1.361] | 9.477% | - | 11.253% [0.915] | 12.605% | - |
| | **GDC MFCC+Δ+ ExtParm.** | 6.132% [1.680] | 6.977% | 26.38% | 9.722% [1.372] | 11.789% | 6.47% |
| | **GDC MFCC+Δ+ΔΔ+ ExtParm.** | 8.491% [1.252] | 9.448% | 0.30% | 10.040% [0.996] | 13.763% | -9.19% |
| | **GSE** | 4.717% [2.191] | 4.943% | 47.85% | 8.568% [1.039] | 9.816% | 22.13% |

**Table 5-18** $HTER_X$ produced for selected configurations on evaluation set, applying different score normalisations

Finally, Figure 5-52 to Figure 5-59, show the DET curves that represent the results obtained in the evaluation set for male and female speakers, which are reflected in Table 5-18.

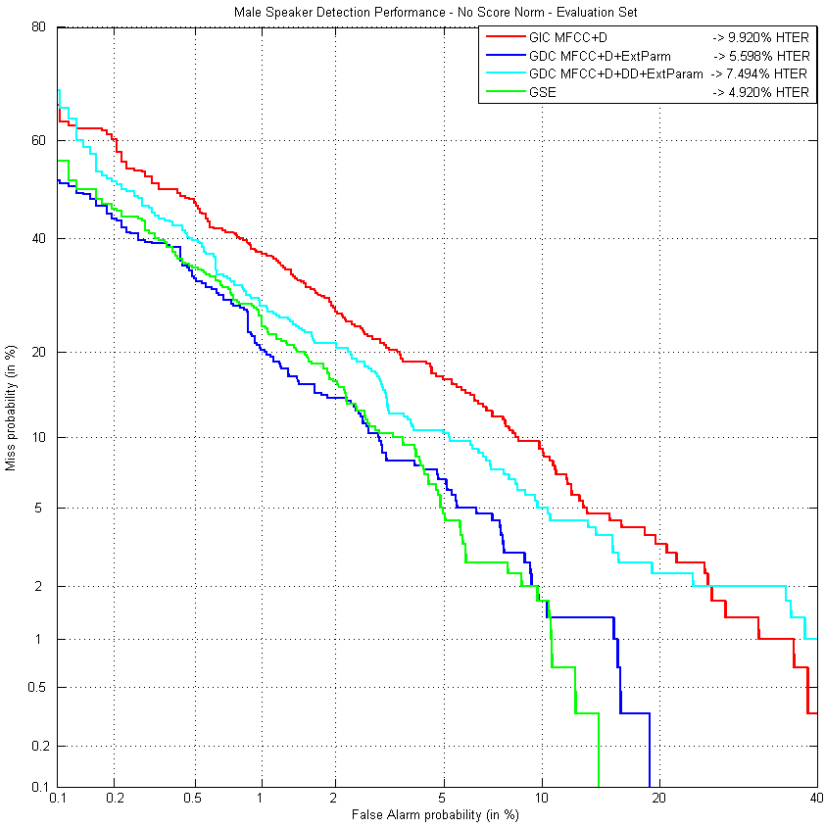



**Figure 5-52** Male's DET curves on HESPERIA evaluation set, without applying any score normalisation technique

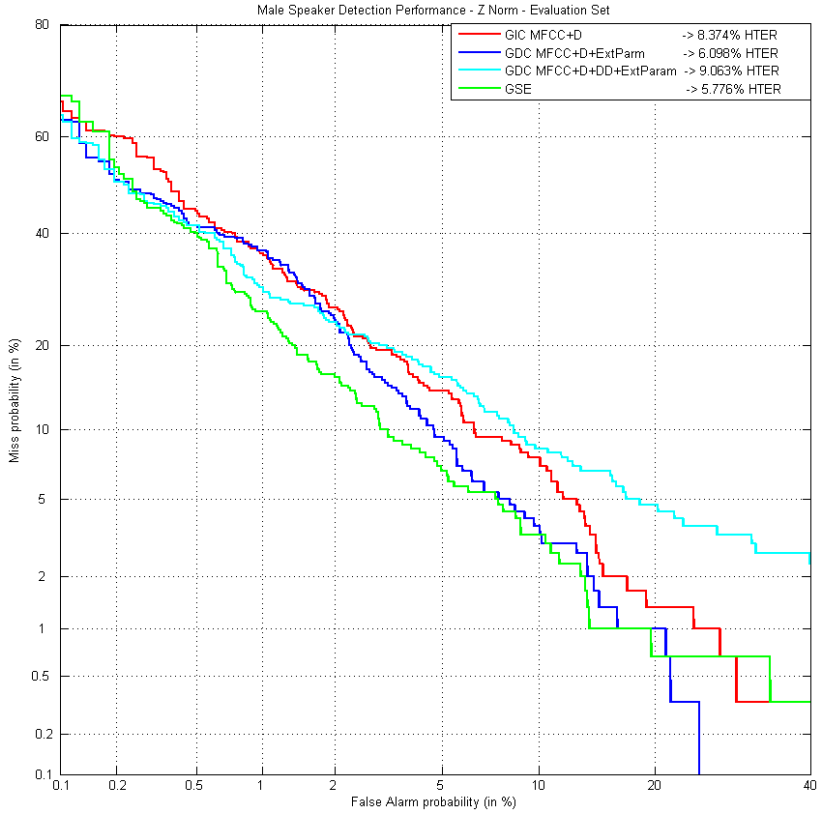



**Figure 5-53** Male's DET curves on HESPERIA evaluation set, applying ZNorm

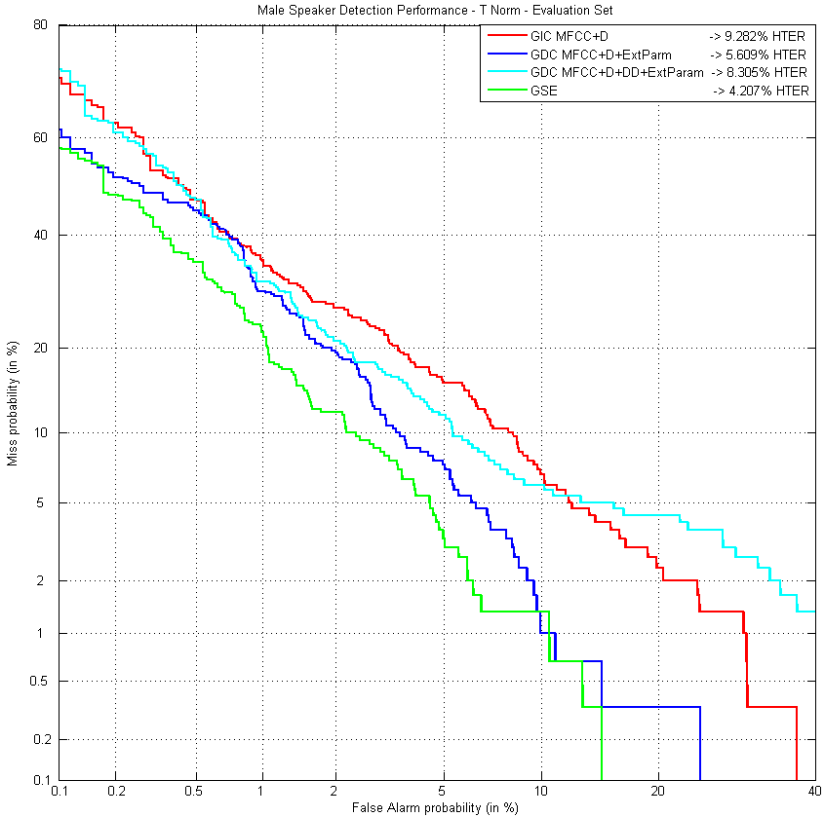



**Figure 5-54** Male's DET curves on HESPERIA evaluation set, applying TNorm

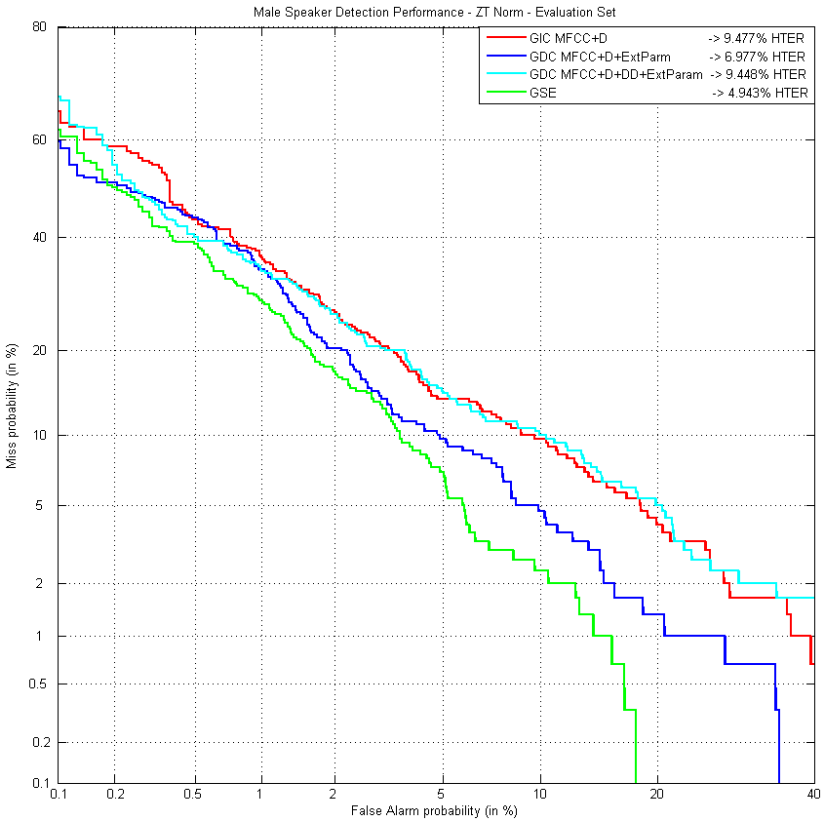



**Figure 5-55** Male's DET curves on HESPERIA evaluation set, applying ZTNorm

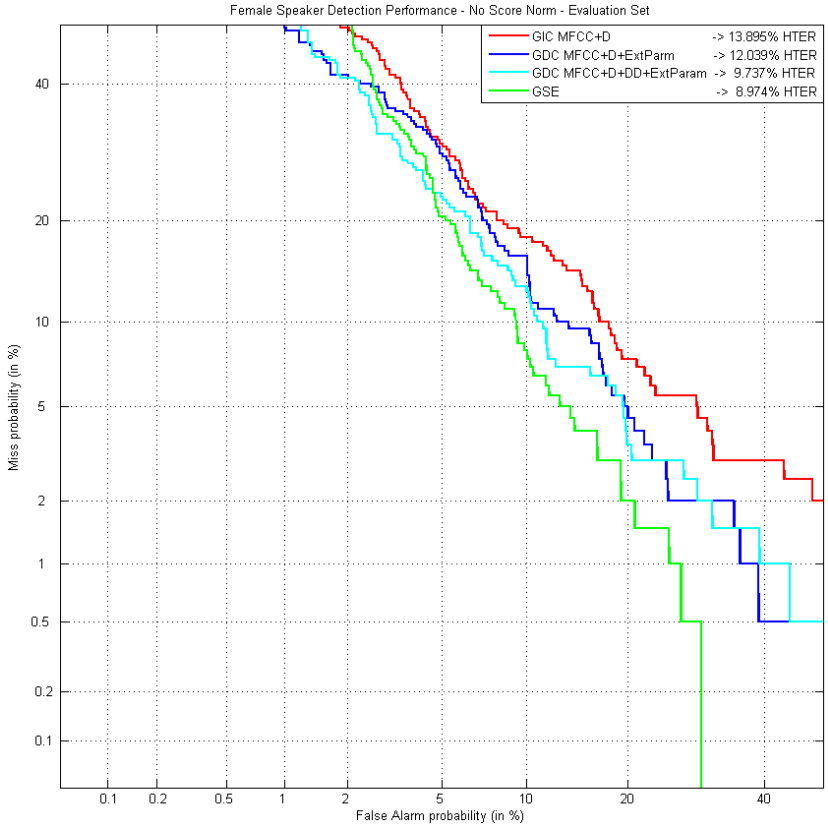



**Figure 5-56** Female's DET curves on HESPERIA evaluation set, without applying any score normalisation technique

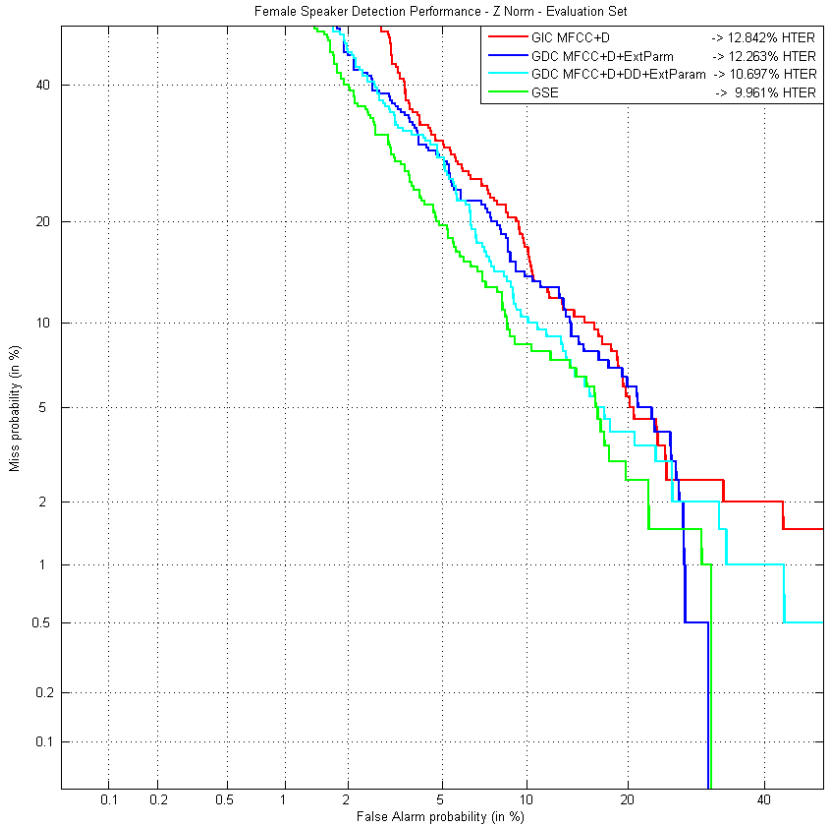



**Figure 5-57** Female's DET curves on HESPERIA evaluation set, applying ZNorm

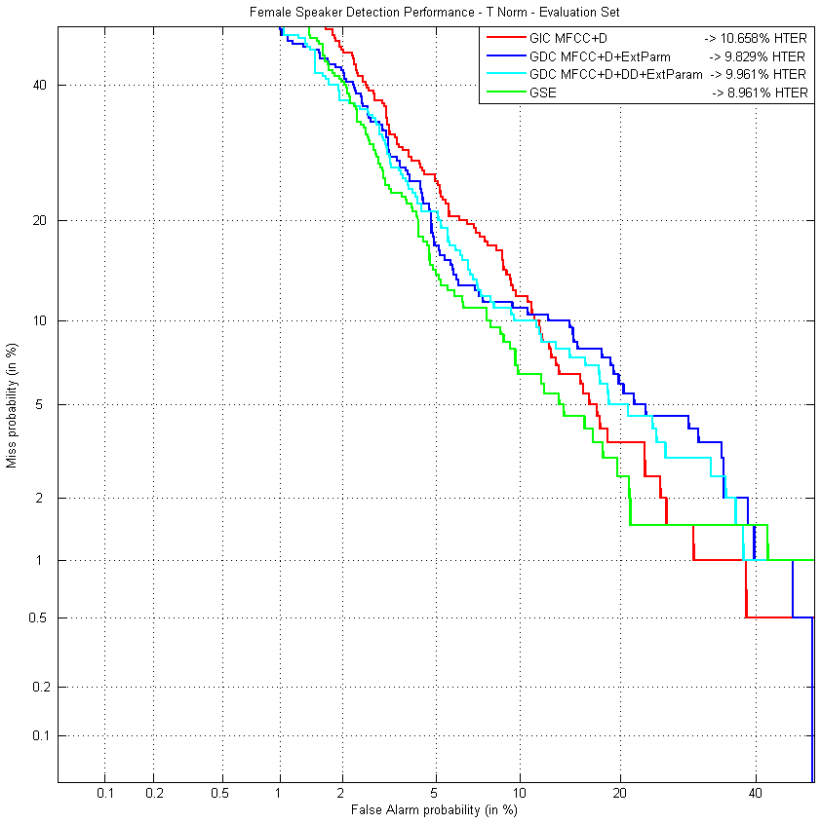



**Figure 5-58** Female's DET curves on HESPERIA evaluation set, applying TNorm

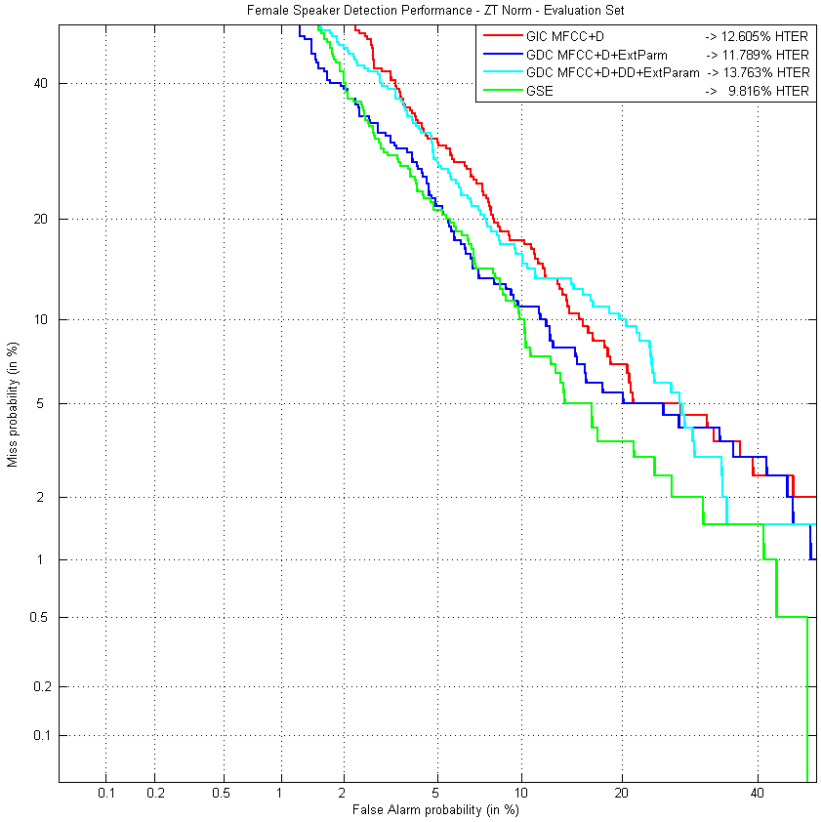



**Figure 5-59** Female's DET curves on HESPERIA evaluation set, applying ZNorm

Like in *Scenario 1* depending on the score normalisation applied there are some cases in which the DET curve associated with the GSE configuration is not always granting a

better performance at all points of the curve. Regarding this fact is worth to note the following: When testing the system in the evaluation set, we have already defined an operation point based on results from the development set. Therefore we are not evaluating the performance of our systems at all points but at the specific one given by $\theta_{dev}$. Furthermore, as previously stated, our study is not focused on minimizing *AUC* but *EER* and thus *HTER*. Anyway, it would be desirable that the DET curve provided by the gender-dependent configuration incorporating GSE information should present better values than the rest of configurations at all points of the curve.

#### 5.2.1.2.1 Brief Conclussions

To conclude this section, we must point out that in the closed-set text-constrained scenario where channel variability has been introduced the recognition rates obtained are clearly worse than in the previous case in which only microphone channel recordings are used. Otherwise, the main conclusion that can be extracted from the set of tests carried out is the same as in the previously presented scenario: The use of a gender-dependent extended-biometric parameterisation in which GSE information have been incorporated provides a clear improvement in terms of recognition rates respect to the use of a classic gender-independent approach.

In a more specific way we can highlight the following aspects. In order to improve recognition rates it is essential to use a gender-dependent parameterization. In the particular case of not applying any score normalization technique, the trend of the previous scenario is maintained regarding the number of channels in the filter bank used to compute the MFCCs. Namely the number of channels in the used filter bank is higher for female speakers than for male speakers; although the number of channels differs from the previous scenario, as the type of involved recordings also differs. Moreover, in this scenario the optimal number of MFCCs for male speakers is higher than for female speakers, contrary to what happened in the previous scenario. This confirms not only the need for gender-dependent parameterization, but the need to adapt the set of parameters for each particular situation.

Additionally, it has been confirmed that the use of extra parameters under some specific combination helps to increase recognition rates. However, in this particular scenario the proposed F3 coefficient does not systematically appear in the optimal configuration for all types of score normalization techniques for male speakers, just for the case of applying TNorm. In the case of female speakers, the use of F3 helps in the increase of recognition rates for all the score normalization techniques, alone or combined with other extra parameters.

Regarding the use of ΔΔ coefficients in this scenario, their use is completely discouraged, as the recognition rates obtained with configurations including these coefficients are worse than in the case of just using MFCCs+Δ, especially for male speakers.

Like in the previous scenario, the results obtained in terms of $EER_X$ and $HTER_X$, allow us to draw some important conclusions. The incorporation of GSE conveniently parameterised into the MFCC+Δ feature vectors, provides an improvement in recognition rates, which remains consistent over the development set and the evaluation set. Again, VTE coefficients do not provide additional benefit for speaker recognition purposes.

Finally, the use of score normalization techniques is clearly influenced by the amount of information available for normalization purposes. In this particular scenario the score

normalization technique providing best results is also gender-dependent. Particularly, the use of ZNorm in the case of male speakers provides a significant improvement over other normalization techniques; whereas in the case of female speakers, the most successful results are obtained when no score normalization is applied. In this regard, it is worth noting that the selected optimal configuration (in terms of number of MFCCs and channels in the filter bank) is not going to remain the same for all the score normalization techniques.

### 5.2.2 Text-Independent Speaker Recognition

Like in the case of the text-constrained speaker recognition scenarios using HESPERIA database, the aim of this section maintains the same goals as the ones defined for those scenarios. In other words, we are going to check whether a gender-dependent characterisation of speakers provides some improvement, in terms of recognition rates, respect to the use of a gender-independent characterisation. For that purpose, different configurations, especially regarding the number of MFCCs and the number of filters used to compute them, have been tested. Additionally, we need to analyse the usefulness of certain extra parameters frequently used in speaker recognition systems such as: frame energy, ΔE or pitch (F0), and a new one, namely, the third formant (F3). Moreover, we meet again the challenge of verifying whether the use of ΔΔ coefficients is justified in the speaker recognition area, but in a text-independent scenario. Last but not least, we need to verify whether the extended-biometric parameters proposed are useful for speaker characterisation purposes, and thus are able to improve speaker recognition systems regarding recognition rates.

For each of the proposed scenarios, we will present the results obtained using both the *Baseline* and the GDEB front-ends, when applied to the GMM-UBM approach. The database ALBAYZIN presents the same restrictions regarding the number of speakers and the amount of available data from each speaker, as the database HESPERIA. For this reason, neither the SV-GMM nor the *i*-vector approaches will be used on these tests. Again, TNorm, ZNorm and ZTNorm score normalisation techniques will be applied to analyse the effect of such techniques in recognition rates.

The system performance will be analysed through the use of the same quality measures defined for the previous scenarios, i.e. *EER* and *HTER* (see Eq. (5-4) to Eq. (5-6)). Based on these metrics we have run a battery of tests using the *Baseline* front-end in order to minimise the *EER*. However, as no cross-gender trials are going to be presented, we may use again the *HEER*:

$$HEER = \frac{EER_M(MFCC, F, \Delta, \Delta\Delta, G, \alpha) + EER_F(MFCC, F, \Delta, \Delta\Delta, G, \alpha)}{2} \qquad \textbf{Eq. (5-8)}$$

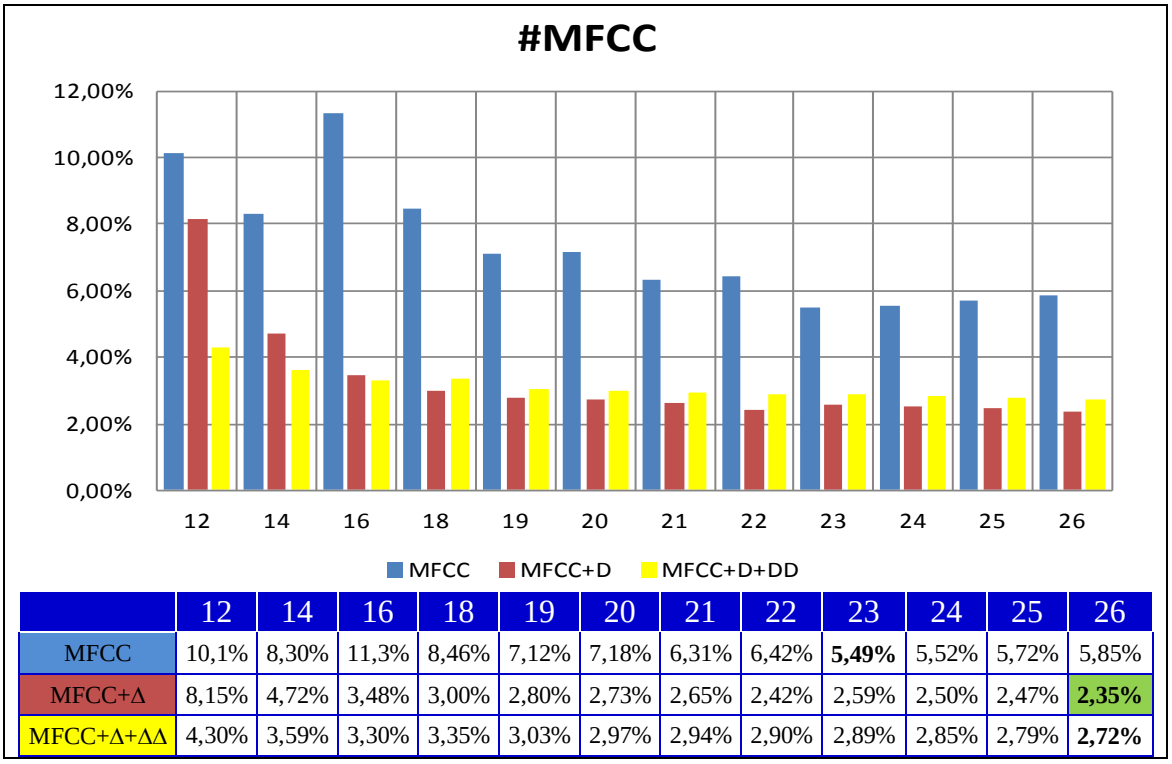
Where *MFCC*={12,14,16,18,19,20,21,22,23,24,25,26} denotes the number of MFCC coefficients computed, *F*={24,28,30,34,38,40,44,48,50} refers to the number of filters on the filter bank used to compute the MFCC, Δ={true/false} and ΔΔ={true/false} refers to the whether the Δ (delta) and ΔΔ (double-delta) coefficients are present on the feature vector, *G*={256,512} refers to number of Gaussians used to build both the UBM and the speaker's models; finally $\alpha$={5,8,10,16,20} represents the relevance factor used to adapt the speaker's model from the UBM using the MAP algorithm. The selected values for each parameter (in brackets) which have been tested are typical values used for the kind of problem that we are facing. It is worth to remember that in Eq. (5-8), we have distinguished between *EER* for male ($EER_M$) and female ($EER_F$) speakers.

Obviously, when using a gender-independent parameterisation, it is necessary to reach a compromise between both *EER* in order to minimise *HEER*. However, when a gender-dependent parameterisation is used, this compromise disappears and the objective is to minimise $EER_M$ and $EER_F$ independently.

DET curves will be used again to evaluate the calibration of the system as they represent the *FRR* against the *FAR*.

In the first place, we are going to present the results obtained on the development set using the *Baseline* front-end, thus in a gender-independent configuration. Working on the development set allows us to tune the recognition system background parameters and meta-parameters. At the same time, we can analyse the usefulness of classical parameters (i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ) in a text-independent scenario, the effect of the number of MFCCs, and the number of filters used on the filter bank to compute them. We must take into account two additional issues. First of all, as we are using a gender-independent setup, it is necessary to reach a compromise between $EER_M$ and $EER_F$, as the objective is to minimise *HEER*. Secondly, in this first approach, no score normalisation techniques have been applied, later on we will face the influence of the different score normalisation processes on the performance of the systems.

The first set of figures represents the results obtained in terms of *HEER*, based on each of the configuration parameters in Eq. (5-7), assuming that we are using the *Baseline* front-end, thus gender-independent configuration (labelled as GIC), and no score normalisation are applied. Like in the previous scenarios, the process followed consists in fixing a value of a specific parameter and test for the rest of configurable parameters whose configuration provides better results in terms of the *HEER*. For instance, regarding the number of MFCCs, we start by fixing its value to 12, and we test all the combinations for the remaining parameters, selecting the one providing the lowest *HEER*. Then we move to the next MFCC value, i.e. 14, and repeat the tests.

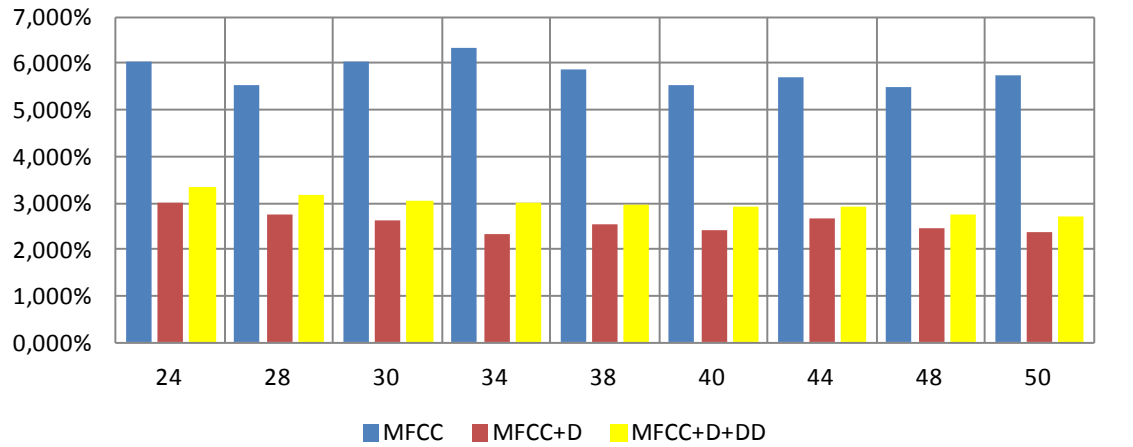



| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 10,1% | 8,30% | 11,3% | 8,46% | 7,12% | 7,18% | 6,31% | 6,42% | **5,49%** | 5,52% | 5,72% | 5,85% |
| MFCC+Δ | 8,15% | 4,72% | 3,48% | 3,00% | 2,80% | 2,73% | 2,65% | 2,42% | 2,59% | 2,50% | 2,47% | **2,35%** |
| MFCC+Δ+ΔΔ | 4,30% | 3,59% | 3,30% | 3,35% | 3,03% | 2,97% | 2,94% | 2,90% | 2,89% | 2,85% | 2,79% | **2,72%** |

**Figure 5-60** *HEER* obtained depending on the number of MFCCs and the use of Δ and ΔΔ (GIC – development set)

In Figure 5-60, we highlight the influence of the number of MFCCs, as well as the influence of the use of Δ and ΔΔ coefficients on the recognition rates. The best results obtained for each configuration, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ, are marked in bold letters; while the best result obtained is marked in green. Like in the text-constrained scenario, MFCCs alone (i.e. without Δ or ΔΔ) are not accurate enough to model speakers, as recognition rates are clearly worse than the ones obtained when complemented with Δ and ΔΔ coefficients. In this case the use of MFCCs+Δ systematically produce better results than MFCCs+Δ+ΔΔ for MFCC values between 18 and 26.

Regarding the number of filters, we operate following the same procedure, i.e. we fix a specific value for *F*, and we test all the combinations of the remaining parameters, selecting the configuration that provides better recognition rates in terms of *HEER* for the specific *F* value. Figure 5-61 shows the results obtained in terms of *HEER* for the different *F* values tested, and for the three tested configurations, i.e. MFCCs, MFCCs+Δ and MFCCs+Δ+ΔΔ. In this case, the number of filter banks used which produces the best results is 48 or 50 for all the tested configurations, although a slightly improvement is obtained in the case of using MFCCs+Δ when 34 filters are used.



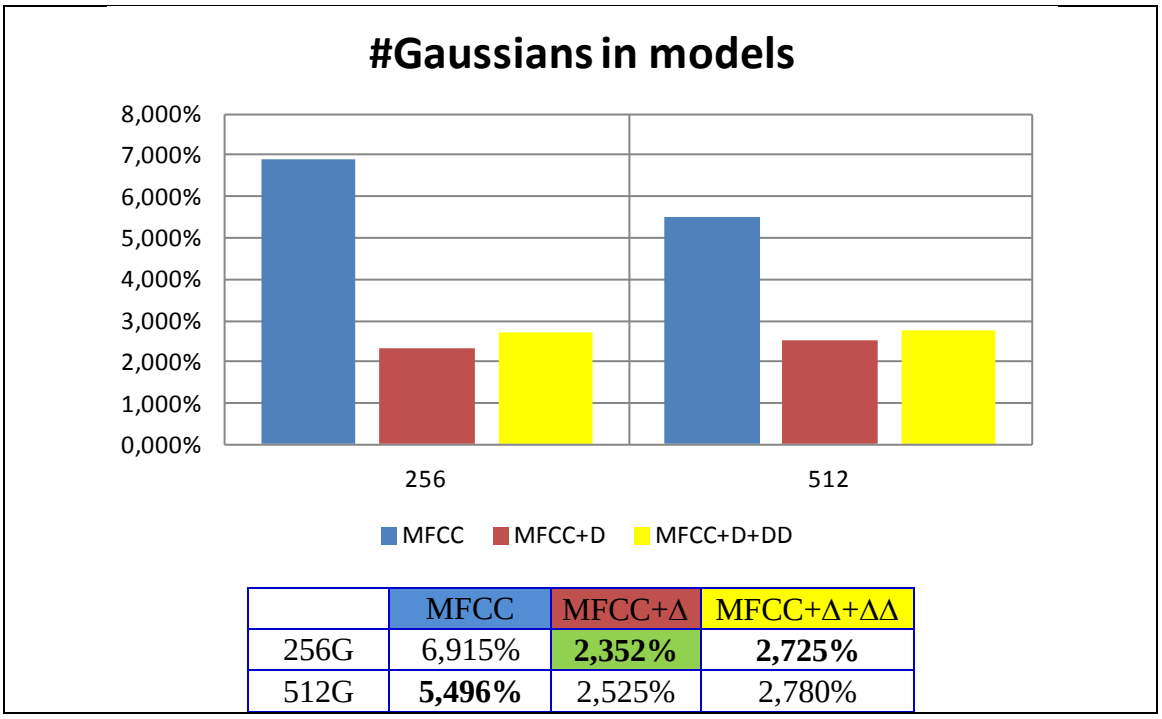

| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 6,02% | 5,52% | 6,03% | 6,31% | 5,85% | 5,54% | 5,69% | **5,49%** | 5,72% |
| MFCC+Δ | 2,99% | 2,75% | 2,62% | **2,35%** | 2,55% | 2,42% | 2,65% | 2,47% | 2,39% |
| MFCC+Δ+ΔΔ | 3,35% | 3,17% | 3,04% | 3,00% | 2,94% | 2,92% | 2,92% | 2,76% | **2,72%** |

**Figure 5-61** *HEER* obtained depending on the number of filters and the use of Δ and ΔΔ (GIC – development set)

If we consider the number of Gaussians used to model the speakers in the gender-independent configuration, we can conclude (see Figure 5-62), that the best results are obtained when 256 Gaussians are used in the case of including Δ or ΔΔ coefficients in the feature vector, while in the case of using just MFCCs alone, the best results are produced when 512 Gaussians are used.

Table 5-19, shows the specific configurations providing the best recognition rates in terms of *HEER*, for the *Baseline* front-end (therefore in a gender-independent configuration), for each set of feature vector parameters (i.e. MFCCs, MFCCs+Δ and

MFCCs+Δ+ΔΔ). Moreover, it also shows the $EER_F$ and $EER_M$ obtained in each case as well as the specific score threshold.



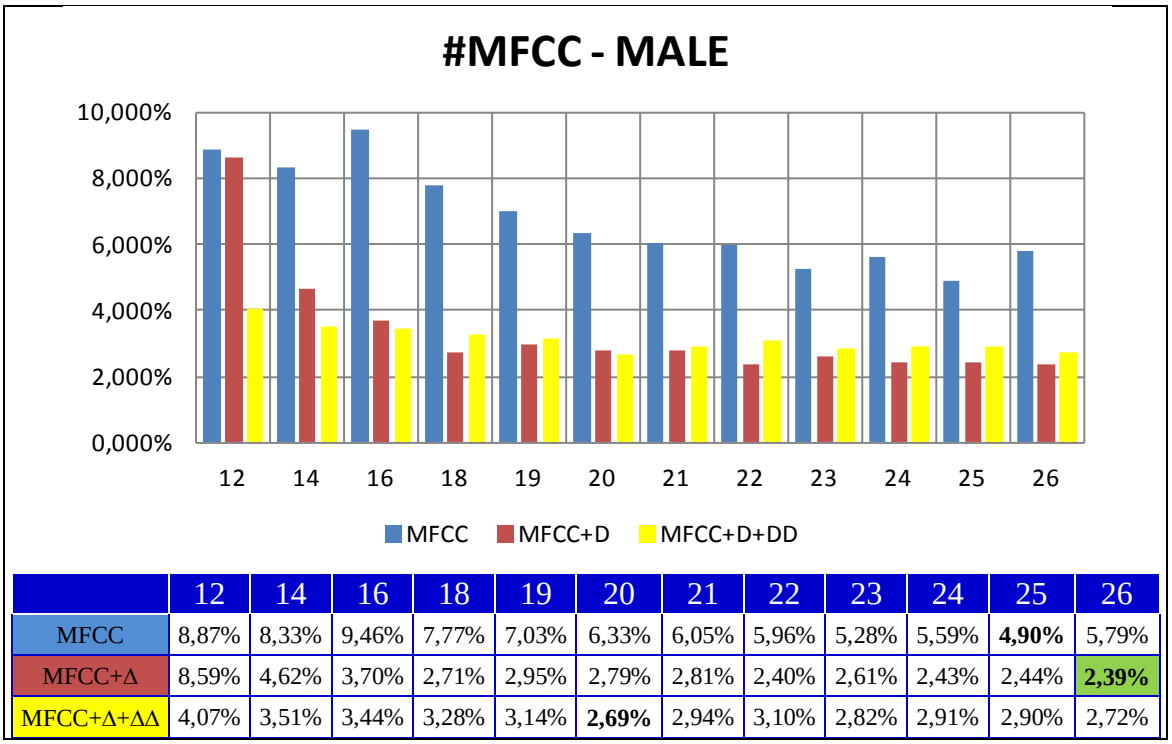

| | MFCC | MFCC+Δ | MFCC+Δ+ΔΔ |
|---|---|---|---|
| 256G | 6,915% | **2,352%** | **2,725%** |
| 512G | **5,496%** | 2,525% | 2,780% |

**Figure 5-62** *HEER* obtained depending on the number of Gaussians and the use of Δ and ΔΔ (GIC – development set)

| Parameters | *F* | *G* | *α* | $EER_M$ [$\theta_M$] | $EER_F$ [$\theta_F$] | *HEER* |
|---|---|---|---|---|---|---|
| 23MFCCs **(GIC MFCC)** | 48 | 512 | 5 | 5.284% [0.146] | 5.708% [0.179] | 5.496% |
| 26MFCCs+Δ **(GIC MFCC+Δ)** | 34 | 256 | 5 | 2.534% [-0.178] | 2.170% [-0.169] | 2.352% |
| 26MFCCs+Δ+ΔΔ **(GIC MFCC+Δ+ΔΔ)** | 50 | 256 | 5 | 3.042% [-0.401] | 2.409% [-0.375] | 2.725% |

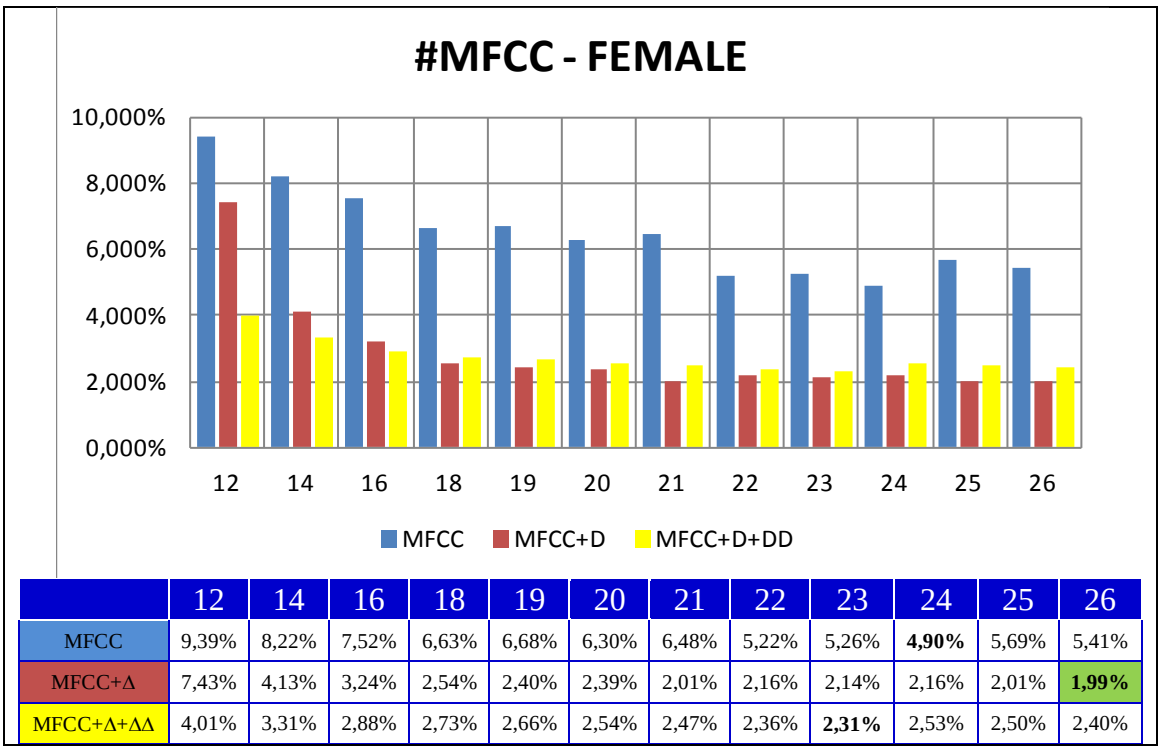
**Table 5-19** Baseline front-end producing the best *HEER* in a gender-independent configuration

So far, we have run the tests using the *Baseline* front-end, thus in a gender-independent configuration setup. Under this constraint, a compromise between $EER_F$ and $EER_M$ must be reached when minimizing *HEER*. If we remove this constraint, allowing the use of a gender-dependent configuration, $EER_F$ and $EER_M$ can be independently minimised, that may result in a reduction of *HEER*. Therefore, like in previous scenarios, we have performed the same analysis already presented, but in a gender-dependent setup (labelled as GDC). Figure 5-63 to Figure 5-67 provide the results obtained in terms of $EER_X$, where $X = \{F,M\}$, when the effect of the number of MFCCs, the number of filters and the number of Gaussians in the model are analysed.



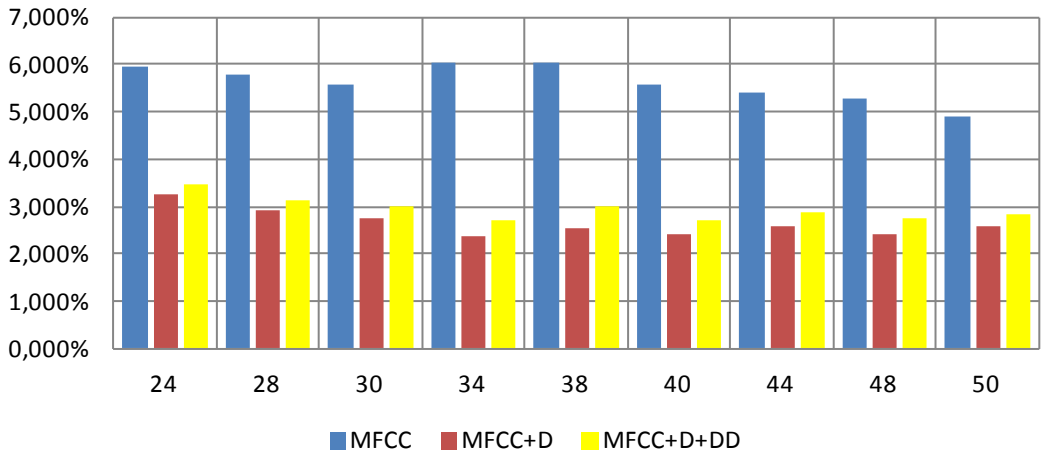

| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 8,87% | 8,33% | 9,46% | 7,77% | 7,03% | 6,33% | 6,05% | 5,96% | 5,28% | 5,59% | **4,90%** | 5,79% |
| MFCC+Δ | 8,59% | 4,62% | 3,70% | 2,71% | 2,95% | 2,79% | 2,81% | 2,40% | 2,61% | 2,43% | 2,44% | **2,39%** |
| MFCC+Δ+ΔΔ | 4,07% | 3,51% | 3,44% | 3,28% | 3,14% | **2,69%** | 2,94% | 3,10% | 2,82% | 2,91% | 2,90% | 2,72% |

**Figure 5-63** $EER_M$ obtained depending on the number of MFCC and the use of Δ and ΔΔ (GDC – development set)



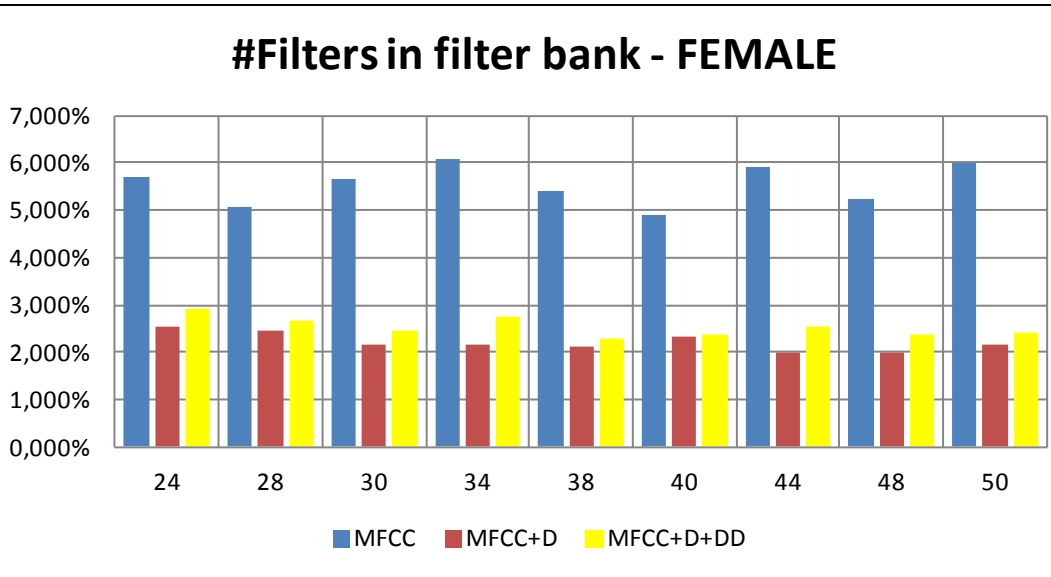

| | 12 | 14 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 9,39% | 8,22% | 7,52% | 6,63% | 6,68% | 6,30% | 6,48% | 5,22% | 5,26% | **4,90%** | 5,69% | 5,41% |
| MFCC+Δ | 7,43% | 4,13% | 3,24% | 2,54% | 2,40% | 2,39% | 2,01% | 2,16% | 2,14% | 2,16% | 2,01% | **1,99%** |
| MFCC+Δ+ΔΔ | 4,01% | 3,31% | 2,88% | 2,73% | 2,66% | 2,54% | 2,47% | 2,36% | **2,31%** | 2,53% | 2,50% | 2,40% |

**Figure 5-64** $EER_F$ obtained depending on the number of MFCC and the use of Δ and ΔΔ (GDC – development set)



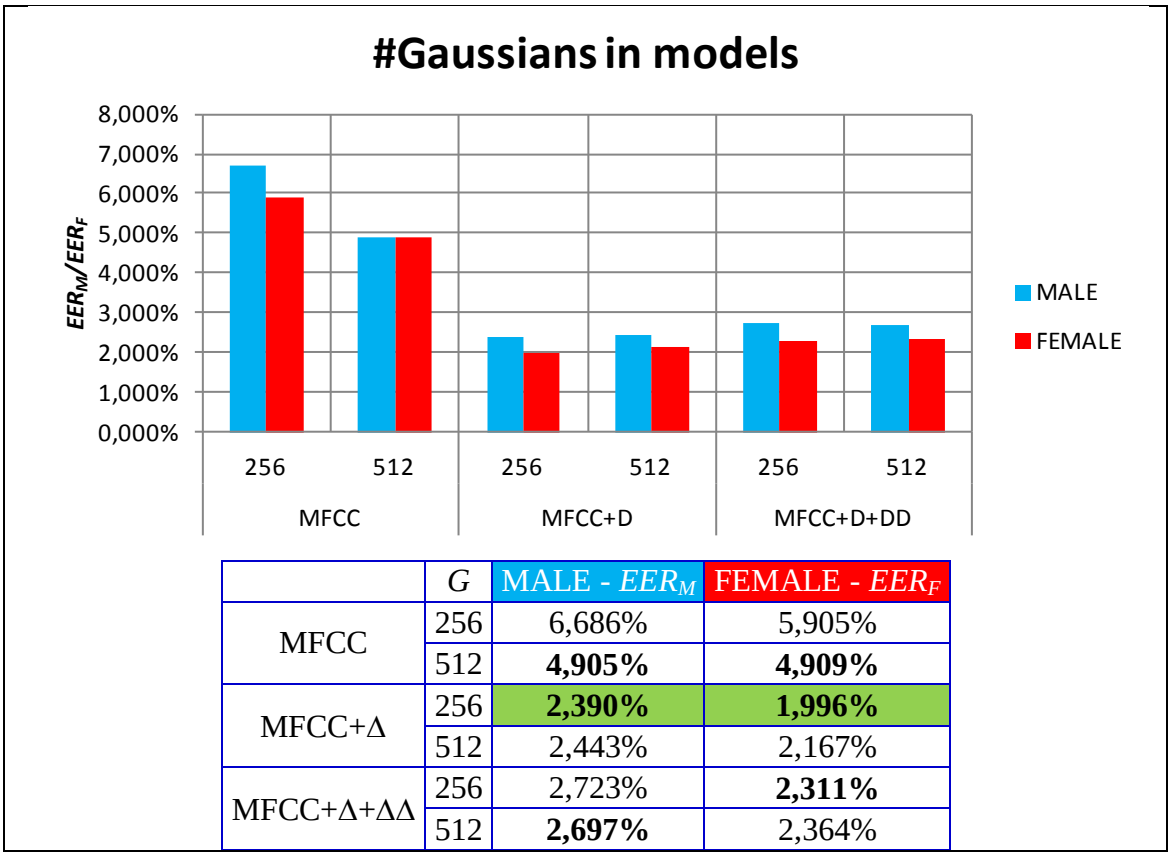

| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 5,966% | 5,792% | 5,591% | 6,053% | 6,027% | 5,576% | 5,394% | 5,284% | **4,905%** |
| MFCC+Δ | 3,254% | 2,924% | 2,746% | **2,390%** | 2,549% | 2,402% | 2,587% | 2,436% | 2,580% |
| MFCC+Δ+ΔΔ | 3,466% | 3,125% | 2,992% | 2,723% | 2,992% | **2,697%** | 2,902% | 2,750% | 2,826% |

**Figure 5-65** $EER_M$ obtained depending on the number of filters and the use of Δ and ΔΔ (GDC – development set)




| | 24 | 28 | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| MFCC | 5,697% | 5,053% | 5,663% | 6,068% | 5,413% | **4,909%** | 5,905% | 5,223% | 5,977% |
| MFCC+Δ | 2,553% | 2,466% | 2,174% | 2,170% | 2,140% | 2,333% | **1,996%** | 2,011% | 2,182% |
| MFCC+Δ+ΔΔ | 2,920% | 2,652% | 2,470% | 2,742% | **2,311%** | 2,356% | 2,545% | 2,364% | 2,409% |

**Figure 5-66** $EER_F$ obtained depending on the number of filters and the use of Δ and ΔΔ (GDC – development set)




| | $G$ | MALE - $EER_M$ | FEMALE - $EER_F$ |
|---|---|---|---|
| MFCC | 256 | 6,686% | 5,905% |
| | 512 | **4,905%** | **4,909%** |
| MFCC+Δ | 256 | **2,390%** | **1,996%** |
| | 512 | 2,443% | 2,167% |
| MFCC+Δ+ΔΔ | 256 | 2,723% | **2,311%** |
| | 512 | **2,697%** | 2,364% |

**Figure 5-67** $EER_M$ (blue) and $EER_F$ (red) obtained depending on the number of Gaussians and the use of Δ and ΔΔ (GDC – development set)

From Figure 5-63 and Figure 5-64, we can conclude that in this text-independent scenario, the number of MFCCs needed to characterise precisely speakers is 26 regardless of gender. Moreover, the use of MFCCs+Δ+ΔΔ configuration does not offer additional benefits neither for male nor for female speakers. Differences arise however in the number of filters included in the filter bank to compute the MFCCs. Particularly, in the case of male speakers (see Figure 5-65) the best results are obtained when 34 filters are used, while in the case of female speakers (see Figure 5-66) this number rises up to 44. This result confirms the trend of previous scenarios, where the number of filters needed to accurately model speakers, is higher for female than for male speakers.

Concerning the number of Gaussians used to build both the UBM and the speaker's models (see Figure 5-67), no general conclusion can be extracted. Nevertheless, the best results are obtained when $G$=256 and the MFCCs+Δ configuration is used.

Therefore, we can conclude that in the set of tests carried out in this scenario, using the development set, a gender-dependent parameterisation presents a clear advantage over the gender-independent parameterisation, in terms of *HEER*, even in the case of using just classical parameters. Table 5-20 provides a comparison between the recognition rates obtained by the system, when a GDC or a GIC is used (the best results obtained so far are highlighted in light green). Additional columns have been added ($EER_X$ RR), which provide the relative reduction obtained by GDC, in terms of $EER_X$, respect to the corresponding GIC. The relative reduction in terms of *HERR* respect to GIC has been indicated in brackets in the *HEER* column.

| Parameters | Gen. | *F* | *G* | *α* | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER* |
|---|---|---|---|---|---|---|---|---|---|
| 23MFCCs **(GIC MFCC)** | M/F | 48 | 512 | 5 | 5.284% [0.146] | - | 5.708% [0.179] | - | 5.496% [-] |
| 25MFCCs **(GDC MFCC)** | M | 50 | 512 | 5 | 4.905% [0.199] | 7.16% | 4.909% [0.177] | 14.00% | 4.907% [10.71%] |
| 24MFCCs **(GDC MFCC)** | F | 40 | 512 | 16 | | | | | |
| | | | | | | | | | |
| 26MFCCs+Δ **(GIC MFCC+Δ)** | M/F | 34 | 256 | 5 | 2.534% [-0.178] | - | 2.170% [-0.169] | - | 2.352% [-] |
| 26MFCCs+Δ **(GDC MFCC+Δ)** | M | 34 | 256 | 16 | 2.390% [-0.001] | 5.68% | 1.996% [-0.166] | 8.02% | 2.193% [6.76%] |
| 26MFCCs+Δ **(GDC MFCC+Δ)** | F | 44 | 256 | 5 | | | | | |
| | | | | | | | | | |
| 26MFCCs+Δ+ΔΔ **(GIC MFCC+Δ+ΔΔ)** | M/F | 50 | 256 | 5 | 3.042% [-0.401] | - | 2.409% [-0.375] | - | 2.725% [-] |
| 20MFCCs+Δ+ΔΔ **(GDC MFCC+Δ+ΔΔ)** | M | 40 | 512 | 20 | 2.697% [0.010] | 11.33% | 2.311% [-0.164] | 4.08% | 2.504% [8.13%] |
| 23MFCCs+Δ+ΔΔ **(GDC MFCC+Δ+ΔΔ)** | F | 38 | 256 | 8 | | | | | |

**Table 5-20** GDC vs. GIC for the ALBAYZIN development set (RR – Relative Reduction)

According to the results shown in Table 5-20 when the feature vector contains just MFCCs coefficients, a gender-dependent configuration produces a relative reduction of 10.71% in terms of *HEER* thanks to a simultaneous reduction on the $EER_M$ (from 5.284% to 4.905%) and $EER_F$ (from 5.708% to 4.909%). In this case, the number of filters used to extract the MFCCs as well as the number of MFCCs is different for male and female speakers. In the case of incorporating Δ coefficients to the feature vectors, we also obtain a relative reduction in terms of *HEER* of 6.76% when a gender-dependent configuration is used. The number of MFCCs is the same for male and female speakers, however the number of filters required to extract them is larger in the case of female speakers. When using the GDC the $EER_M$ drops from 2.534% to 2.390%, while $EER_F$ is reduced from 2.170% to 1.996%. Finally, GDC also provides better results when using MFCCs+Δ+ΔΔ, in terms of *HEER*, than GIC.

Figure 5-68 and Figure 5-69 show the DET curves of male and female speakers, for both configurations (GIC and GDC), from the different sets of classical parameters used so far, and specified in Table 5-20. It must be reminded that the goal of the test is the reduction of *EER* and not the *Area Under the Curve* (AUC – see Section 1.4), thus it may happen that GIC shows better results than GDC for some of the points of the curve. However, GDC will always produce better or at least equal results than GIC in terms of *EER*.

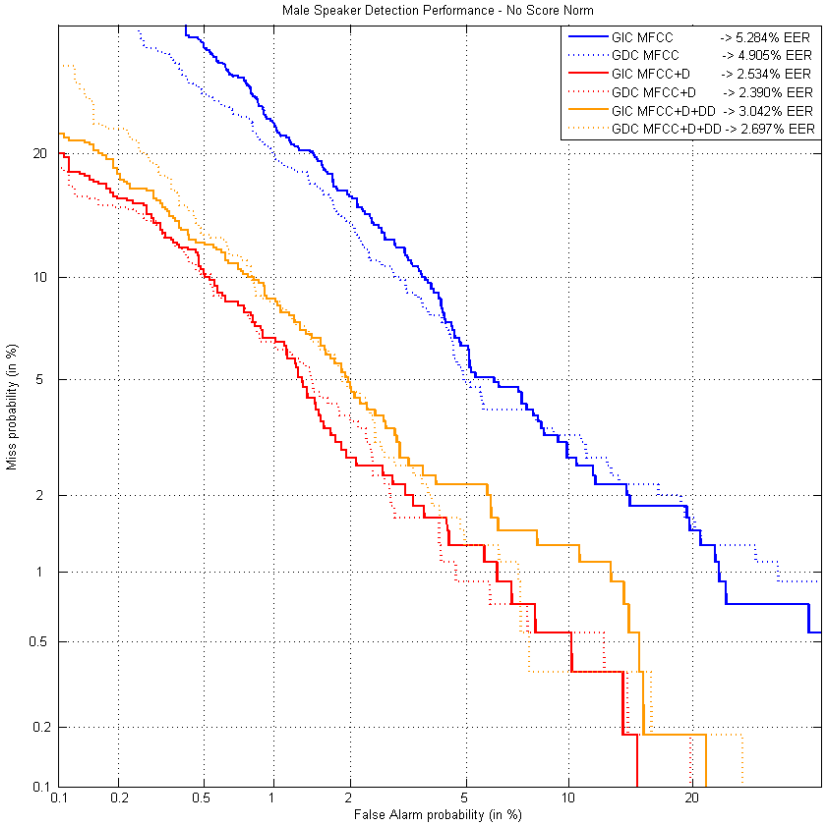



**Figure 5-68** DET curve for classic parameters on ALBAYZIN male development set for GIC and GDC

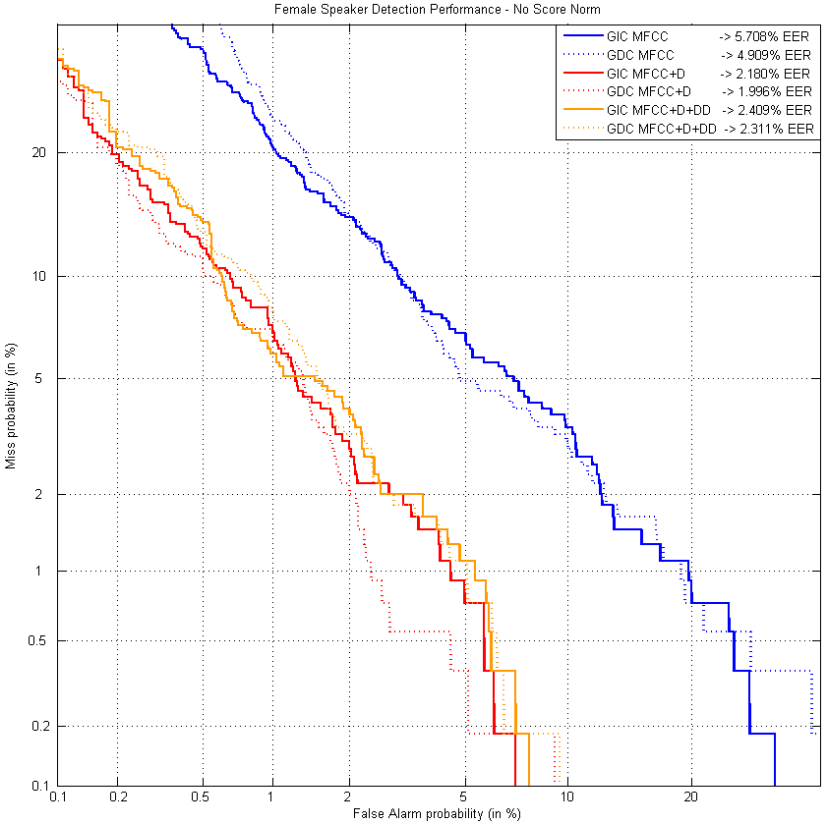



**Figure 5-69** DET curve for classic parameters on ALBAYZIN female development set for GIC and GDC

In this text-independent scenario, we have also analysed the effect of adding the extra parameters, i.e. Energy, Δ Energy, Pitch (F0) and third formant (F3), both on the gender-dependent and gender-independent configurations. All combinations of these extra parameters have been included in the feature vectors, however, only the most relevant results, in terms of $EER_M$, $EER_F$ and *HEER*, obtained on this set of tests are reflected in Table 5-21.

| Parameters | Gen. | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|
| **GIC MFCC** | M/F | - | 5.284% [0.146] | - | 5.708% [0.179] | - | 5.496% [-] |
| **GIC MFCC** | M/F | ΔE+F0+F3 | 4.564% [0.176] | 13.62% | 5.367% [0.149] | 5.97% | 4.966% [9.64%] |
| **GDC MFCC** | M | - | 4.905% [0.199] | 7.16% | 4.909% [0.177] | 14.00% | 4.907% [10.71%] |
| | F | - | | | | | |
| **GDC MFCC** | M | E+ΔE+F0+F3 | 4.405% [0.204] | 16.63% | 4.837% [0.166] | 15.26% | 4.621% [15.92%] |
| | F | F0 | | | | | |
| | | | | | | | |
| **GIC MFCC+Δ** | M/F | - | 2.534% [-0.178] | - | 2.170% [-0.169] | - | 2.352% [-] |
| **GIC MFCC+Δ** | M/F | E+ΔE | 2.163% [-0.035] | 14.64% | 1.992% [-0.028] | 8.20% | 2.078% [11.67%] |
| **GDC MFCC+Δ** | M | - | 2.390% [-0.001] | 5.68% | 1.996% [-0.166] | 8.02% | 2.193% [6.76%] |
| | F | - | | | | | |
| **GDC MFCC+Δ** | M | E+ΔE | 2.163% [-0.035] | 14.64% | 1.818% [-0.113] | 16.2% | 1.991% [15.37%] |
| | F | E+ΔE+F0+F3 | | | | | |
| | | | | | | | |
| **GIC MFCC+Δ+ΔΔ** | M/F | - | 3.042% [-0.401] | - | 2.409% [-0.375] | - | 2.725% [-] |
| **GIC MFCC+Δ+ΔΔ** | M/F | ΔE+F0 | 2.352% [-0.238] | 22.66% | 2.542% [-0.222] | -5.50% | 2.447% [10.21%] |
| **GDC MFCC+Δ+ΔΔ** | M | - | 2.697% [0.010] | 11.33% | 2.311% [-0.164] | 4.08% | 2.504% [8.13%] |
| | F | - | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | E+F0 | 2.205% [0.032] | 27.52% | 1.807% [0.016] | 25.00% | 2.006% [26.40%] |
| | F | E+ΔE | | | | | |

**Table 5-21** $EER_M$, $EER_F$ and *HEER* obtained on development set when extra parameters are included on the feature vectors for GIC and GDC

The best results for each set of classical parameters used (i.e. MFCCs, MFCCs+Δ, MFCCs+Δ+ΔΔ) are highlighted in green. Although not reflected in the previous table, it must be noted that not all the tested combinations of extra parameters provide a reduction in terms of $EER_M$, $EER_F$ and *HEER*. Additionally, depending on the configuration, the extra parameters providing the best results may be different.

From the results reflected in Table 5-21, we can point out several conclusions. First of all, it is clear that the incorporation of extra parameters provides an extra benefit, helping to improve recognition rates, no matter whether MFCCs, MFCCs+Δ, or MFCCs+Δ+ΔΔ parameters are used. Secondly, the best results in terms of $EER_M$, $EER_F$, and *HEER* are obtained in all the cases when using a GDC. It must be also noted that, like in previous scenarios, the use of MFCCs+Δ+ΔΔ parameters in a gender-dependent setup, becomes again the most accurate parameterisation for female speakers. Last but

not least, in this text-independent scenario, the extra parameters E+ΔE are the ones providing improvements in recognition rates for both male and female speakers. This may somehow justify their use by most speaker recognition systems. However, it must be noted that this conclusion is valid in a gender-dependent setup, where male speakers are parameterised using MFCCs+Δ, and female speakers are parameterised using MFCCs+Δ+ΔΔ. If we consider the state-of-the-art parameterisation, i.e. GIC MFCCs+Δ+ΔΔ+E+ΔE, it is clear as reflected in Table 5-22, that there are better options when it comes to reducing *HEER*. Particularly, GIC MFCCs+Δ+ΔΔ+ΔE+F0 provide better results in terms of *HEER*, while GDC MFCCs+Δ+ExtParam. provide better results in terms of $EER_M$ and $EER_F$, with significantly less number of parameters and computational cost.

| Parameters | Gen. | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER* | *HEER* RR |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ+ΔΔ** | M/F | - | 3.042% [-0.401] | - | 2.409% [-0.375] | - | 2.725% | - |
| **GIC MFCC+Δ+ΔΔ** | M/F | ΔE+F0 | 2.352% [-0.238] | 22.66% | 2.542% [-0.222] | -5.50% | 2.447% | 10.21% |
| **GIC MFCC+Δ+ΔΔ** | M/F | E+ΔE | 2.947% [-0.184] | 3.11% | 2.178% [-0.152] | 9.59% | 2.563% | 5.97% |

**Table 5-22** $EER_M$, $EER_F$, and *HEER* obtained on the development set comparing different GIC MFCC+Δ+ΔΔ configurations

Figure 5-70 and Figure 5-71 respectively provide the DET curves for male and female speakers, comparing the most relevant configurations for the present study, i.e., GIC MFCCs+Δ, GIC MFCCs+Δ+ExtParam., GDC MFCCs+Δ+ExtParam., GIC MFCCs+Δ+ΔΔ, GIC MFCCs+Δ+ΔΔ+ExtParam., GDC MFCCs+Δ+ΔΔ+ExtParam.. In other words, we compare gender-dependent with gender-independent configurations, while we check as well the influence of ΔΔ coefficients and the extra proposed parameters. Two important comments must be made concerning these plots. First of all, it is possible that GIC and GDC were the same for a specific configuration, thus in the DET plots only one curve is going to be visible (male MFCCs+Δ+ExtParam.). Additionally, we must remember that the goal is the reduction of EER and not the *Area Under the Curve* (AUC – see Section 1.4), thus it may happen that GIC can show better results than GDC for some of points of the curve. However, GDC will always achieve better or at least equal results than GIC in terms of EER.

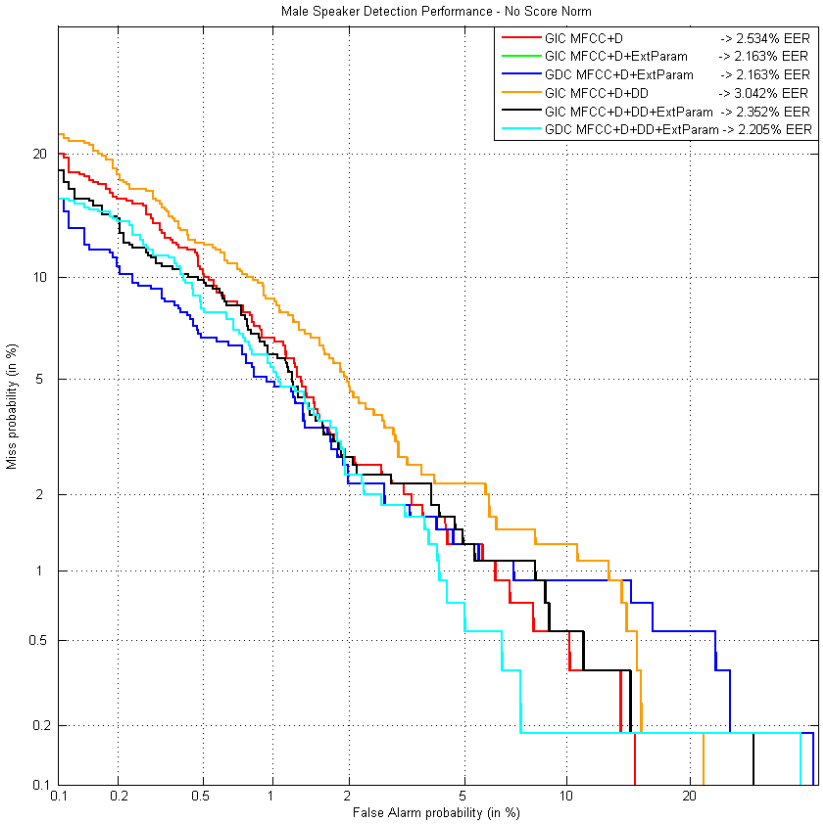



**Figure 5-70** DET curves comparing different sets of parameters under GIC and GDC without extended biometrics on the male development set

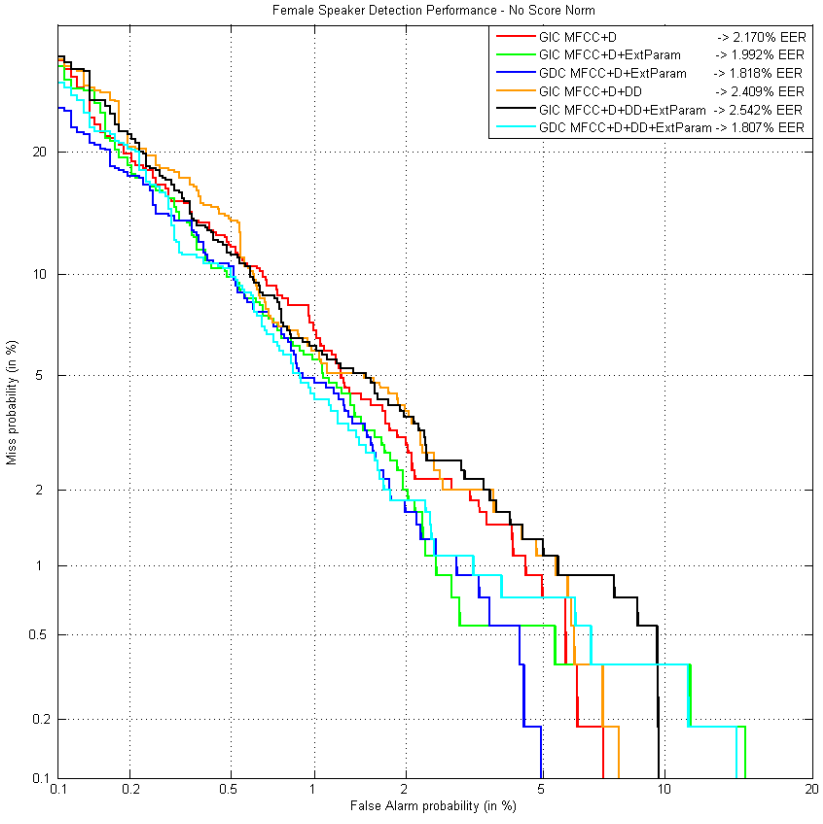



**Figure 5-71** DET curves comparing different sets of parameters under GIC and GDC without extended biometrics on the female development set

Next, we verify the viability of using the extended-biometric parameters extracted by the GDEB front-end, for speaker recognition purposes in the text-independent scenario. Like in the text-constrained scenario, the approach that has been followed consists in incorporating the extended biometric coefficients (information extracted from the vocal tract and glottal source estimates) into the best gender-dependent configuration selected so far without ΔΔ parameters, in two stages. First we incorporate a set of parameters extracted from the glottal source estimate (labelled as GSE), and once a specific configuration improving previous results is found, we continue by incorporating parameters extracted from the vocal tract estimate (labelled as VTE).

We proceed this way because the vocal tract estimate is more related to the message carried out by voice, rather than to the biometry of the speakers; therefore as we are dealing with text-independent trials, GSE is supposed to provide more benefits than VTE in terms of recognition rates. Additionally, we have already seen that even in the case of text-constrained scenarios, VTE does not provide additional improvement in recognition rates. Once more, we have ruled out the use of ΔΔ in this section, as we believe that the proposed GDEB parameterisation may represent the speaker more accurately than the one including ΔΔ coefficients, like in previous scenarios.

Although multiple configurations have been tested, regarding the multiple variables that can be tuned in the GDEB front-end, Table 5-23 shows the ultimate configurations chosen for each gender, as well as the recognition rates obtained in each case in terms of $EER_M$, $EER_F$ and *HEER*. Additionally, the relative reduction (RR) in terms of $EER_X$ and *HEER*, compared to the GIC MFCCs+Δ configuration is also provided.

DET curves corresponding to the results presented in Table 5-23 are depicted in Figure 5-72 for male speakers and Figure 5-73 for female speakers. Clearly, the parameterisation generated by the GDEB front-end, in this case just including information from the glottal source estimate in the form of mel-frequency cepstrum coefficients, is the one providing the best results on the development set for male and female speakers. Specifically, the GSE setup, provides a relative reduction in terms of *EER* respect to the best configuration using classical parameters of 30% in the case of male speakers (GDC MFCCs+Δ+ExtParam.) and 20% in the case of female speakers (GDC MFCCs+Δ+ΔΔ+ExtParam.). Regarding the specific setup selected for GSE, it must be noted that the number of filters used to compute MFCCs from the GSE is almost the same for both male and female speakers; however, the number of MFCCs used is larger in the case of male speakers.

| Parameters | Gen. | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | - | - | 2.534% [-0.178] | - | 2.170% [-0.169] | - | 2.352% [-] |
| **GDC MFCC+Δ** | M | - | E+ΔE | 2.163% [-0.035] | 14.64% | 1.818% [-0.113] | 16.23% | 1.991% [15.37%] |
| | F | - | E+ΔE+F0+F3 | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | - | E+F0 | 2.205% [0.032] | 13.00% | 1.807% [0.016] | 16.75% | 2.006% [14.73%] |
| | F | - | E+ΔE | | | | | |
| **GSE** | M | **Source-Tract Sep. Alg:** Prediction Order: 10 Forgetting Factor: 0.995 **GSE:** 13-Channel Filter bank 8 MFCC | E+ΔE | 1.504% [-0.131] | 40.65% | 1.451% [-0.145] | 33.15% | 1.477% [37.19%] |
| | F | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 12-Channel Filter bank 4 MFCC | E+ΔE+F0+F3 | | | | | |

**Table 5-23** $EER_M$, $EER_F$, and *HEER* obtained on development set (no score normalisation), comparing classical parameters with extra parameters and extended biometric parameters

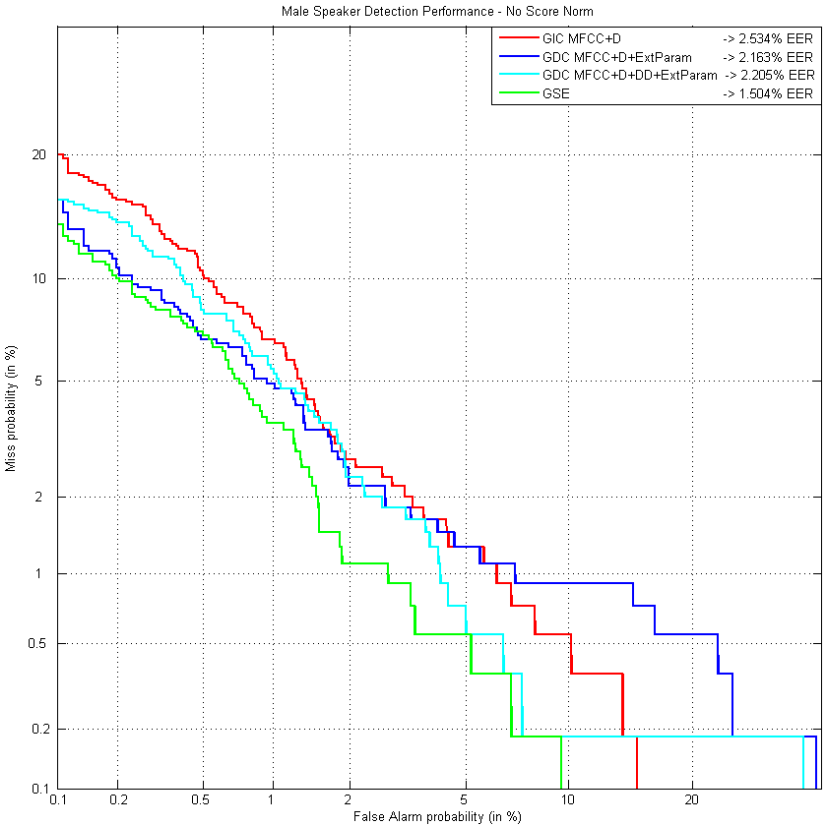



**Figure 5-72** DET curves comparing classical parameters and GDEB on ALBAYZIN development set for male speakers

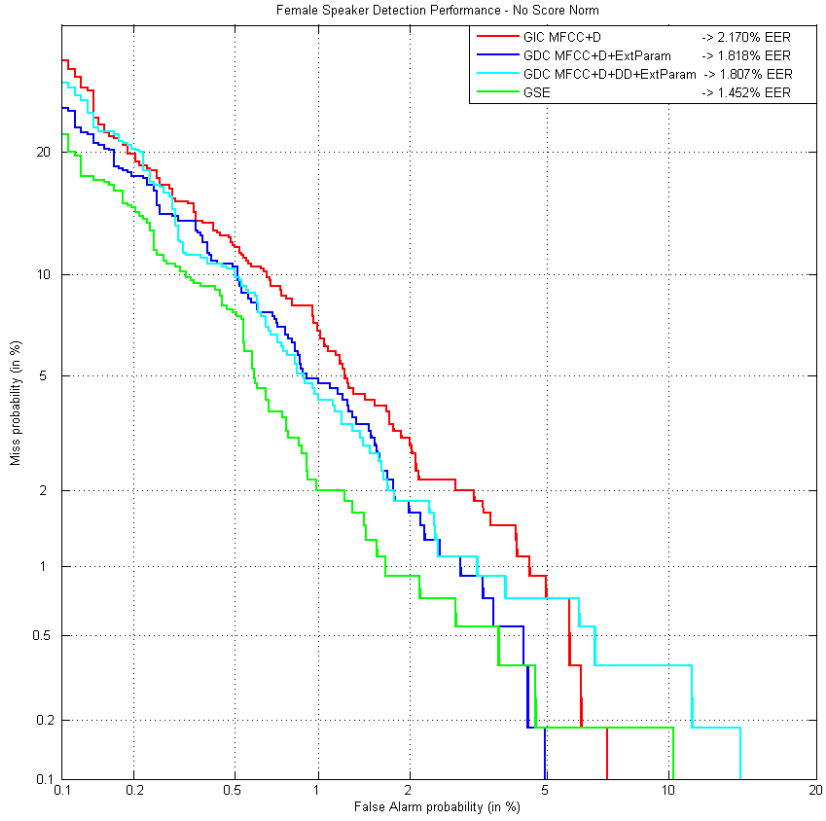



**Figure 5-73** DET curves comparing classical parameters and GDEB on ALBAYZIN development set for female speakers

Like in previous scenarios, we need to verify that the improvement derived from incorporating GSE information into the feature vector is systematically obtained and is not the result of an isolated and specific configuration. Figure 5-74 provides in green solid line the minimum $EER_X$ (y-axis) obtained when GSE is incorporated into the feature vector in the form of MFCCs for male speakers. Different numbers of $MFCC_G$={2,4,6,8,10} have been tested, which have been computed applying a filter bank with different numbers of filters $F_G$=[4...23] (x-axis). Each point in the x-axis represents the minimum *EER* obtained for a specific value of $F_G$, regardless the $MFCC_G$ values. Figure 5-75 provides the same information for female speakers. Clearly, from the depicted results, the use of GSE systematically results in an improvement of recognition rates regardless the gender of speakers.

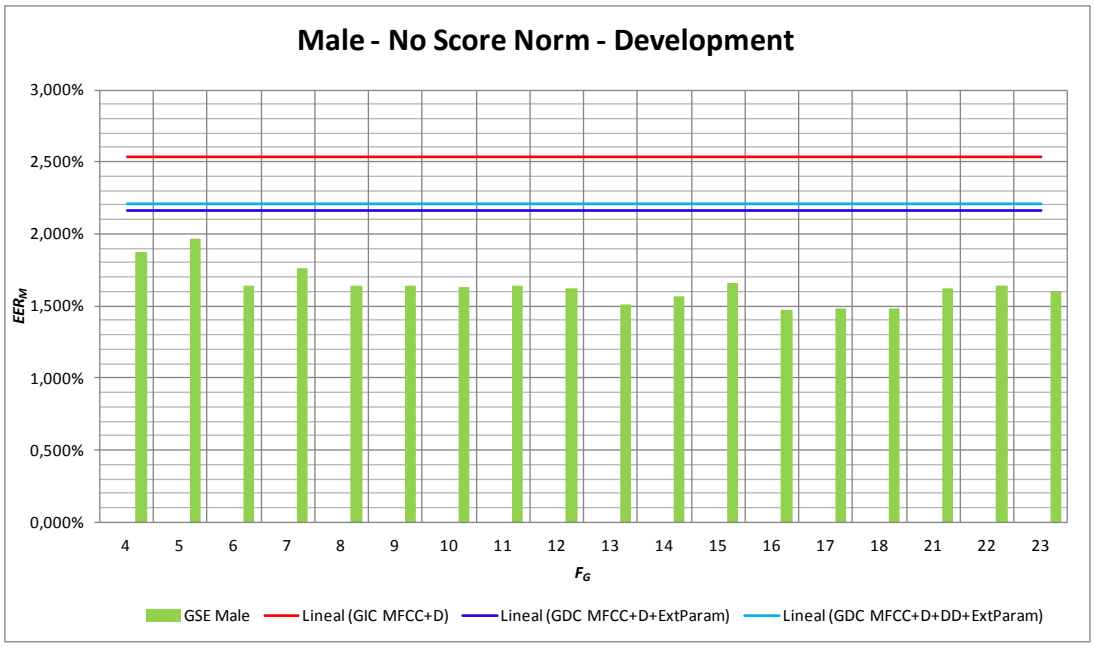



**Figure 5-74 Influence of the GSE configuration on the $EER_M$ (development set)**

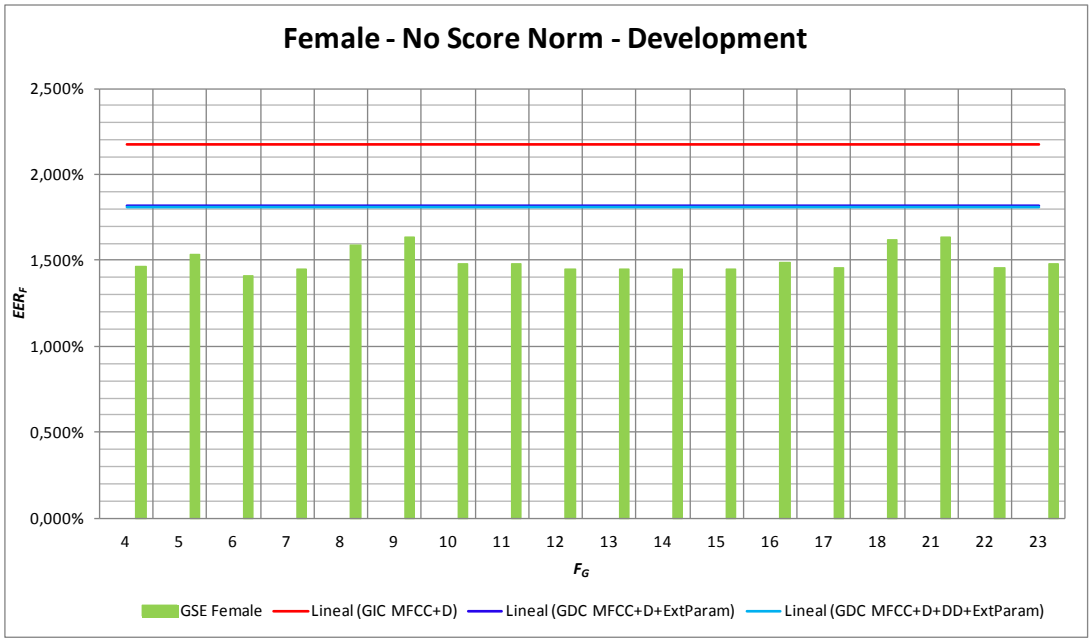



**Figure 5-75** Influence of the GSE configuration on the $EER_F$ (development set)

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=40, *MFCC*=22, *G*=256, *α*=5 | - | - | 2.405% [2.477] | - | 1.818% [2.886] | - | 2.112% [-] |
| **GDC MFCC+Δ** | M | *F*=38, *MFCC*=25, *G*=512, *α*=20 | - | E+F3 | 1.848% [2.794] | 23.15% | 1.655% [3.228] | 8.95% | 1.752% [17.04%] |
| | F | *F*=48, *MFCC*=21, *G*=512, *α*=16 | - | ΔE | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=44, *MFCC*=25, *G*=256, *α*=8 | - | ΔE+F0 | 2.504 [2.406] | -4.09% | 1.818% [2.811] | 0.00% | 2.161% [-2.33%] |
| | F | *F*=30, *MFCC*=24, *G*=256, *α*=5 | - | ΔE | | | | | |
| **GSE** | M | *F*=38, *MFCC*=25, *G*=256, *α*=8 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 13-Channel Filter bank 10 MFCC | E+F3 | 1.496% [2.777] | 37.79% | 1.231% [3.030] | 32.29% | 1.364% [35.42%] |
| | F | *F*=48, *MFCC*=21, *G*=256, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 14- Channel Filter bank 10 MFCC | ΔE | | | | | |

**Table 5-24** $EER_M$, $EER_F$, and *HEER* obtained on the development set (ZNorm), comparing classical parameters with extra parameters and extended biometric parameters

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=50, *MFCC*=26, G=256, *α*=5 | - | - | 2.000% [1.004] | - | 1.455% [1.118] | - | 1.727% [-] |
| **GDC MFCC+Δ** | M | *F*=44, *MFCC*=24, G=256, *α*=5 | - | E+ΔE+F0 | 1.807% [1.199] | 9.65% | 1.424% [1.238] | 2.08% | 1.616% [6.46%] |
| | F | *F*=50, *MFCC*=26, *G*=256, *α*=5 | - | E+ΔE | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=38, *MFCC*=25, *G*=256 , *α*=5 | - | ΔE+F0 | 2.042% [0.490] | -2.08% | 2.000% [0.847] | -37.50% | 2.021% [-16.99%] |
| | F | *F*=44, *MFCC*=22, *G*=256, *α*=5 | - | E+ΔE+F0+F3 | | | | | |
| **GSE** | M | *F*=44, *MFCC*=24, *G*=256, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 17-Channel Filter bank/10 MFCC | E+ΔE+F0 | 1.288% [1.252] | 35.60% | 1.133% [1.151] | 22.13% | 1.210% [29.93%] |
| | F | *F*=50, *MFCC*=26, *G*=256, *α*=5 | **Source-Tract Sep. Alg:** Prediction Order: 20 Forgetting Factor: 0.995 **GSE:** 7-Channel Filter bank /8 MFCC | E+ΔE | | | | | |
| **GSE+VTE** | F | *F*=50, *MFCC*=26, *G*=256, α=5 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 18-Channel Filter bank /6MFCC **VTE:** 40-Channel Filter bank /2MFCC | E+ΔE | - | - | 1.091% [1.270] | 25.02% | - |

**Table 5-25** $EER_M$, $EER_F$, and *HEER* obtained on the development set (TNorm), comparing classical parameters with extra parameters and extended biometric parameters

| Parameters | Gen. | Classic Parameters set up | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | M/F | *F*=34, MFCC=26, *G*=256, *α*=5 | - | - | 2.000% [1.847] | - | 1.655% [2.233] | - | 1.828% [-] |
| **GDC MFCC+Δ** | M | *F*=40, MFCC=22, *G*=256, *α*=8 | - | ΔE+F0 | 1.636% [1.986] | 18.18% | 1.455% [2.305] | 12.12% | 1.545% [15.44%] |
| | F | *F*=44, MFCC=26, *G*=256, *α*=5 | - | F0+F3 | | | | | |
| **GDC MFCC+Δ+ΔΔ** | M | *F*=40, *MFCC*=22, *G*=256, *α*=10 | - | F3 | 2.129% [1.731] | -6.43% | 2.027% [2.030] | -22.42% | 2.078% [-13.67%] |
| | F | *F*=40, *MFCC*=24, *G*=256, *α*=10 | - | - | | | | | |
| **GSE** | M | *F*=40, *MFCC*=22, *G*=512, *α*=8 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 12-Channel Filter bank/8MFCC | ΔE+F0 | 1.273% [2.031] | 36.36% | 1.273% [2.304] | 23.11% | 1.273% [30.36%] |
| | F | *F*=44, *MFCC*=26, *G*=256, *α*=16 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 13-Channel Filter bank /6MFCC | F0+F3 | | | | | |
| **GSE+VTE** | M | *F*=40, *MFCC*=22, *G*=256, *α*=16 | **Source-Tract Sep. Alg:** Prediction Order: 16 Forgetting Factor: 0.995 **GSE:** 13-Channel Filter bank /8MFCC **VTE:** 18-Channel Filter bank /6MFCC | ΔE+F0 | 1.114% [2.092] | 44.31% | - | - | - |

**Table 5-26** $EER_M$, $EER_F$, and *HEER* obtained on the development set (ZTNorm), comparing classical parameters with extra parameters and extended biometric parameters

Another factor that must be analysed is the influence of score normalisation algorithms in recognition rates. For this reason, the same experiments that have been presented so far (when no score normalisation is applied), have been conducted but applying ZNorm, TNorm, and ZTNorm. Table 5-24 to Table 5-26, provide the equivalent results that have been reported in Table 5-23, when the different score normalisations are applied.

Form the results shown on Table 5-23 to Table 5-26, it is clear that no matter whether score normalisation is applied or not, the best results in terms of $EER_X$ or *HEER* are always obtained when GSE information is incorporated on the feature vectors. Moreover, it can be asserted that the use of ΔΔ coefficients does not provide any additional benefit. Additionally, the use of score normalisation algorithms provides a general improvement on all recognition rates: specifically, TNorm provides the best results in terms of $EER_F$ while ZTNorm provides the best results in terms of $EER_M$.

In order to complete previous results, Figure 5-76 and Figure 5-77 provide DET curves for male and female speakers, respectively, so the performance of the *Baseline* front-end (in a GIC MFCCs+Δ setup) can be compared with the GDEB front-end when different score normalisation techniques are applied and only information from the glottal source estimate is included.

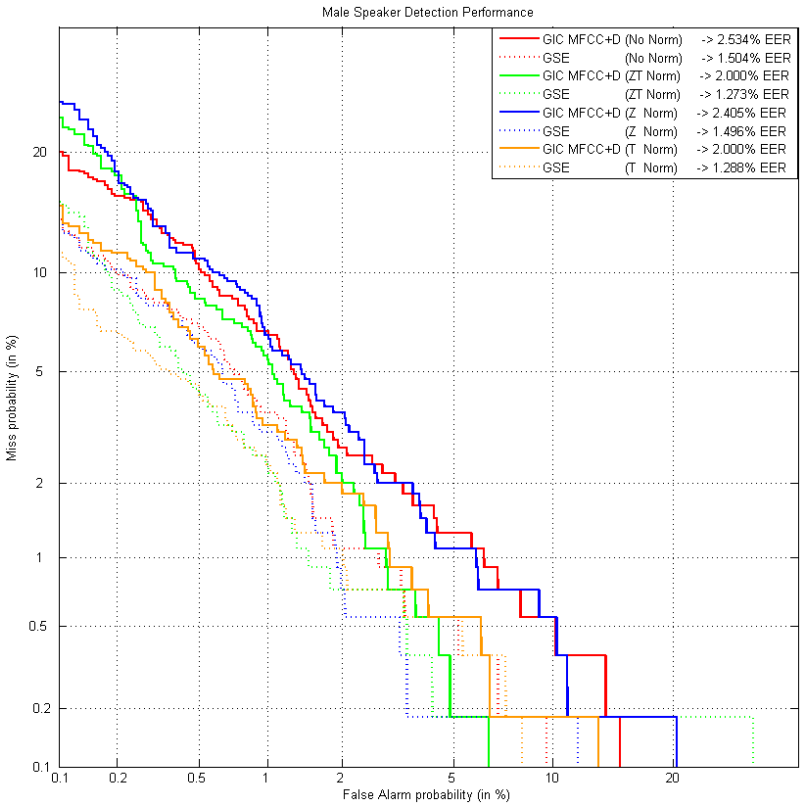



**Figure 5-76** DET curves for *Baseline* and GDEB front-end, applying different score normalisation techniques (male development set)

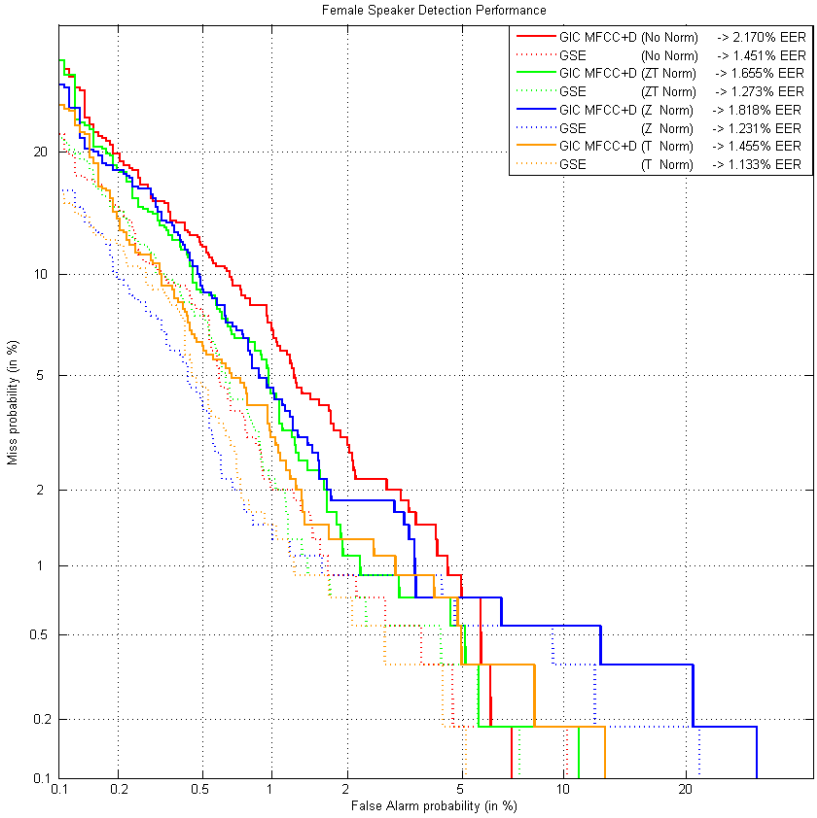



**Figure 5-77** DET curves for *Baseline* and GDEB front-end, applying different score normalisation techniques (female development set)

To complete the study on the ALBAYZIN database, there are still two additional tasks to perform. First of all, we have to verify the usefulness of the information provided by the vocal tract estimate (VTE), which we have not yet analysed. As we are working on a text-independent scenario, chances are on VTE not been as useful as GSE in order to accurately characterise a speaker. For this reason we only run an additional test based on the most successful male and female configurations in terms of $EER_M$ and $EER_F$, i.e. GSE ZTNorm for male speakers and GSE TNorm for female speakers.

In the case of the tests run on male speakers, the use of GSE information combined with VTE information systematically produces recognition rates that outperform the recognition rates obtained by the GIC MFCCs+Δ and the GDC MFCCs+Δ+ExtParam. configurations. However, it is difficult to find a configuration that outperforms the recognition rates obtained when incorporating GSE information into the feature vector, which leads us to conclude that the GSE provides more relevant speaker's biometric information than the VTE. Nevertheless, we have found few configurations providing better recognition rates in terms of $EER_M$ (see Table 5-26). Specifically, adding 6 MFCCs extracted from the VTE using 16 filters, a relative reduction of 44.31% in terms of $EER_M$, respect to GIC MFCCs+Δ is obtained, which means that a relative reduction of 12.5% is obtained respect to the configuration in which only GSE information is used. Figure 5-78 provides the DET curves comparing results shown on Table 5-26.

The same explanation is valid for the case of female speakers, as there are a few configurations, including VTE information, that provide better recognition rates in terms of $EER_F$, than the ones obtained with the different configurations tested so far. Specifically (see Table 5-25), adding 2 MFCCs extracted from the VTE using 40 filters, a relative reduction of 25% in terms of $EER_F$, respect to GIC MFCCs+Δ is obtained,

which means that a relative reduction of 3.67% is obtained respect to the configuration in which only GSE information is used. Figure 5-79 provides the DET curves comparing results shown on Table 5-25.

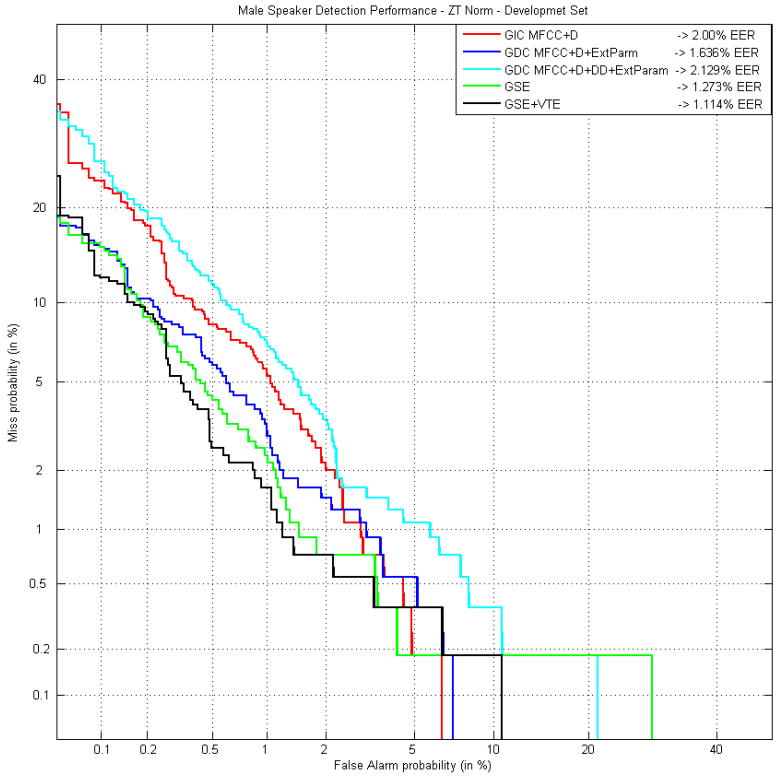



**Figure 5-78** DET curves comparing classical parameters and GDEB on ALBAYZIN development set for male speakers and ZTNorm

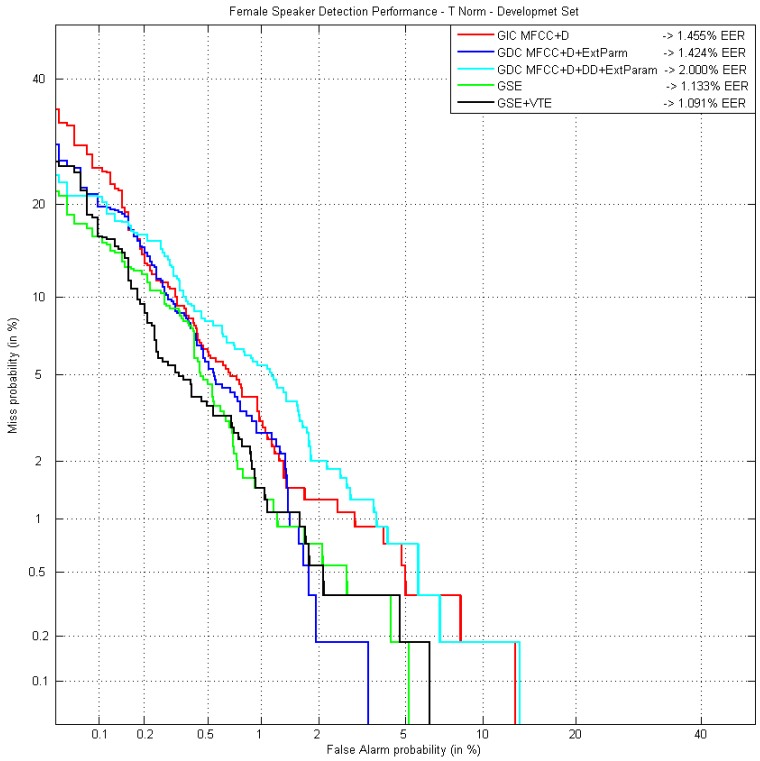



**Figure 5-79** DET curves comparing classical parameters and GDEB on ALBAYZIN development set for female speakers and TNorm

Finally, once the settings are fixed for both *Baseline* and GDEB front-ends, and a score threshold is established on the development set, the actual system performance can be checked out using the evaluation set. As we are facing the systems to new unknown data, we can verify if the behaviour of the speaker recognition system with the selected parameters holds for the evaluation set, or if the results obtained are affected by overtraining on the development set. Table 5-27 provides the results obtained on the evaluation set using the different configuration previously selected (see Table 5-23 to Table 5-26) for different score normalisation techniques, and the selected threshold.

The results obtained in terms of $HTER_X$ lead to two important conclusions. First of all, we can assert that the use of GSE information conveniently parameterised provides an improvement in recognition rates which remains consistent over the development set, and the evaluation set. Specifically, for the female case when applying TNorm we achieve a relative reduction of 25% in terms of $EER_F$ (from $EER_F$=1.455%, when classical gender-independent characterisation is used, to $EER_F$=1.091%), which is transformed into a relative reduction of 20% in terms of $HTER_F$, when moving to evaluation set (from $HTER_F$=2.835%, when classical gender-independent characterisation is used, to $HTER_F$=2.262%). For the case of male speakers, when ZTNorm is applied, we achieve a relative reduction of 36% in terms of $EER_M$ (from $EER_F$=2.000%, when classical gender-independent characterisation is used, to $EER_M$=1.273%), which is transformed into a relative reduction of 29% in terms of $HTER_M$, when moving to evaluation set (from $HTER_M$=2.783%, when classical gender-independent characterisation is used, to $HTER_M$=1.977%).

Surprisingly, the inclusion on VTE information on the feature vector seems to be helpful for both male and female speakers. However, this result should be treated with caution, especially by two important reasons. First of all, unlike in the case of information extracted from GSE, the use of VTE information only provides a slight improvement under limited and specific settings on the development set. Secondly, this improvement in terms of $EER_X$ is not always reflected on the evaluation set. As it has been already pointed out, the VTE provides information which is more related to the phonetic content of the message, therefore, in order to be useful in the speaker recognition task (especially in a text-independent scenario), the training information is required to provide sufficient phonetic coverage representative of the message that can be found lately on trials. Although some improvements have been reported on the development and evaluation sets, clearly, the training information for a proper VTE use seems to be not sufficient or adequate.

Additionally, the results show that the improvement obtained on the development set when using the $\Delta\Delta$ coefficients and no score normalisation was applied for male speakers, were due to an overtraining on the development set, as the results obtained for the same configuration on the evaluation set were far worse than the ones obtained by the rest of tested configurations. Therefore, the use of $\Delta\Delta$ coefficients which offered some advantage, particularly for female speakers in a text-constrained context, appears to be clearly inefficient in a text-independent scenario.

| Score Norm | Parameters | $EER_M$ [$\theta_M$] | $HTER_M$ | $HTER_M$ RR | $EER_F$ [$\theta_F$] | $HTER_F$ | $HTER_F$ RR |
|---|---|---|---|---|---|---|---|
| No Norm | **GIC MFCC+Δ** | 2.534% [-0.178] | 3.347% | - | 2.170% [-0.169] | 3.250% | - |
| | **GDC MFCC+Δ+ExtParm.** | 2.163% [-0.035] | 3.089% | 7.70% | 1.818% [-0.113] | 3.094% | 4.79% |
| | **GDC MFCC+Δ+ΔΔ+ExtParm.** | 2.205% [0.032] | 5.565% | -66.26% | 1.807% [0.016] | 8.383% | -157.92% |
| | **GSE** | 1.504% [-0.131] | 2.189% | 34.61% | 1.451% [-0.145] | 2.673% | 17.74% |
| ZTNorm | | | | | | | |
| | **GIC MFCC+Δ** | 2.000% [1.847] | 2.783% | - | 1.655% [2.233] | 3.081% | - |
| | **GDC MFCC+Δ+ExtParm.** | 1.636% [1.986] | 2.432% | 12.59% | 1.455% [2.305] | 2.870% | 6.85% |
| | **GDC MFCC+Δ+ΔΔ+ExtParm.** | 2.129% [1.731] | 3.513% | -26.22% | 2.027% [2.030] | 3.287% | -6.68% |
| | **GSE** | 1.273% [2.031] | 1.977% | 28.94% | 1.273% [2.304] | 2.709% | 12.07% |
| | **GSE+VTE** | 1.114% [2.092] | 1.917% | 31.12% | - | - | - |
| TNorm | | | | | | | |
| | **GIC MFCC+Δ** | 2.000% [1.004] | 2.806% | - | 1.455% [1.118] | 2.835% | - |
| | **GDC MFCC+Δ+ExtParm.** | 1.807% [1.199] | 2.555% | 8.93% | 1.424% [1.238] | 2.598% | 8.36% |
| | **GDC MFCC+Δ+ΔΔ+ExtParm.** | 2.042% [0.490] | 3.077% | -9.66% | 2.000% [0.847] | 2.865% | -1.02% |
| | **GSE** | 1.288% [1.252] | 1.812% | 35.42% | 1.133% [1.151] | 2.289% | 19.28% |
| | **GSE+VTE** | - | - | - | 1.091% [1.270] | 2.262% | 20.21% |
| ZNorm | | | | | | | |
| | **GIC MFCC+Δ** | 2.045% [2.477] | 3.388% | - | 1.818% [2.886] | 3.075% | - |
| | **GDC MFCC+Δ+ExtParm.** | 1.848% [2.794] | 3.040% | 10.25% | 1.655% [3.228] | 3.203% | -4.16% |
| | **GDC MFCC+Δ+ΔΔ+ExtParm.** | 2.504% [2.406] | 3.290% | 2.89% | 1.818% [2.811] | 3.199% | -4.04% |
| | **GSE** | 1.496% [2.777] | 1.980% | 41.55% | 1.231% [3.030] | 2.635% | 14.31% |

**Table 5-27** $HTER_X$ obtained for selected configurations on evaluation set, applying different score normalisations

Finally, Figure 5-80 to Figure 5-87, provide the DET curves that represent the results obtained in the evaluation set for male and female speakers, which are reflected in Table 5-27.

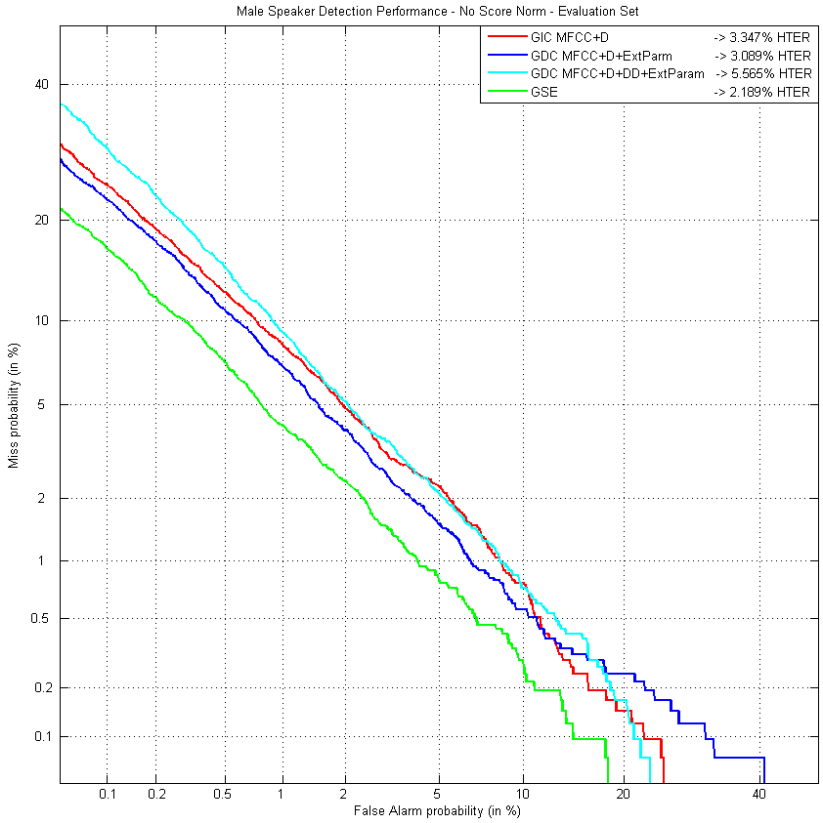



**Figure 5-80** Male DET curves on ALBAYZIN evaluation set, without applying any score normalisation technique

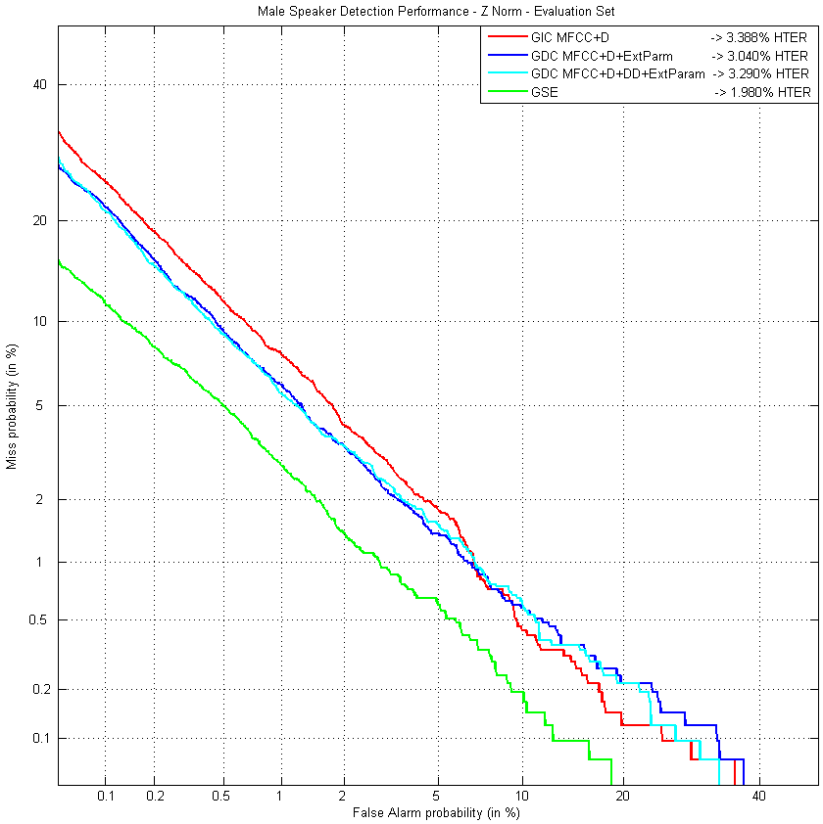



**Figure 5-81** Male DET curves on ALBAYZIN evaluation set, applying ZNorm

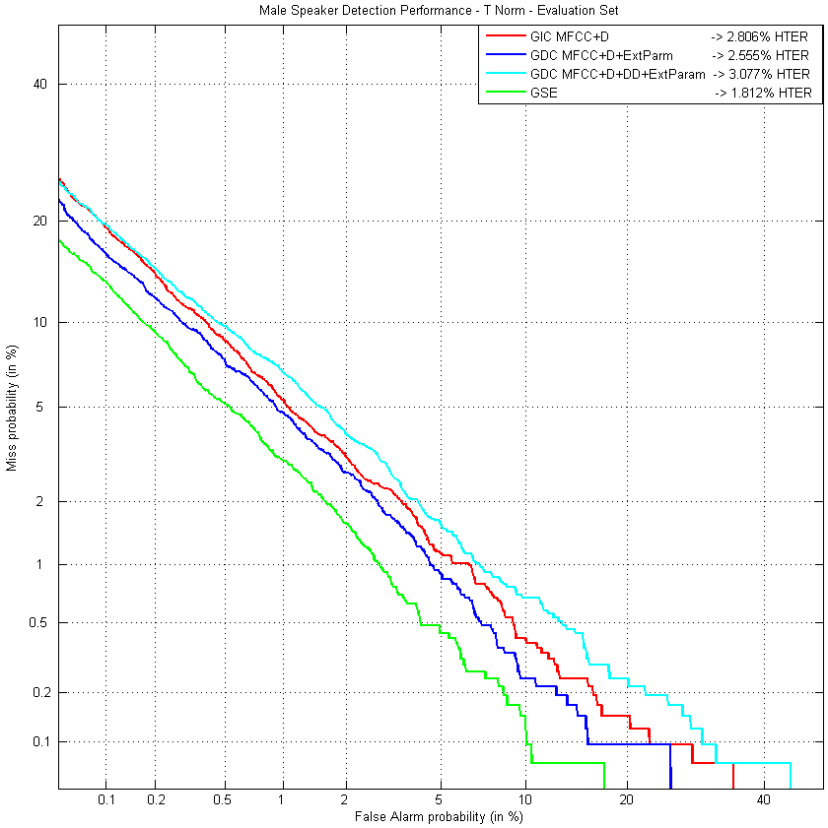



**Figure 5-82** Male DET curves on ALBAYZIN evaluation set, applying TNorm

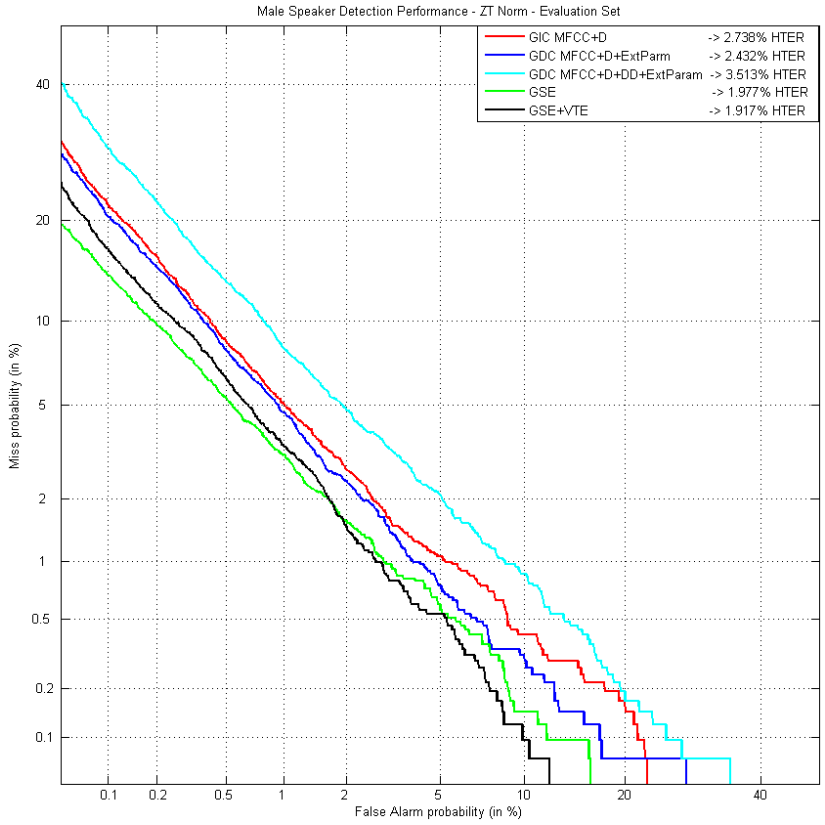



**Figure 5-83** Male DET curves on ALBAYZIN evaluation set, applying ZTNorm

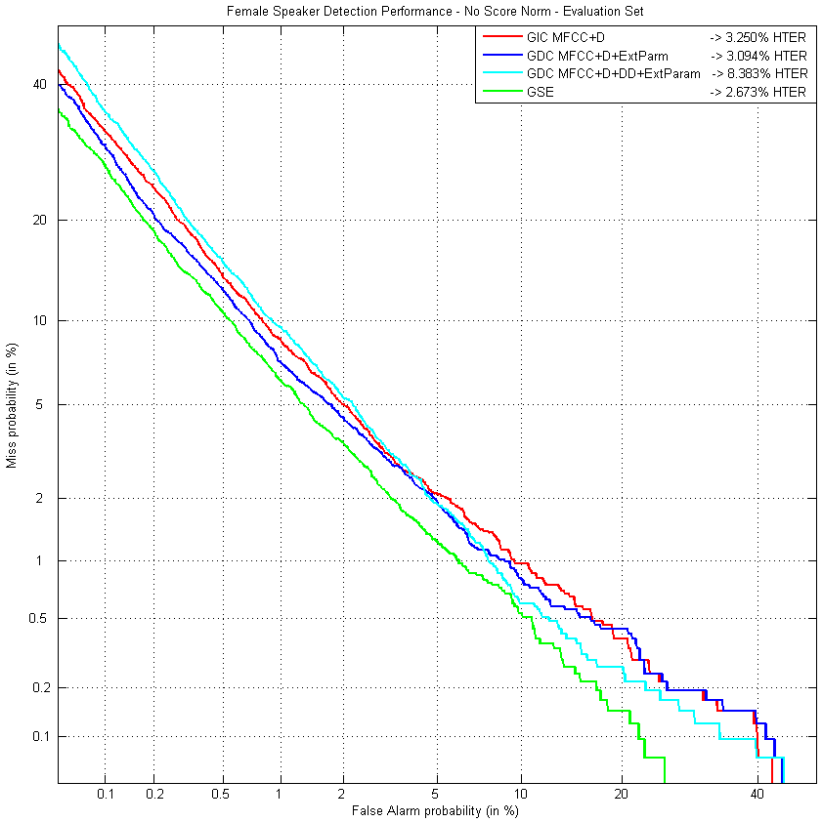



**Figure 5-84** Female DET curves on ALBAYZIN evaluation set, without applying any score normalisation technique

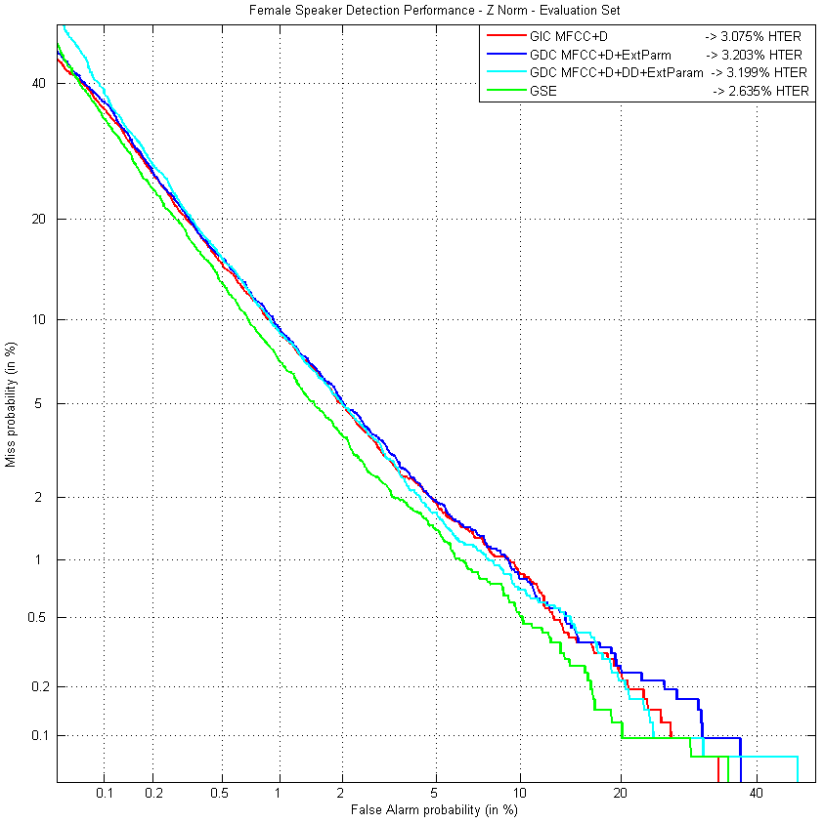



**Figure 5-85** Female DET curves on ALBAYZIN evaluation set, applying ZNorm

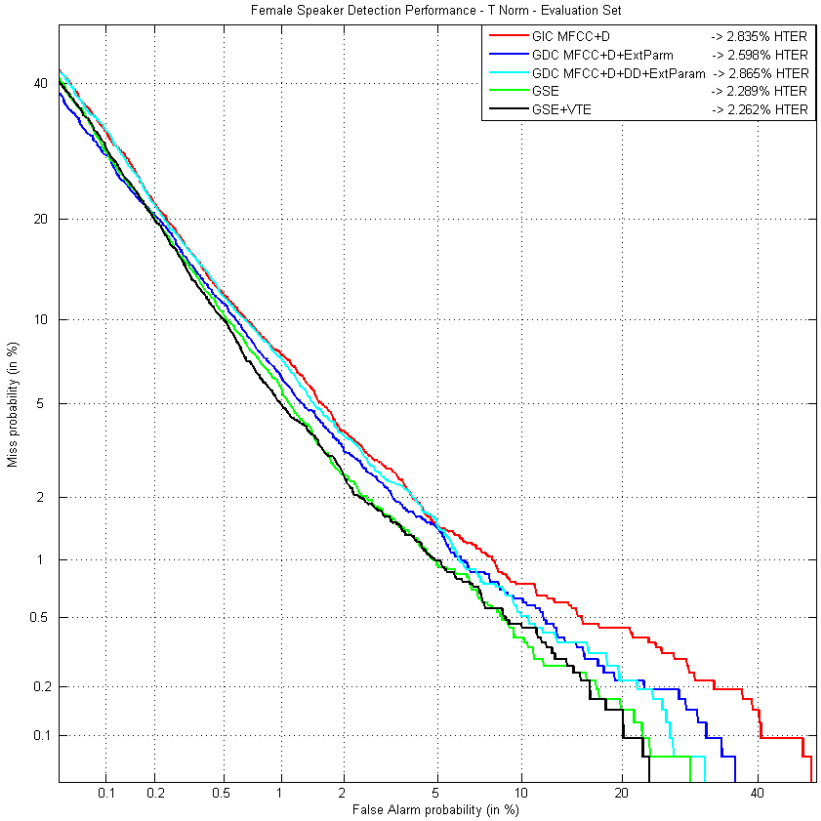



**Figure 5-86** Female DET curves on ALBAYZIN evaluation set, applying TNorm

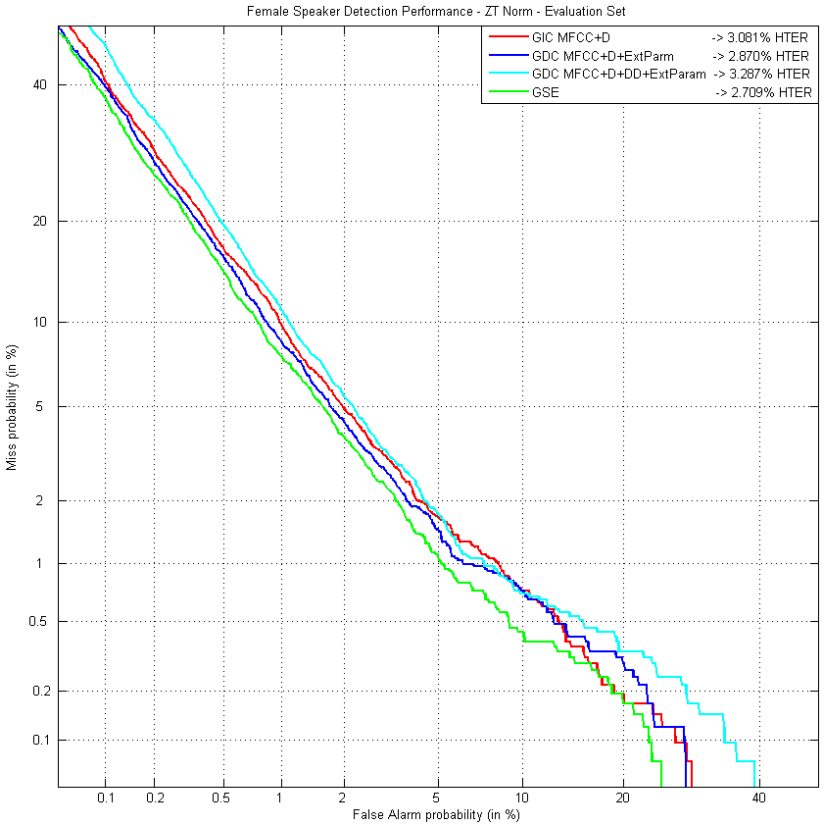



**Figure 5-87** Female DET curves on ALBAYZIN evaluation set, applying ZTNorm

There is an important aspect that must be noticed based on the DET curves previously presented, which refers to the behaviour of the system on the evaluation set. When

testing the system in the evaluation set, we have already defined an operation point based on development set results. Therefore we are not evaluating the performance of our systems at all points but at the specific one given by $\theta_{dev}$. However, as the interested reader may have already noticed, in this text-independent scenario, the DET curves provided by the gender-dependent configuration incorporating GSE (or GSE+VTE) information presents better results than the rest of configurations at almost all points of the curve for all of the score normalisation algorithms applied and both genders. Therefore, in this scenario we have not only minimised *EER* and thus *HTER*, but also *AUC*.

#### *5.2.2.1 Brief Conclussions*

To conclude this section, we must point out that in the closed-set text-independent scenario designed using the ALBAYZIN database, the use of a gender-dependent extended biometric parameterisation in which VTE and GSE information have been incorporated provides a clear improvement in terms of recognition rates respect to the use of the classic gender-independent approach. This improvement remains consistent over the development set and the evaluation set.

In a more specific way we can highlight the following aspects. In order to improve recognition rates it is essential to use a gender-dependent parameterization. The trend shown in previous scenarios regarding the number of channels in the filter bank used to compute the MFCCs is confirmed in this scenario as well. In other words, the number of channels in the optimal filter bank is higher for female speakers than for male speakers, regardless the score normalization technique applied. Regarding the number of MFCCs (in the MFCCs+Δ configuration), it remains in the range of 21 to 26 for both male and female speakers, being the optimal value different depending on the gender, except for the case of not applying any score normalization technique in which *MFCC*=26 for both genders.

Like in the previous scenario, the use of ΔΔ coefficients is completely discouraged as the recognition rates obtained with configurations including these coefficients, are worse than in the case of using just MFCCs+Δ, especially for male speakers.

Additionally, it has been confirmed that the use of extra parameters under some specific combination helps to increase recognition rates. However, in this particular scenario the proposed F3 coefficient does not systematically appear in the optimal configuration for all types of score normalization techniques for male speakers, just for the case of applying ZNorm. However, F0 appears to be an important coefficient for male speakers in this scenario. In the case of female speakers, the use of F3 helps in the increase of recognition rates for almost all of the score normalization techniques, alone or combined with other extra parameters.

Regarding the use of score normalisation techniques, it must be noted that any of those tested, provide an extra value in the reduction of error rates, but obviously they entail additional computational costs to the speaker recognition system. The fact that TNorm provides better results than the ZNorm or even ZTNorm for female speakers, in terms of reduction of error rates, suggests that the number of impostors selected (25) is not enough or even representative of the set of alternative speakers. Therefore the obtained performance is been penalised if compared with TNorm for which 625 recordings are used in the normalisation process, and seems to cover better the impostor space.

Finally, it must be noted that contrary to what might be expected, the used of specific MFCCs extracted from the vocal tract estimate, provides extra value in improving recognition rates in this text-independent scenario.

### 5.2.3 Text-Independent Speaker Recognition in Mobile Environments

In this section we are going to present the results and conclusions obtained on the competition on Speaker Recognition in Mobile Environments using the MOBIO database, which took place during the 6th IAPR International Conference on Biometrics (ICB-2013) [Khoury,2013], since the set of tests designed were aimed at the participation in this international evaluation. Nevertheless, we will also analyse the performance of the *Baseline* front-end versus the GDEB front-end.

As previously presented, the evaluation plan splits the database into 3 different sets: the background training set used to learn the background parameters of the algorithm (UBM, subspaces, etc.) or for normalisation purposes, the development set supposed to be used to tune meta-parameters of the algorithm, and the evaluation set used to analyse the performance of recognition systems. In order to evaluate the performance of the systems like in previous scenarios, *EER* and *HTER* are used as quality measures (see Eq. (5-4) to Eq. (5-6)) in the development and evaluation set respectively.

Based on these metrics we have run a battery of tests using the *Baseline* front-end in order to minimise the *EER*. However, as no cross-gender trials are going to be present, we can use again the Half Equal Error Rate metric, *HEER*:

$$HEER = \frac{EER_M(MFCC, F, \Delta, \Delta\Delta, G, \alpha) + EER_F(MFCC, F, \Delta, \Delta\Delta, G, \alpha)}{2} \qquad \textbf{Eq. (5-9)}$$

In this case, we are not going to perform a deep search like in previous scenarios. Instead, we will make use of previous experience to limit the number of tests. Thus, *MFCC*={16,18,20,21,22,23,24,25,26,27,28}, as in previous scenarios we have confirmed that MFCC values lower than 16 do not usually provide any improvement; *F*={30,34,38,40,44,48,50}, for the same reason; $\Delta$ and $\Delta\Delta$ have been set to true and false respectively, as we have verified that in a text-independent context the use of $\Delta\Delta$ does not provide any additional benefit. In the case of the number of Gaussians used to build both the UBM and the speaker's models, we have also used 1024 Gaussians (in addition to 512 and 256), as the amount of available training data is higher than in the previous scenarios. Finally the relevance factor has been increased to six values $\alpha$={5,8,10,16,20,24}.

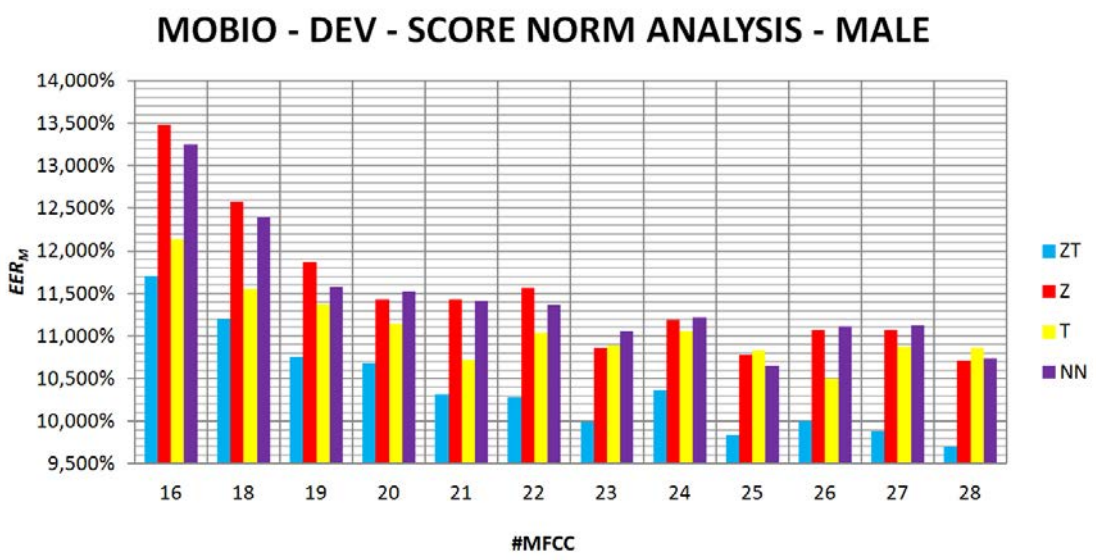
We will present the results obtained using both the *Baseline* and the GDEB front-ends, when applied to the GMM-UBM approach. Neither the SV-GMM nor the *i*-vector approaches will be used on these tests mainly due to the limited amount of different speakers in the database.

Regarding the score normalization techniques we have tested all that have been used in previous scenarios, i.e. ZNorm, TNorm, ZTNorm, and obviously the case in which no score normalization is applied. However, the analysis of the different types of score normalizations will be limited to the first stage, and will not be extended to the tests in which extended biometrics are used. We can proceed this way as in previous scenarios, we have found that the score normalization that obtains better performance when applied to a gender-dependent configuration extended with extra parameters (namely, E, $\Delta$E, F0, and F3), will continue to offer the most successful results when extended biometrics are incorporated.

Although obvious, the *Baseline* and the GDEB front-end present the same behaviour when we work with classical parameters. Thus we can go straight to the analysis of the gender-dependent configuration in terms of $EER_X$, and simply reference the configuration which would provide the most successful results in terms of *HEER*, when a gender-independent configuration is used, so we can set a baseline for comparing results. We can proceed this way as from previous results it is clear that a gender-dependent configuration always provides better or at least the same results than a gender-independent one.

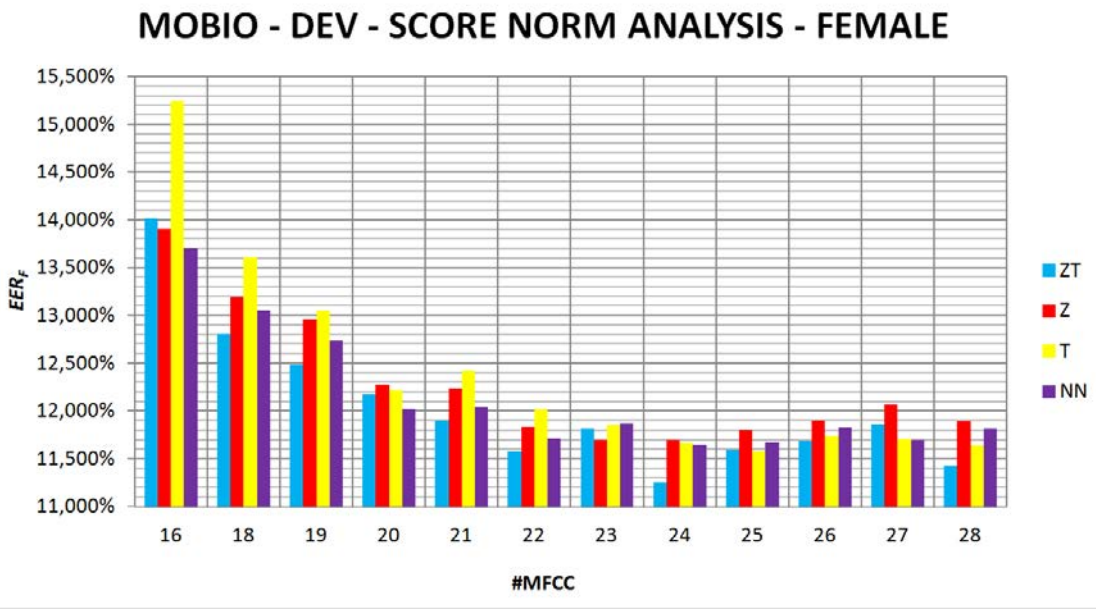
The first set of figures represents the results obtained in terms of $EER_X$, based on each of the configuration parameters in Eq. (5-9), assuming that we are using the GDEB front-end, thus a gender-dependent configuration (labelled as GDC) but just including classical parameters into the feature vector. Like in the previous scenarios, the process followed consists in fixing a value for a specific parameter, and test for the rest of configurable parameters which configuration provides better results in terms of *EER*.

First of all we highlight the influence of the number of MFCCs on the recognition rates depending on the gender of the speaker. Additionally, the influence of the score normalization algorithms has been analysed (light blue for ZTNorm, red for ZNorm, yellow for TNorm and purple for the case in which no score normalization is applied).




| | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZT Norm | 11,70% | 11,19% | 10,76% | 10,68% | 10,31% | 10,27% | 9,99% | 10,36% | 9,83% | 10,00% | 9,88% | 9,70% |
| Z Norm | 13,48% | 12,57% | 11,86% | 11,42% | 11,42% | 11,56% | 10,86% | 11,18% | 10,77% | 11,07% | 11,07% | 10,71% |
| T Norm | 12,13% | 11,55% | 11,37% | 11,14% | 10,72% | 11,03% | 10,88% | 11,05% | 10,83% | 10,49% | 10,87% | 10,86% |
| No Norm | 13,24% | 12,39% | 11,57% | 11,52% | 11,40% | 11,36% | 11,05% | 11,21% | 10,65% | 11,11% | 11,12% | 10,73% |

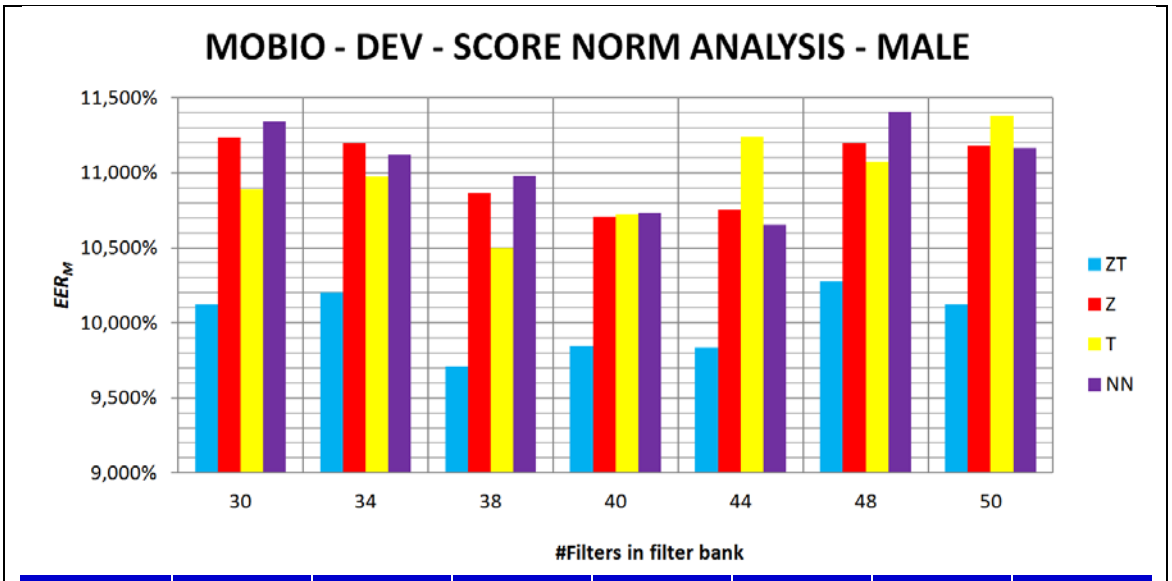
**Figure 5-88** $EER_M$ obtained depending on the number of MFCCs (GDC – development set) and for different score normalization algorithms




| | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZT Norm | 14,02% | 12,80% | 12,48% | 12,17% | 11,90% | 11,58% | 11,81% | 11,25% | 11,59% | 11,69% | 11,85% | 11,42% |
| Z Norm | 13,90% | 13,18% | 12,96% | 12,27% | 12,23% | 11,83% | 11,69% | 11,69% | 11,80% | 11,90% | 12,06% | 11,89% |
| T Norm | 15,24% | 13,60% | 13,05% | 12,22% | 12,42% | 12,01% | 11,85% | 11,66% | 11,57% | 11,74% | 11,70% | 11,63% |
| No Norm | 13,69% | 13,05% | 12,73% | 12,02% | 12,03% | 11,71% | 11,86% | 11,64% | 11,67% | 11,82% | 11,69% | 11,81% |

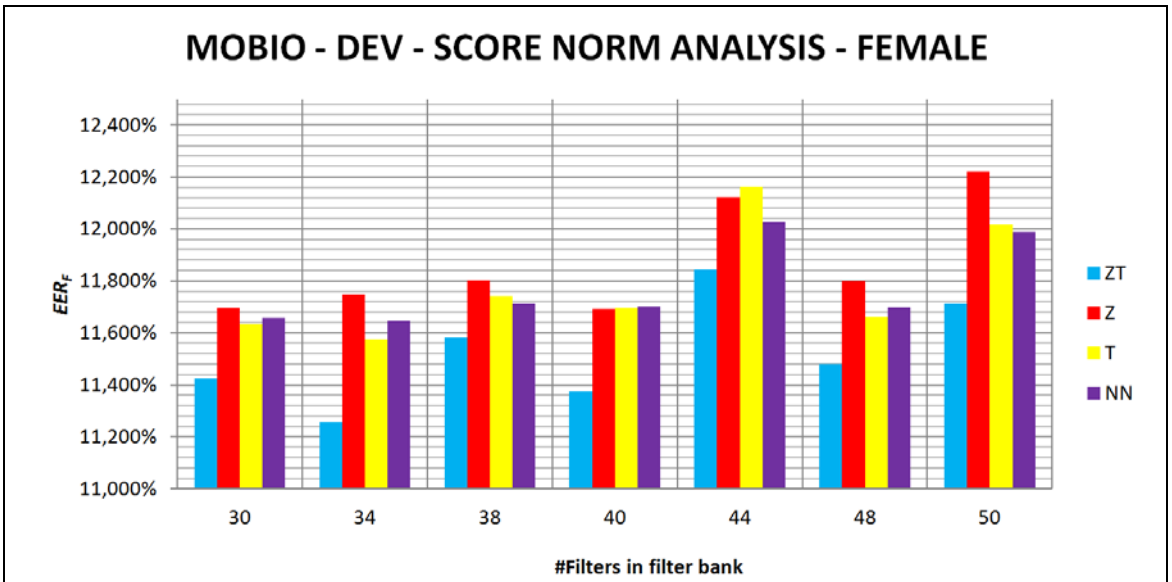
**Figure 5-89** $EER_F$ obtained depending on the number of MFCCs (GDC – development set) and for different score normalization algorithms

Figure 5-88 shows that in the case of male speakers, a minimum in terms of $EER_M$ is reached for the specific case of using 28 MFCCs when ZTNorm is applied. Additionally, regardless the score normalization algorithm applied, most successful results are obtained for MFCC values higher than 25. Regarding the score normalization algorithms, it is clear that the use of ZTNorm (light blue columns) systematically produces better recognition rates than any other tested normalization. In the case of female speakers, the configuration that provides the most successful results in terms of $EER_F$ is the one which uses ZTNorm for score normalization purposes and 24 MFCCs. In this case, the number of MFCCs seems to be quite stable regardless the score normalization algorithm applied. However, although the minimum is reached for the ZTNorm configuration, this score normalization is not always providing the most successful results (regarding the number of MFCCs), which indicates that the amount of available information for normalization purposes in the case of female speakers is not enough. Finally, it must be noted that as previously mentioned, feature vectors include both MFCCs and Δ, which actually means that in the case of male speakers the dimensionality of the feature vector, that will generate an accurate model is 56 (28 MFCCs+ 28 ΔMFCCs), while in the case of female speakers is 48 (24MFCCs+ 24 ΔMFCCs).




| | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|
| ZTNorm | 10,121% | 10,202% | 9,708% | 9,845% | 9,837% | 10,276% | 10,119% |
| ZNorm | 11,235% | 11,193% | 10,865% | 10,710% | 10,752% | 11,193% | 11,183% |
| TNorm | 10,889% | 10,974% | 10,499% | 10,722% | 11,240% | 11,073% | 11,383% |
| No Norm | 11,342% | 11,122% | 10,977% | 10,736% | 10,654% | 11,406% | 11,167% |

**Figure 5-90** $EER_M$ obtained depending on the number of filters in the filter bank (GDC – development set) and for different score normalization algorithms

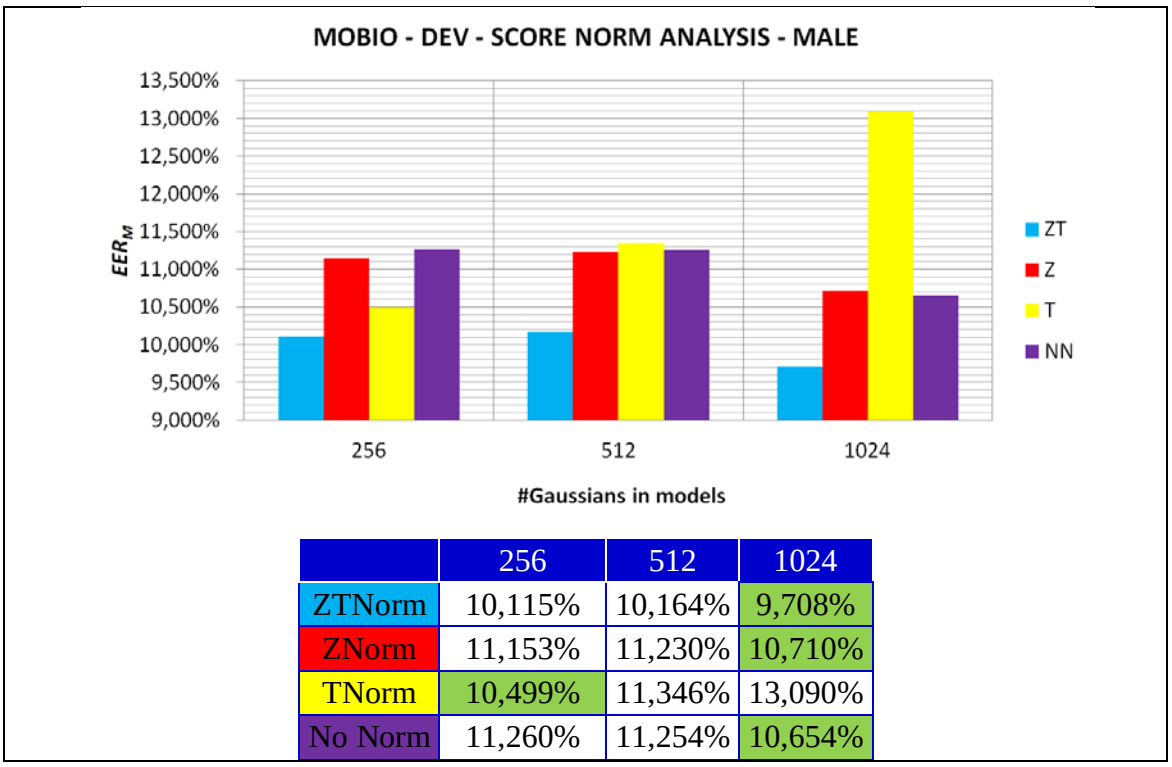



| | 30 | 34 | 38 | 40 | 44 | 48 | 50 |
|---|---|---|---|---|---|---|---|
| ZTNorm | 11,425% | 11,256% | 11,581% | 11,376% | 11,843% | 11,480% | 11,712% |
| ZNorm | 11,694% | 11,748% | 11,804% | 11,693% | 12,123% | 11,799% | 12,221% |
| TNorm | 11,634% | 11,576% | 11,741% | 11,696% | 12,162% | 11,660% | 12,018% |
| No Norm | 11,657% | 11,646% | 11,713% | 11,701% | 12,025% | 11,698% | 11,987% |

**Figure 5-91** $EER_F$ obtained depending on the number of filters in the filter bank (GDC – development set) and for different score normalization algorithms

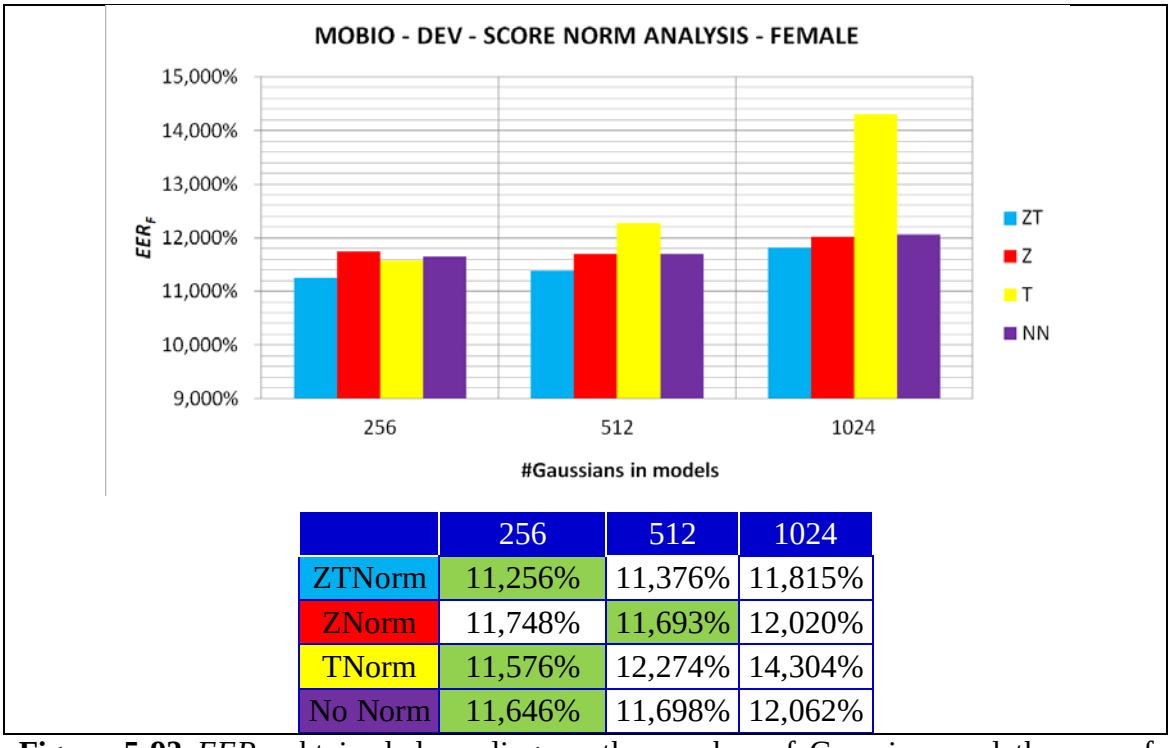
If we consider the number of filters forming the filter bank used to evaluate MFCCs, differences arise again depending on the gender of the speakers. Specifically, Figure

5-90 shows that for male speakers a minimum, in terms of $EER_M$, is reached in the case of using 38 filters in the filter bank, whereas in the case of female speakers the most successful results are obtained with a filter bank of 34 filters.




| | 256 | 512 | 1024 |
|---|---|---|---|
| ZTNorm | 10,115% | 10,164% | 9,708% |
| ZNorm | 11,153% | 11,230% | 10,710% |
| TNorm | 10,499% | 11,346% | 13,090% |
| No Norm | 11,260% | 11,254% | 10,654% |

**Figure 5-92** $EER_M$ obtained depending on the number of Gaussians and the use of different score normalizations (GDC – development set)




| | 256 | 512 | 1024 |
|---|---|---|---|
| ZTNorm | 11,256% | 11,376% | 11,815% |
| ZNorm | 11,748% | 11,693% | 12,020% |
| TNorm | 11,576% | 12,274% | 14,304% |
| No Norm | 11,646% | 11,698% | 12,062% |

**Figure 5-93** $EER_F$ obtained depending on the number of Gaussians and the use of different score normalizations (GDC – development set)

Regarding the number of Gaussians used to build the UBM as well as the models of the target speakers, we found again some differences concerning gender. However, these

differences can be explained by the fact that the amount of data available for female speakers is less than for male speakers in the training set, thus leading to a need of less Gaussians to generate the UBM.

After this brief analysis, we have verified that the setup providing most successful results in terms of $EER_X$, and thus in terms of *HEER*, is different depending on the gender of the speaker under analysis. If we had used the *Baseline* front-end (thus a gender-independent configuration – GIC), the selected setup would have been different and what is more, it would have provided worse recognition rates in terms of *HEER*, and in terms of $EER_X$. Table 5-28 provides a comparison between the recognition rates obtained by the system, when a GDC or a GIC is used (the best results obtained so far are highlighted in light green), and when different score normalizations are applied. Additional columns have been added ($EER_X$ RR), which provide the relative reduction obtained by GDC, in terms of $EER_X$, respect to the corresponding GIC. The relative reduction in terms of *HERR* respect to GIC has been indicated in brackets in the *HEER* column.

Additionally, Figure 5-94 and Figure 5-95 show the DET plots for both male and female speakers from both configurations GIC and GDC. It must be noted that, although having the same name, GDC is different depending on the gender (see Table 5-28). In the case of male speakers, see Figure 5-94, it is clear that except for the case of applying TNorm, the gender dependent configuration (dot line) provides more successful results than the gender independent configuration (solid line); not only in terms of *EER*, but for almost all the points of the DET curve. In the case of female speakers, as the difference between GIC and GDC in terms of $EER_F$ is quite small, differences between corresponding DET curves are also minimal and hardly noticeable.

| Score Norm | Parameters | Gen. | Classic Parameters set up | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR | *HEER [RR]* |
|---|---|---|---|---|---|---|---|---|
| ZTNorm | | | | | | | | |
| | **GIC MFCC+Δ** | M/F | *F*=34, MFCC=24, *G*=256, *α*=20 | 10.594% [1.556] | - | 11.256% [1.545] | - | 10.925% [-] |
| | **GDC MFCC+Δ** | M | *F*=38, MFCC=28, *G*=1024, *α*=24 | 9.708% [1.406] | 8.36% | 11.256% [1.545] | 0.00% | 10.482% [4.05%] |
| | | F | *F*=34, MFCC=24, *G*=256, *α*=20 | | | | | |
| ZNorm | | | | | | | | |
| | **GIC MFCC+Δ** | M/F | *F*=30, MFCC=27, *G*=1024, *α*=24 | 11.256% [1.663] | - | 12.062% [2.145] | - | 11.659% [-] |
| | **GDC MFCC+Δ** | M | *F*=40, MFCC=28, *G*=1024, *α*=24 | 10.710% [1.704] | 4.85% | 11.693% [2.168] | 3.06% | 11.202% [3.92%] |
| | | F | *F*=40, MFCC=24, *G*=512, *α*=10 | | | | | |
| TNorm | | | | | | | | |
| | **GIC MFCC+Δ** | M/F | *F*=38, MFCC=26, *G*=256, *α*=10 | 10.499% [1.014] | - | 11.741% [0.640] | - | 11.120% [-] |
| | **GDC MFCC+Δ** | M | *F*=38, MFCC=26, *G*=256, *α*=10 | 10.499% [1.014] | 0.00% | 11.576% [0.674] | 1.40% | 11.038% [0.74%] |
| | | F | *F*=34, MFCC=25, *G*=256, *α*=10 | | | | | |
| No Norm | | | | | | | | |
| | **GIC MFCC+Δ** | M/F | *F*=30, MFCC=27, *G*=256, *α*=24 | 11.567% [-0.007] | - | 11.693% [0.004] | - | 11.630% [-] |
| | **GDC MFCC+Δ** | M | *F*=44, MFCC=25, *G*=1024, *α*=24 | 10.654% [0.014] | 7.89% | 11.646% [0.011] | 0.40% | 11.150% [4.12%] |
| | | F | *F*=34, MFCC=24, *G*=256, *α*=24 | | | | | |

**Table 5-28** GDC vs. GIC for MOBIO development set (RR – Relative Reduction)

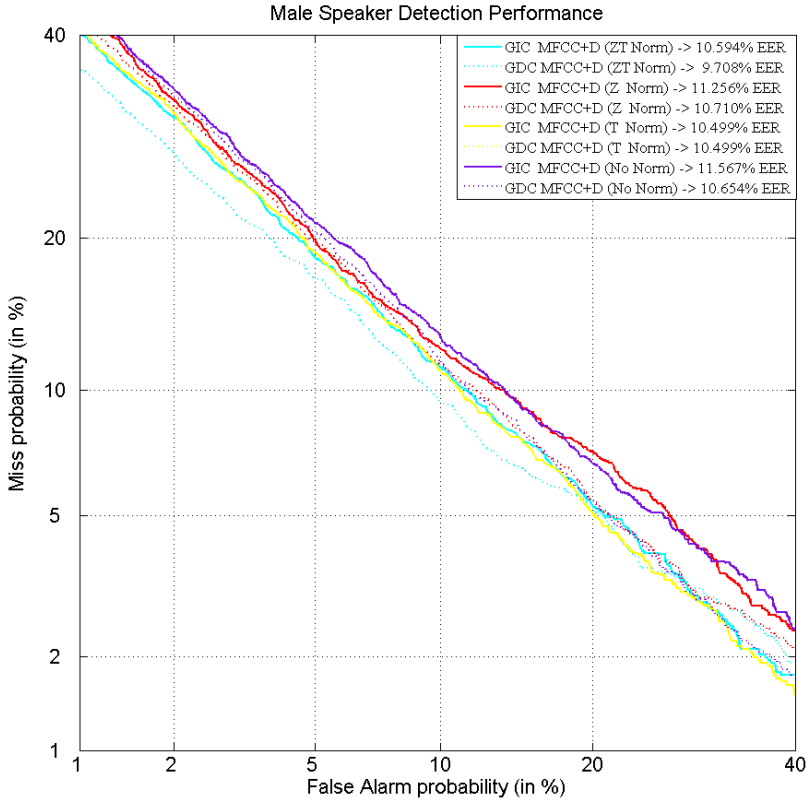



**Figure 5-94** DET curve for classic parameters on MOBIO male development set for GIC and GDC

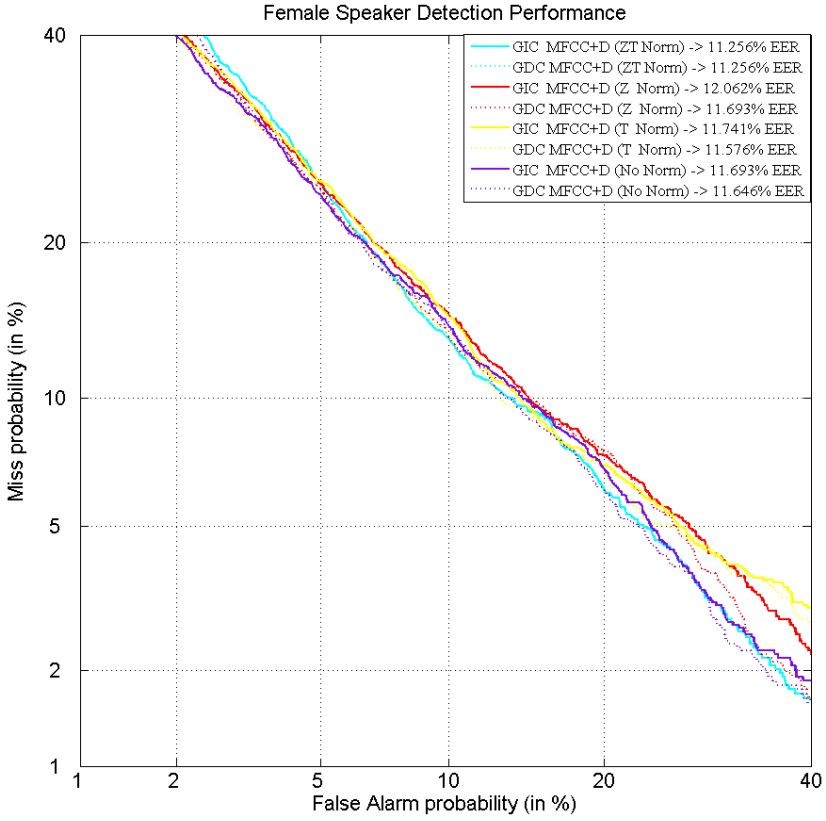



**Figure 5-95** DET curve for classic parameters on MOBIO female development set for GIC and GDC

In order to improve the recognition rates obtained using a gender-dependent setup, we introduced the extra parameters already tested in previous scenarios, namely Energy, ΔEnergy, Pitch (F0) and third formant estimate (F3). We run a set of tests in which each of these extra parameters were included in the feature vector defined by the GDC either alone or combined with the others. For each test all the score normalization techniques were applied, i.e. ZTNorm, ZNorm, TNorm and No Norm (which means that no score normalization technique was applied). From Figure 5-96 (male speakers) and Figure 5-97 (female speakers) it is clear that not all the combinations of the extra parameters provide an improvement in terms of $EER_X$. In these sets of graphs, the horizontal line represents the *EER* obtained for the gender dependent configuration for the selected score normalization; while each column represents a specific combination of extra parameters added to the GDC.

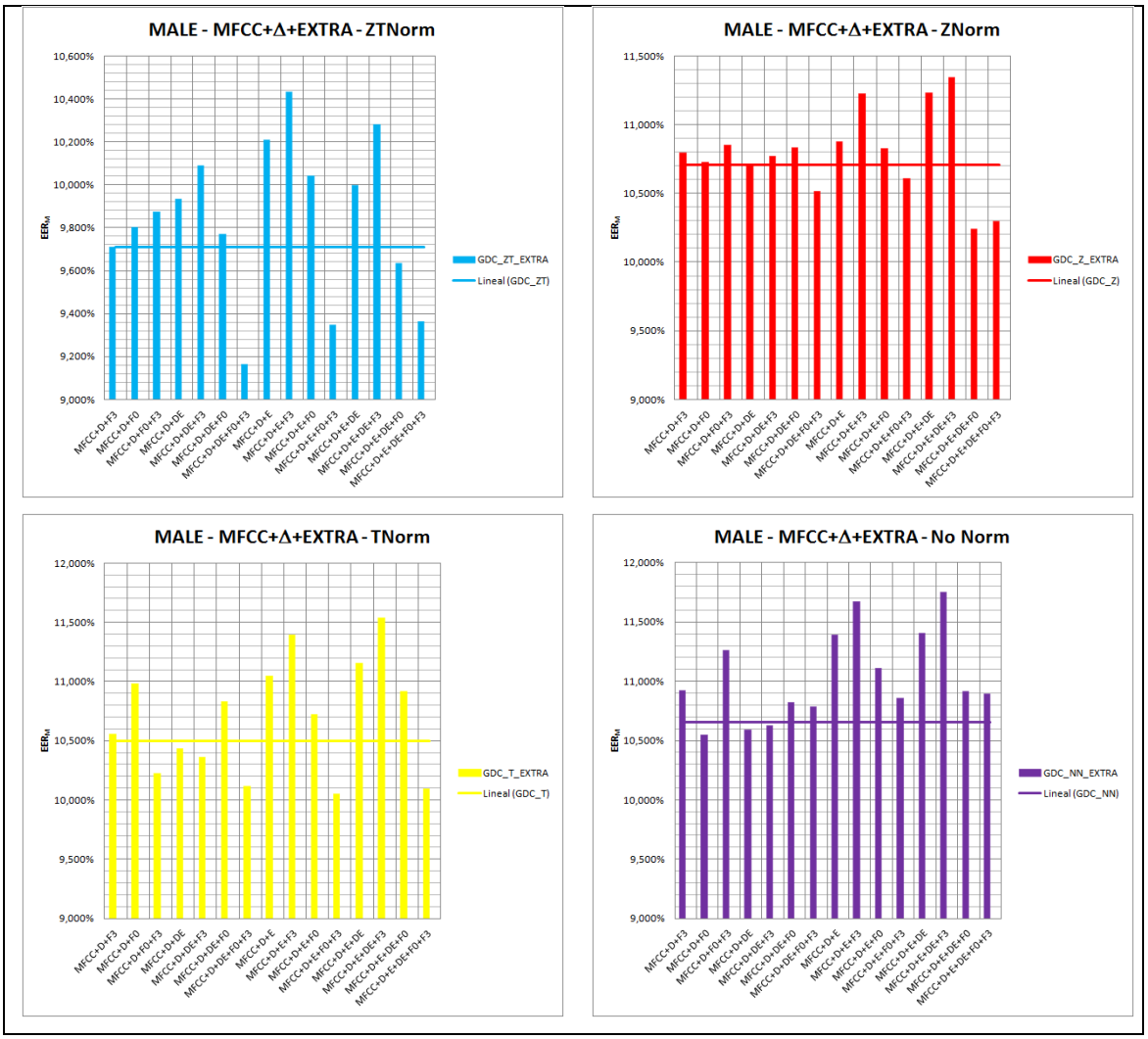



**Figure 5-96** $EER_M$ obtained for the tests which incorporate E, ΔE, F0 and F3 in the feature vectors

In the case of female speakers, only two specific configurations outperformed the GDC MFCCs+Δ, namely, GDC MFCCs+Δ+F0 for the case of applying ZNorm, and GDC MFCCs+Δ+F0+F3 for the case of not applying any score normalization technique. Moreover, this last configuration is the one providing the best results obtained so far on the development set for female speakers. In the case of male speakers, there are multiple configurations that using some combination of the proposed extra parameters provide a

clear improvement in terms of recognition rates, respect to the GDC MFCCs+Δ. Specifically, the inclusion of ΔE+F0+F3, E+F0+F3, or E+ΔE+F0+F3 provide an improvement no matter the score normalization technique applied.

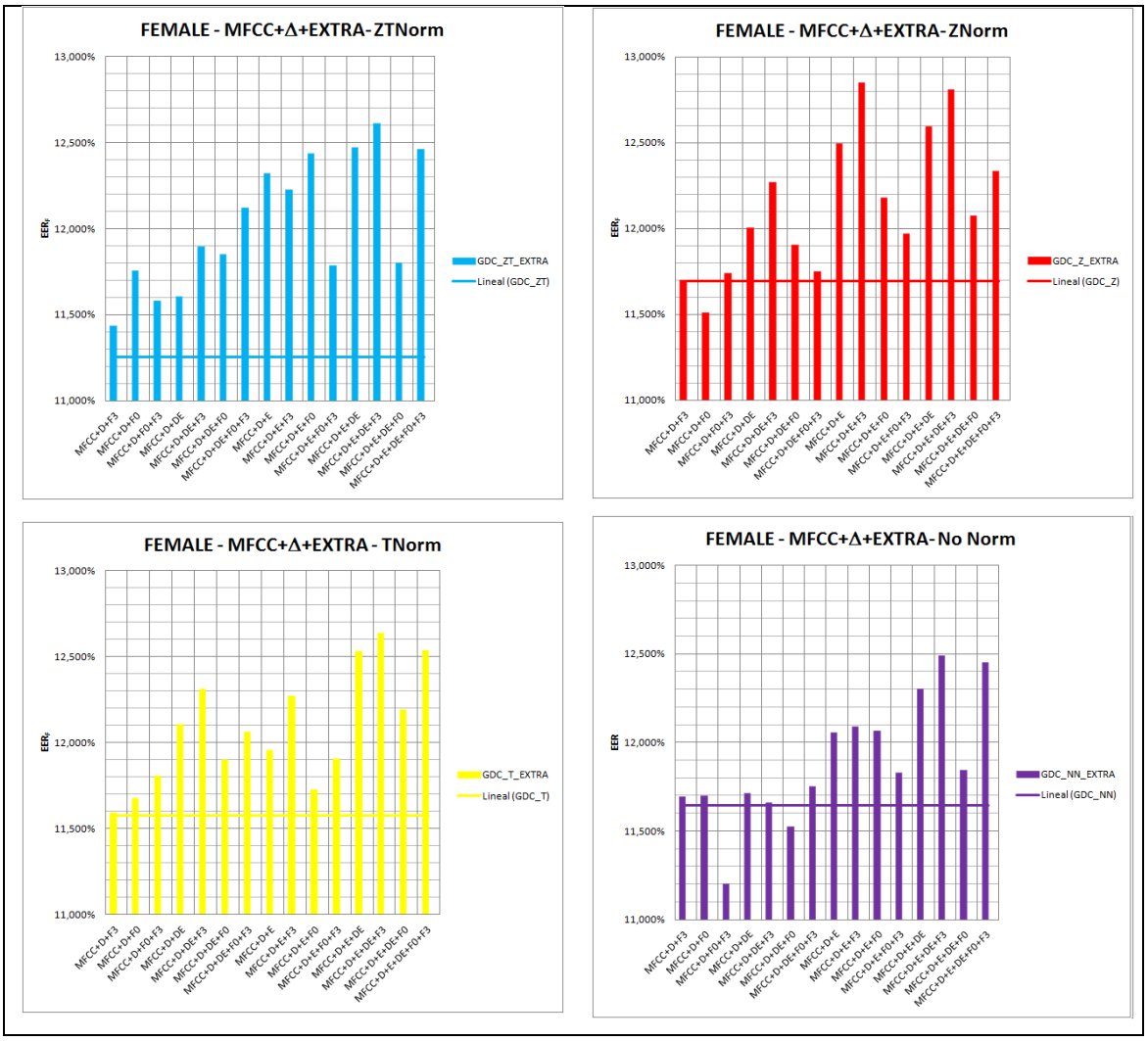



**Figure 5-97** $EER_F$ obtained for the tests which incorporate E, ΔE, F0 and F3 in the feature vectors

Table 5-29 reflects, the most relevant results for each score normalization technique, taking into account that for each setup, different configurations have been tested regarding the number of Gaussians and relevance factor values, keeping only the one providing the best result. Similarly, regarding the use of extra parameters, only the configuration providing the most successful result has been included in Table 5-29. It must be noted that in the case of female speakers and when ZTNorm or TNorm is applied, no improvement is obtained in terms of $EER_F$, when the extra parameters are used.

| Score Norm | Parameters | Gen. | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR | $EER_F$ [$\theta_F$] | $EER_F$ RR |
|---|---|---|---|---|---|---|---|
| ZTNorm | | | | | | | |
| | **GIC MFCC+Δ** | M/F | - | 10.594% [1.556] | - | 11.256% [1.545] | - |
| | **GDC MFCC+Δ** | M | - | 9.708% [1.406] | 8.36% | 11.256% [1.545] | 0.00% |
| | | F | - | | | | |
| | **GDC MFCC+Δ** | M | ΔE+F0+F3 | 9.165% [1.597] | 13.49% | - | - |
| | | F | - | | | | |
| ZNorm | | | | | | | |
| | **GIC MFCC+Δ** | M/F | - | 11.256% [1.663] | - | 12.062% [2.145] | - |
| | **GDC MFCC+Δ** | M | - | 10.710% [1.704] | 4.85% | 11.693% [2.168] | 3.06% |
| | | F | - | | | | |
| | **GDC MFCC+Δ** | M | E+ΔE+F0 | 10.238% [1.739] | 9.04% | 11.508% [2.164] | 4.59% |
| | | F | F0 | | | | |
| TNorm | | | | | | | |
| | **GIC MFCC+Δ** | M/F | - | 10.499% [1.014] | - | 11.741% [0.640] | - |
| | **GDC MFCC+Δ** | M | - | 10.499% [1.014] | 0.00% | 11.576% [0.674] | 1.40% |
| | | F | - | | | | |
| | **GDC MFCC+Δ** | M | E+F0+F3 | 10.048% [1.043] | 4.30% | - | - |
| | | F | - | | | | |
| No Norm | | | | | | | |
| | **GIC MFCC+Δ** | M/F | - | 11.567% [-0.007] | - | 11.693% [0.004] | - |
| | **GDC MFCC+Δ** | M | - | 10.654% [0.014] | 7.89% | 11.646% [0.011] | 0.40% |
| | | F | - | | | | |
| | **GDC MFCC+Δ** | M | F0 | 10.547% [0.013] | 8.82% | 11.201% [0.016] | 4.21% |
| | | F | F0+F3 | | | | |

**Table 5-29** *EER* obtained for the tests which incorporate E, ΔE, F0 and F3 in the feature vectors (best results highlighted in green)

As reflected in Table 5-29, it is possible for both male and female speaker to find a setup, using these extra parameters, that improves the recognition rates obtained by the GDC MFCCs+Δ configuration. Particularly, for male speakers, the use of ΔE, F0 and F3, generates a relative reduction of 13.5% in terms of $EER_M$ respect to GIC, therefore a relative reduction of 5.6% respect to GDC MFCCs+Δ, in the case of applying ZTNorm. For female speakers, the inclusion of F0 and F3 extra parameters provides a relative reduction of 4.21% respect to GIC, and 3.82% respect to GDC MFCCs+Δ, when no score normalization is applied.

Anyway, from the results shown in Table 5-29, we can draw the following conclusions that confirm previous results on the presented scenarios. Specifically, the use of E+ΔE is still not the best option despite being used in most of speaker recognition systems. Besides, we find again that F3 parameter, whose use is proposed in this thesis, keeps on being an interesting option in order to improve recognition rates, for both male and female speakers.

The next step, like in the previous scenario, consists in introducing what we have called extended-biometric parameters extracted by the GDEB front-end. The approach that has been followed, consists in incorporating, the set of parameters extracted from the glottal source estimate (labelled as GSE) into the most successful setup, i.e. GDC MFCCs+Δ+ΔE+F0+F3 in the case of male speakers and GDC MFCCs+Δ+F0+F3 in the case of female speakers. Once a specific configuration improving previous results is found, we continue by incorporating parameters extracted from the vocal tract estimate (VTE). As previously noted, from this point on we are going to run the test using the score normalization that provides most successful results for each gender. Therefore, in the case of male speakers, we are going to apply ZTNorm, and in the case of female speakers, we are not going to apply any score normalization as the data available for normalization purposes seems to be not enough to improve recognition rates.

Although multiple configurations have been tested, Table 5-30 (male) and Table 5-31 (female) show the final configurations chosen for each gender, as well as the recognition rates obtained in each case in terms of *EER* (best highlighted in green). Additionally, the relative reduction in terms of *EER*, if compared to the GIC is also presented, provided that the comparison with GIC instead of GDC is based on the fact that GIC is considered the state-of-the-art to beat in the front-end subsystem. In brackets, in the *EER* column, it is also shown the resulting scoring threshold for the evaluation phase.

| Parameters | GSE+VTE set up | Extra Parameters | $EER_M$ [$\theta_M$] | $EER_M$ RR |
|---|---|---|---|---|
| **GIC MFCC+Δ** | - | - | 10.594% [1.556] | - |
| **GDC MFCC+Δ** | - | - | 9.708% [1.406] | 8.36% |
| **GDC MFCC+Δ** | - | ΔE+F0+F3 | 9.165% [1.597] | 13.49% |
| **GSE** | **Source-Tract Sep. Alg:** Prediction Order: 24 Forgetting Factor: 0.995 **GSE:** 7-Channel Filter bank /6MFCC | ΔE+F0+F3 | 8.332% [1.619] | 21.35% |
| **GSE+VTE** | **Source-Tract Sep. Alg:** Prediction Order: 24 Forgetting Factor: 0.995 **GSE:** 14-Channel Filter bank /8MFCC **VTE:** 14-Channel Filter bank /2MFCC | ΔE+F0+F3 | 8.496% [1.506] | 19.80% |

**Table 5-30** $EER_M$ obtained on development set (ZTNorm), comparing classical parameters with extra parameters and extended biometric parameters

| Parameters | GSE+VTE set up | Extra Parameters | $EER_F$ [$\theta_F$] | $EER_F$ RR |
|---|---|---|---|---|
| **GIC MFCC+Δ** | - | - | 11.693% [0.004] | - |
| **GDC MFCC+Δ** | - | - | 11.646% [0.011] | 0.40% |
| **GDC MFCC+Δ** | - | F0+F3 | 11.201% [0.016] | 4.21% |
| **GSE** | **Source-Tract Sep. Alg:** Prediction Order: 36 Forgetting Factor: 0.995 **GSE:** 22-Channel Filter bank/4 MFCC | F0+F3 | 10.643% [0.010] | 8.98% |
| **GSE+VTE** | **Source-Tract Sep. Alg:** Prediction Order: 36 Forgetting Factor: 0.995 **GSE:** 22-Channel Filter bank/4 MFCC **VTE:** 25-Channel Filter bank/2 MFCC | F0+F3 | 11.016% [0.023] | 5.79% |

**Table 5-31** $EER_F$ obtained on development set (No Norm), comparing classical parameters with extra parameters and extended biometric parameters

The DET curves that represent the results obtained with each of the previously presented configurations (see Table 5-30 and Table 5-31) are depicted in Figure 5-98 for male speakers and Figure 5-99 for female speakers. Clearly, the proposed gender-dependent extended biometric parameterisation, in this case incorporating information just from the glottal source estimate in form of MFCCs, is the configuration that provides the most successful results in the development set for both male and female speakers. The different tests carried out including the VTE parameters are worse than the results obtained using GSE parameters, but still improve recognition rates of GIC, as expected. Specifically, for the male speakers, the use of GSE setup, thus a gender-dependent configuration incorporating extended biometric features, provides a relative reduction of 21% in terms of $EER_M$, respect to the gender-independent configuration. Whereas in the case of female speakers, the use of the GSE setup allows for a relative close to 9% in terms of $EER_F$, respect to the gender-independent configuration.

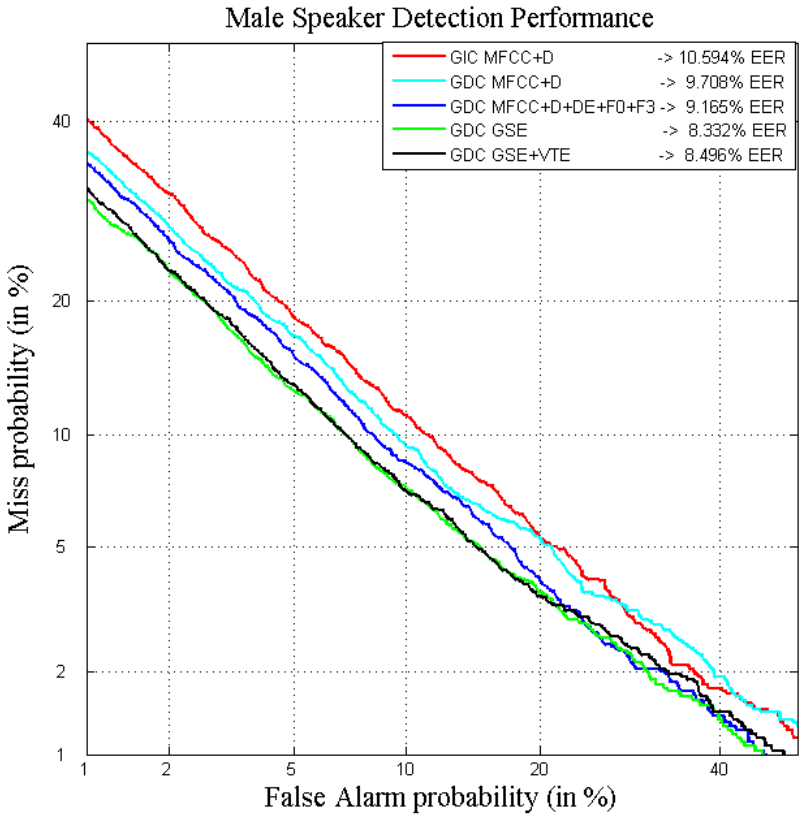



**Figure 5-98** DET curves comparing classical parameters and GDEB on MOBIO's development set for male speakers and ZTNorm

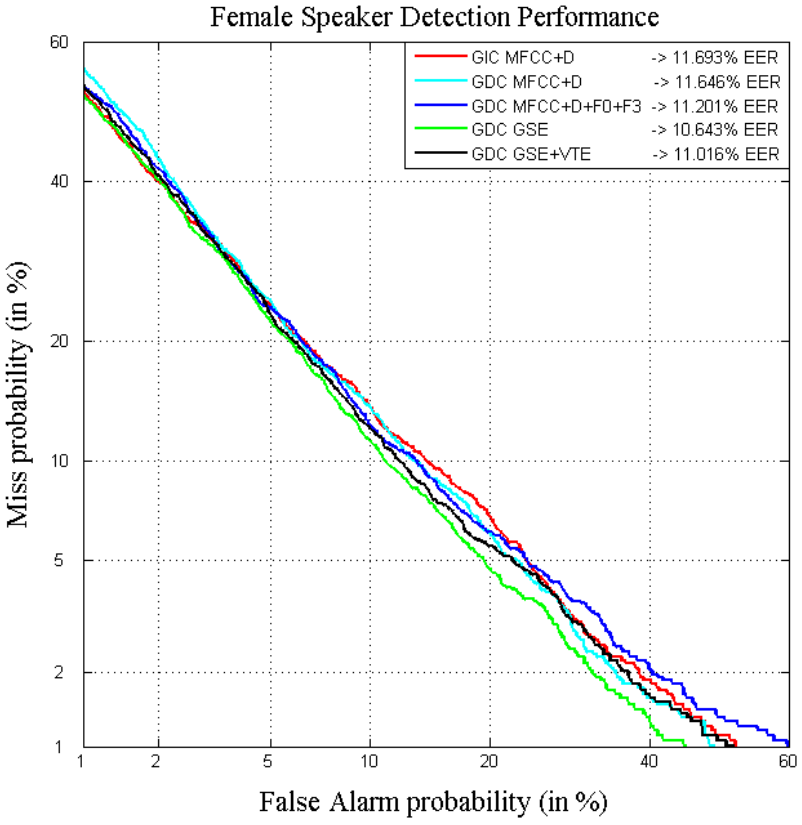



**Figure 5-99** DET curves comparing classical parameters and GDEB on MOBIO's development set for female speakers and No Norm

Like in previous scenarios, we have verified that the improvement derived from incorporating GSE information into the feature vector is systematically obtained and is not the result of an isolated and specific configuration. Figure 5-100 provides the minimum $EER_X$ (y-axis) obtained when GSE is incorporated into the feature vector in form of MFCCs for male speakers (light green). Different numbers of $MFCCs_G$={4,6,8,10} have been tested, which have been computed applying a filter bank with different number of filters $F_G$=[4...18] (x-axis). Each point in the x-axis represents the minimum *EER* obtained for a specific value of $F_G$, regardless $MFCCs_G$ value. Figure 5-101 provides the same information for female speakers, although in this case, $MFCCs_G$={2,4,6} and $F_G$=[2...23]. These values have been selected based on previous experience. Clearly, the use of GSE systematically derives in an improvement of recognition rates regardless the gender of speakers, as deduced from the depicted results.

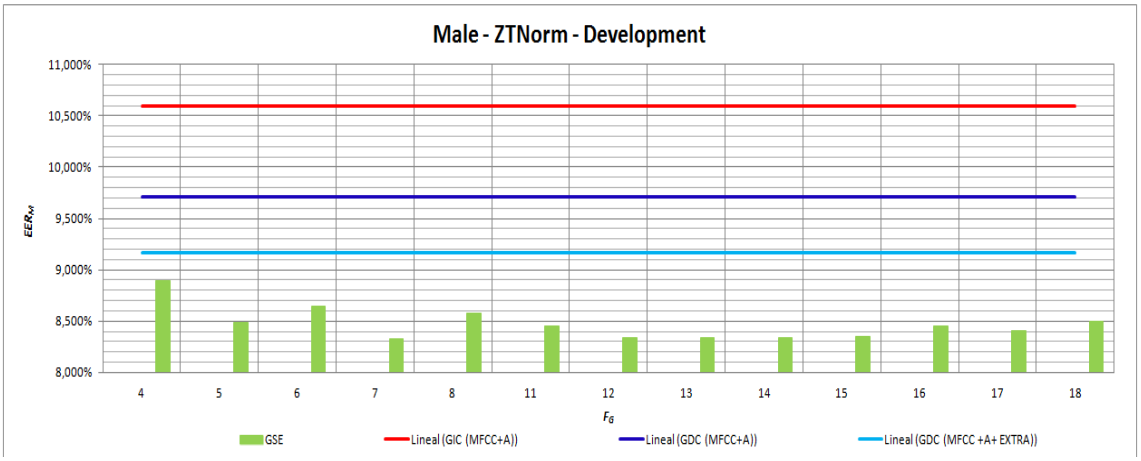



**Figure 5-100** Influence of GSE configuration on the $EER_M$ (development set)

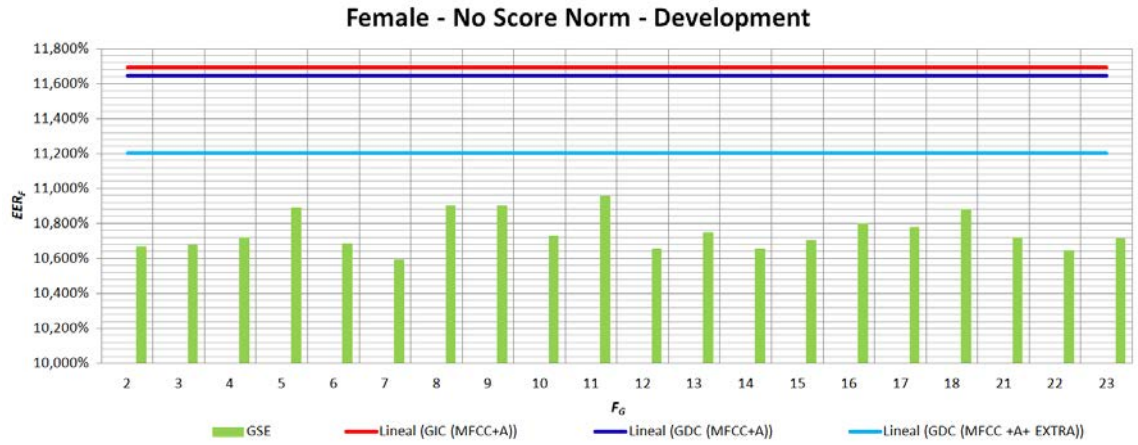



**Figure 5-101** Influence of GSE configuration on the $EER_F$ (development set)

The results obtained in terms of *EER* on the development set should be treated with caution, especially by the fact that, although higher than in HESPERIA and ALBAYZIN databases, the number of speakers in the training set (used for training the UBM and for score normalisation purposes) is quite limited. This can lead to an overtraining in development set, which we have previously seen in HESPERIA, especially if the UBM obtained is not representative of the speaker on the evaluation set, i.e. it is not a universal model.

Once we have established the configurations providing the best results, in terms of *EER*, on the development set, the next step, consists in checking the behaviour of the system on the evaluation set. Figure 5-32 shows the results obtained from the evaluation set using the different configurations previously specify in Table 5-30 and Table 5-31.

| Parameters | $EER_M$ [$\theta_M$] | $HTER_M$ | $HTER_M$ RR | $EER_F$ [$\theta_F$] | $HTER_F$ | $HTER_F$ RR |
|---|---|---|---|---|---|---|
| **GIC MFCC+Δ** | 10.594% [1.556] | 9.23% | - | 11.693% [0.004] | 14.09% | - |
| **GDC MFCC+Δ** | 9.708% [1.406] | 8.79% | 4.73% | 11.646% [0.011] | 15.09% | -7.14% |
| **GDC MFCC+Δ+ExtParm.** | 9.165% [1.597] | 8.87% | 3.84% | 11.201% [0.016] | 14.85% | -5.38% |
| **GSE** | 8.332% [1.619] | 8.38% | 9.19% | 10.643% [0.010] | 13.10% | 6.99% |
| **GSE+VTE** | 8.496% [1.506] | 8.64% | 6.35% | 11.325% [1.441] | 13.15% | 6.62% |

**Table 5-32** $HTER_X$ obtained for the selected configurations on evaluation set, applying ZTNorm – male – and No Norm – female.

The results obtained clearly show that the fact of using a gender-dependent parameterisation which includes information from the glottal source and the vocal tract estimates, provides an improvement in recognition rates with respect to the gender-independent characterisation of speakers, both on development an evaluation set. Although most successful results are obtained when only glottal source information is used. This improvement is especially significant in the case of male speakers, for whom we obtained a relative reduction of 9% in terms of *HTER* from an *HTER*=9.23% when the classical gender-independent characterisation is used down to *HTER*=8.38% when GSE setup is used. In the case of female speakers, this reduction is lower, close to 7%. Another aspect that must be noticed is the fact that the result obtained for GDC MFCCs+Δ and GDC MFCCs+Δ +ExtParam., do not provide an improvement in the evaluation in terms of *HTER* in the case of female speakers. This may be due, as previously cited, to an overtraining on development set.

The DET curves that represent the results obtained in evaluation set, with each of the previously presented configurations are depicted in Figure 5-102 for male speakers and Figure 5-103 for female speakers.

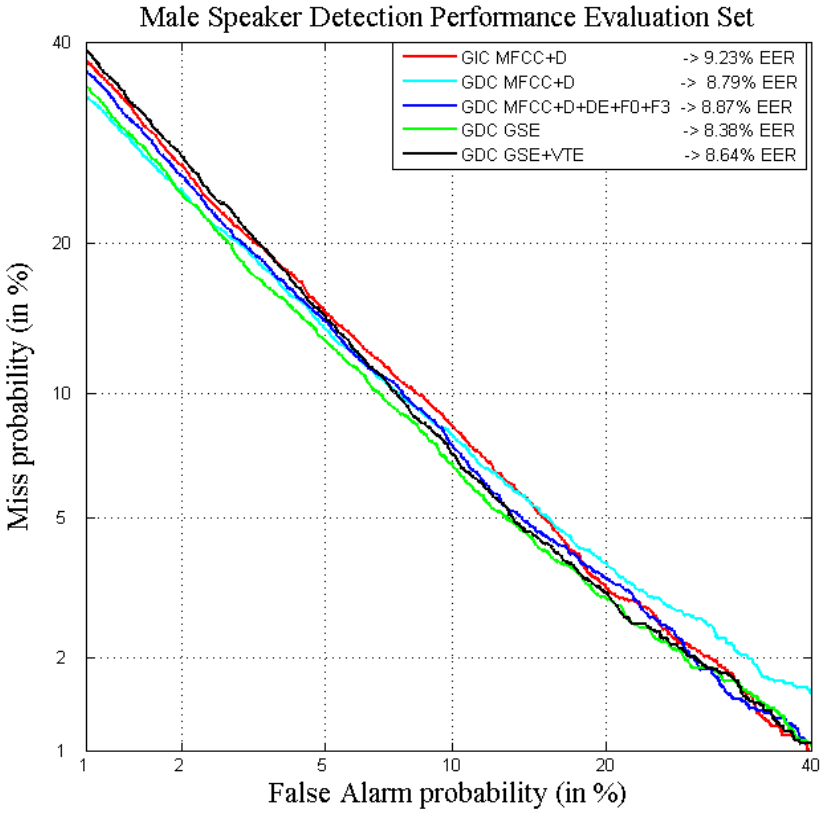



**Figure 5-102** Male DET curves on MOBIO evaluation set, applying ZTNorm

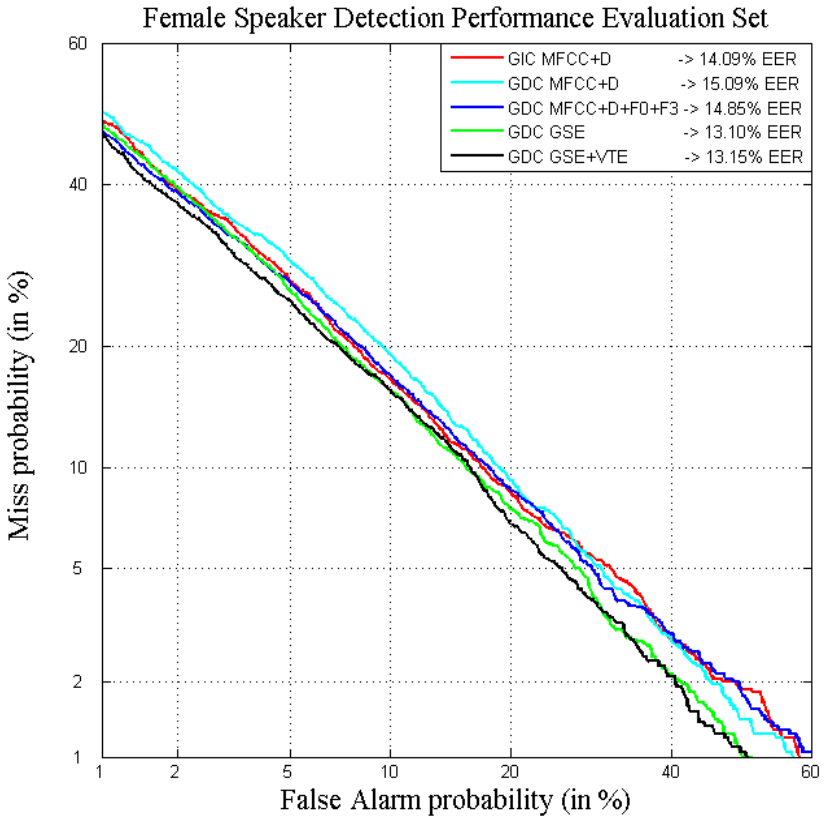



**Figure 5-103** Female DET curves on MOBIO evaluation set, without applying score normalization.

Finally, as the presented tests have been designed in the context of an international evaluation contest, where different speaker recognition systems are tested under the same conditions, it seems appropriate and necessary to compare the results obtained by our system with the results of other sites participating in the SRE. Table 5-33 summarises the results obtained in the SRE by different systems in both development and evaluation sets [Khoury,2013]. It must be noted that although it is supposed that all systems work under the same conditions (defined in the SRE plan) this is not actually so. Specifically, systems marked with + are those using external/additional training data, allowing them to create, probably, more appropriate UBMs and have additional information for normalisation purposes. Moreover, systems marked with *, are actually fusions of multiple systems, thus requiring the application of adequate fusion techniques thereby increasing the complexity and the cost in terms of running time needed to generate recognition results, if compared with the simple but effective system proposed here.

| | FEMALE | | MALE | |
|---|---|---|---|---|
| System | DEV-*EER* | EVAL-*HTER* | DEV-*EER* | EVAL-*HTER* |
| Alpineon* | 7.982% | 10.678% | 5.040% | 7.076% |
| ATVS+ | 16.836% | 17.858% | 14.881% | 15.429% |
| CPqD* | 14.348% | 15.987% | 11.824% | 10.214% |
| CDTA | 19.471% | 22.640% | 12.738% | 19.404% |
| GIAPSI[1] | 10.643% | 13.107% | 8.332% | 8.382% |
| **GIAPSI[2]** | **11.320%** | **12.590%** | **8.496%** | **8.631%** |
| EHU | 17.937% | 19.511% | 11.310% | 10.058% |
| IDIAP | 12.011% | 14.269% | 9.960% | 10.032% |
| L2F* | 13.484% | 22.140% | 10.599% | 11.129% |
| L2F-EHU* | 11.005% | 17.266% | 7.889% | 8.191% |
| Mines-Telecom+ | 11.429% | 11.633% | 10.198% | 9.109% |
| Phonexia+ | 8.364% | 14.181% | 9.601% | 10.779% |
| RUN+ | 25.405% | 23.112% | 24.643% | 22.524% |
| Fusion LLR | 3.556% | 6.986% | 2.897% | 4.767% |

**Table 5-33** EER % on the development (DEV) set and half total error rate (*HTER* %) on the evaluation (EVAL) set for the systems participating in 2013 SRE in Mobile Environments.

Regarding the front-end used by the participating sites, it must be pointed out that all systems but the Alpineon system which uses 3 different cepstral-based features (MFCCs, LFCCs and PLPs), the all remainder systems use the same features, i.e. MFCCs in a gender-independent configuration. The only exception was our system (GIAPSI) which as presented over this section uses the GDEB front-end. The last row of Table 5-33 shows the results obtained in terms of *EER* and *HTER* of the fusion of all primary systems using linear logistic regression.

As reflected in Table 5-33, after a post-processing subsequent to the delivery of results, we got a better parameter adjustment on the development set, which resulted also in a slight improvement on the evaluation set, in the case of male speakers.

[1] Original results.

[2] Results achieved after post-processing.

#### *5.2.3.1 Brief Conclussions*

In the text-independent scenario on mobile environments designed using the MOBIO database, the use of a gender-dependent extended biometric parameterisation in which VTE and GSE information have been incorporated provides a clear improvement in terms of recognition rates respect to the use of the classic gender-independent approach. This improvement remains consistent over the development set and the evaluation set. However, the most successful results are achieved when just GSE parameters are included in the feature vectors.

In order to improve recognition rates it is essential to use a gender-dependent parameterization. This result in the need to use different configurations in terms of the number of MFCCs as well as in the number of channels in the filter bank used to compute the MFCCs, for both male and female speakers. However, in this specific scenario, the number of optimal filters used in the filter bank to compute MFCCs is higher in the case of male speakers than in the case of female speakers. Likewise, the optimal number of MFCCs is higher for male speakers than for female speakers, though in both cases is located in the range of 24 to 28.

In this scenario we have not test the effect of $\Delta\Delta$ coefficients as in previous scenarios their use has not offered additional benefits in terms of EER.

Regarding the use of extra parameters, it has been confirmed that some specific combination of them helps to increase recognition rates. These combinations usually include the F3 coefficient proposed in this thesis as a new parameter for speaker recognition.

Concerning the evaluation it must be noted that, despite having developed a simple recognition system (based on the UBM-GMM paradigm), the fact of having obtained a better speaker characterisation based on gender-dependent extended biometric parameters, allows us to get very competitive results. Moreover, the only systems that improve our recognition rates are those that either performed a fusion of multiple systems or used additional data for training. Besides, according to the published results our system gets the best simple system performance on male speakers. However, results seem to be still far away for the best results that can be obtained fusing all the presented systems (male evaluation set HEER=6.986%, female evaluation set HEER=4.767%), providing still some room for improvement by using the presented gender-dependent extended biometric front-end but incorporating additional external data for training and normalisation purposes, aspect which is essential to also apply more complex classifiers such as GSV or *i*-vectors.

### 5.2.4 NIST SRE Evaluations

Due to the special characteristics of this experiment, especially regarding the number of different speakers, the number of available recordings per speakers, and the number of trials to be tested, we have followed a slightly different approach.

As previously presented, we have to deal with three different sets: the background training set used to learn the background parameters of the algorithm (UBM, subspaces, etc.) or for normalisation purposes, the development set supposed to be used to tune meta-parameters of the algorithm, and the evaluation set used to analyse the performance of recognition systems. In order to evaluate the performance of the systems like in previous scenarios, *EER* is used as a quality measure in the development and evaluation set respectively, as no decision needs to be made on the evaluation set.

Therefore, DET curves, where the *FRR* is plotted against the *FAR* are going to play a key role to evaluate the calibration of the verification system.

Similarly to the previously presented scenarios, no cross-gender trials are going to be performed, therefore, the *HEER* metric can be used again. Obviously, when using a gender-independent parameterisation, it is necessary to reach a compromise between both *EER* in order to minimise *HEER*. However, when using a gender-dependent parameterisation, this compromise disappears and the objective is to minimise $EER_M$ and $EER_F$ independently.

$$HEER = \frac{EER_M(conf\_male) + EER_F(conf\_female)}{2} \qquad \textbf{Eq. (5-10)}$$

It must be noted that the configurable parameters present in Eq. (5-9), have been deliberately removed in Eq. (5-10), and substituted by a generic configuration. This change is based on the impossibility of carrying out a deep search for the best configuration as it was done in previous scenarios, given the amount of data to be handled in this specific case. Therefore, we are going to proceed by selecting some specific configurations that are the state-of-the-art both in speaker recognition and NIST SREs. Additionally, the availability of a huge amount of data with a high variability in terms of speakers and channels of transmission, allows us to test different classification methods, namely GMM-UBM, SV-GMM, and *i*-vector approaches, but on a limited number of configurations.

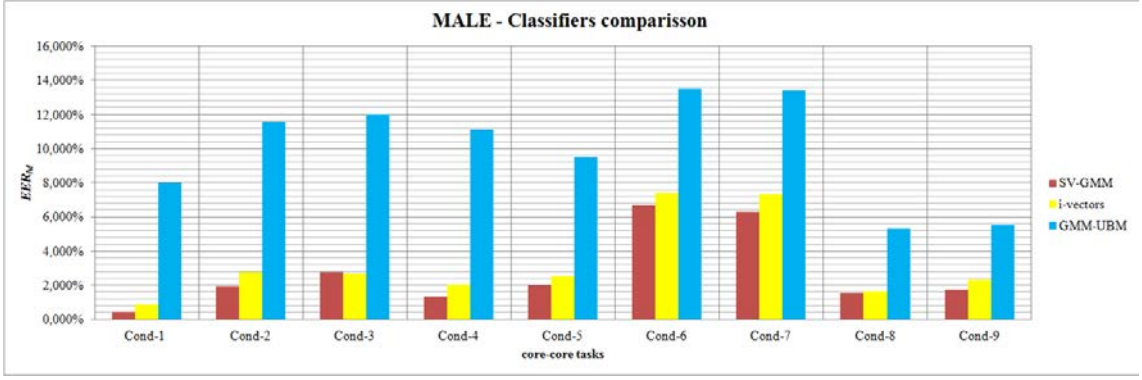
Regarding score normalisation techniques, we have focused our efforts in two specific score normalisation techniques depending on the classification method used. Specifically, we have combined the use of the GMM-UBM approach with the ZTNorm. In previous scenarios, we have verified that its use in text-independent environments, in which enough amount of normalisation data is available, offers greater advantages respect to the other score normalisation algorithms. In the case of the *i*-vector approach, we have combined it with Symmetric Normalisation (SNorm – see section 1.5.3), as it usually provides the most successful results for this kind of classification method. In the case of using the SV-GMM classification method, we have tested both the SNorm approach and the SNorm combined with ZTNorm.

The first set of tests carried out in this scenario was directed to determine which of the classification methods offers better performance. To this end, we proposed a test using the *Baseline* front-end, thus applying a gender-independent configuration, with a typical configuration (as reported in multiple research papers) regarding number of MFCCs, and number of filters in the filter bank. In this case, the selected configuration was: *MFCC*={19}, *F*={40}, Δ={true} and ΔΔ={false}. It must be noted that ΔΔ has been deliberately set to false, even though most speaker recognition systems use this parameters. We have already seen, and it will be also shown later, that these parameters are not especially helpful for speaker recognition purposes in a text-independent scenario.

Regarding the configuration of the three tested classifiers, we have used typical parameters reported in the state-of-the-art. Specifically, for the GMM-UBM approach, 1024 Gaussians have been used to build both the UBM and the speaker's models, whereas as relevance factor different values have been tested, $\alpha$={8,12,16,20}, selecting the one providing better recognition rates. In the case of the SV-GMM, the same configuration described for the GMM-UBM approach has been used. Additionally, although applying a gender-independent approach, gender dependent total subspaces, of

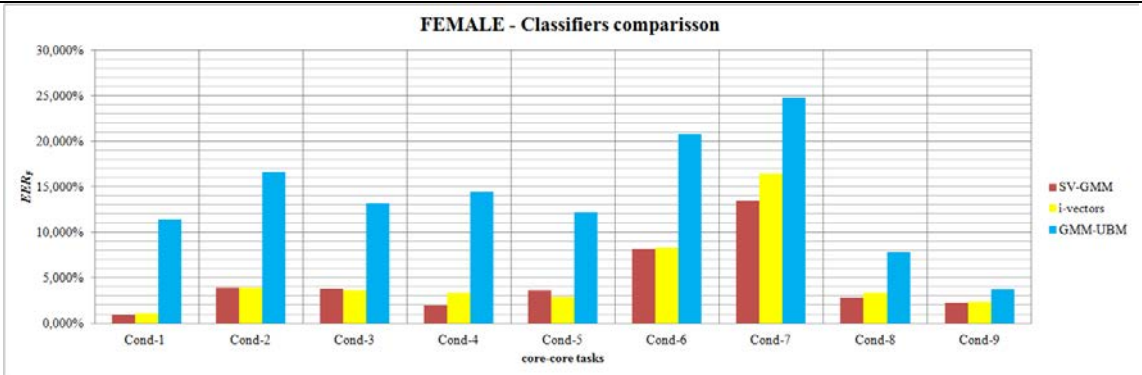
400 dimension, were generated after applying LDA to an 800 (rank of PCA matrix) dimension space calculated via classical eigenanalysis (PCA) from background data. Three different *total spaces* were considered, namely Tel (telephone only), Mic (phone-mic and interview-mic), and Tel-Mic, where phone-mic and interview-mic were included besides telephone data. WCCN has also been applied for inter-session variability compensation. Finally, for the *i*-vector approach, the same configuration described for the SV-GMM approach is valid, taken into account that the classical eigenanalysis is replaced by the total variability analysis (T-matrix).

The following plots show the performance in terms of $EER_X$, for each of the classifiers under the different conditions previously defined. The most successful results for each gender and each condition are highlighted in green.




| | cond-1 | cond-2 | cond-3 | cond-4 | cond-5 | cond-6 | cond-7 | cond-8 | cond-9 |
|---|---|---|---|---|---|---|---|---|---|
| GMM-UBM | 7.988% | 11.580% | 12.028% | 11.138% | 9.473% | 13.484% | 13.409% | 5.305% | 5.508% |
| SV-GMM | 0.432% | 1.924% | 2.745% | 1.323% | 1.987% | 6.685% | 6.303% | 1.536% | 1.757% |
| *i*-vectors | 0.855% | 2.771% | 2.680% | 2.049% | 2.508% | 7.387% | 7.321% | 1.636% | 2.324% |

**Figure 5-104** $EER_M$ obtained depending on the classifier - NIST SRE 2010, Cond-1 to Cond-9




| | cond-1 | cond-2 | cond-3 | cond-4 | cond-5 | cond-6 | cond-7 | cond-8 | cond-9 |
|---|---|---|---|---|---|---|---|---|---|
| GMM-UBM | 11.435% | 16.594% | 13.132% | 14.434% | 12.205% | 20.765% | 24.770% | 7.820% | 3.733% |
| SV-GMM | 0.944% | 3.928% | 3.826% | 1.967% | 3.670% | 8.183% | 13.408% | 2.780% | 2.290% |
| *i*-vectors | 1.108% | 3.911% | 3.596% | 3.387% | 2.894% | 8.382% | 16.472% | 3.349% | 2.333% |

**Figure 5-105** $EER_F$ obtained depending on the classifier - NIST SRE 2010, Cond-1 to Cond-9

The results obtained by the GMM-UBM classifier are clearly worse that the other two evaluated classifiers as depicted in Figure 5-104 and Figure 5-105.

In order to carry out a straight comparison of the two remaining classification methods, we have defined and additional metric, Average Condition Error Rate, *ACER*:

$$ACER = \frac{\sum_{cond} EER_{cond}}{\#cond} \qquad \textbf{Eq. (5-11)}$$

Which may be defined as the average of the *EER* obtained for the different conditions reflected in which the development set is divided. Obviously this is not the most accurate metric that can be defined to measure the actual behaviour of the system, as the number of trials on each condition is neglected. However, minimizing *EER*, which is the primary objective, will contribute to minimise also *ACER*. So we can assume that the system achieving lower *ACER* will show better overall performance. Table 5-34 gives the *ACER* values obtained based on the values reflected in the tables contained in **¡Error! No se encuentra el origen de la referencia.** to Figure 5-105.

| Genre | *MALE* | *FEMALE* |
|---|---|---|
| Classifier | *ACER* *[RR]* | *ACER* *[RR]* |
| GMM-UBM | 9.990% [-] | 13.876% [-] |
| SV-GMM | 2.744% [72.54%%] | 4.555% [67.17%] |
| *i*-vectors | 3.281% [67.16] | 5.048% [63.62%] |

**Table 5-34** Results obtained in terms of *ACER* by the different classification methods for each gender

There are some conclusions that can be drawn both from the results obtained and reflected in Figure 5-104, Figure 5-105 and, Table 5-34, and from the process itself. Regarding the results, it must be noted that for most of the conditions in which the development set has been divided, no matter whether we are facing male or female trials, the SV-GMM approach usually shows better performance, providing lower values for *EER*. Additionally as reflected in Table 5-34, SV-GMM is the classifier providing lower *ACER* among the three classifiers analysed. Specifically, in the case of male speakers, SV-GMM provides a relative reduction in terms of *ACER* up to 72.5% with respect to the GMM-UBM approach and more than 16% with respect to the *i*-vector approach; while in the case of female speakers, this relative reduction is closed to 67% with respect to the GMM-UBM system and close to 10% with respect to the *i*-vector based system.

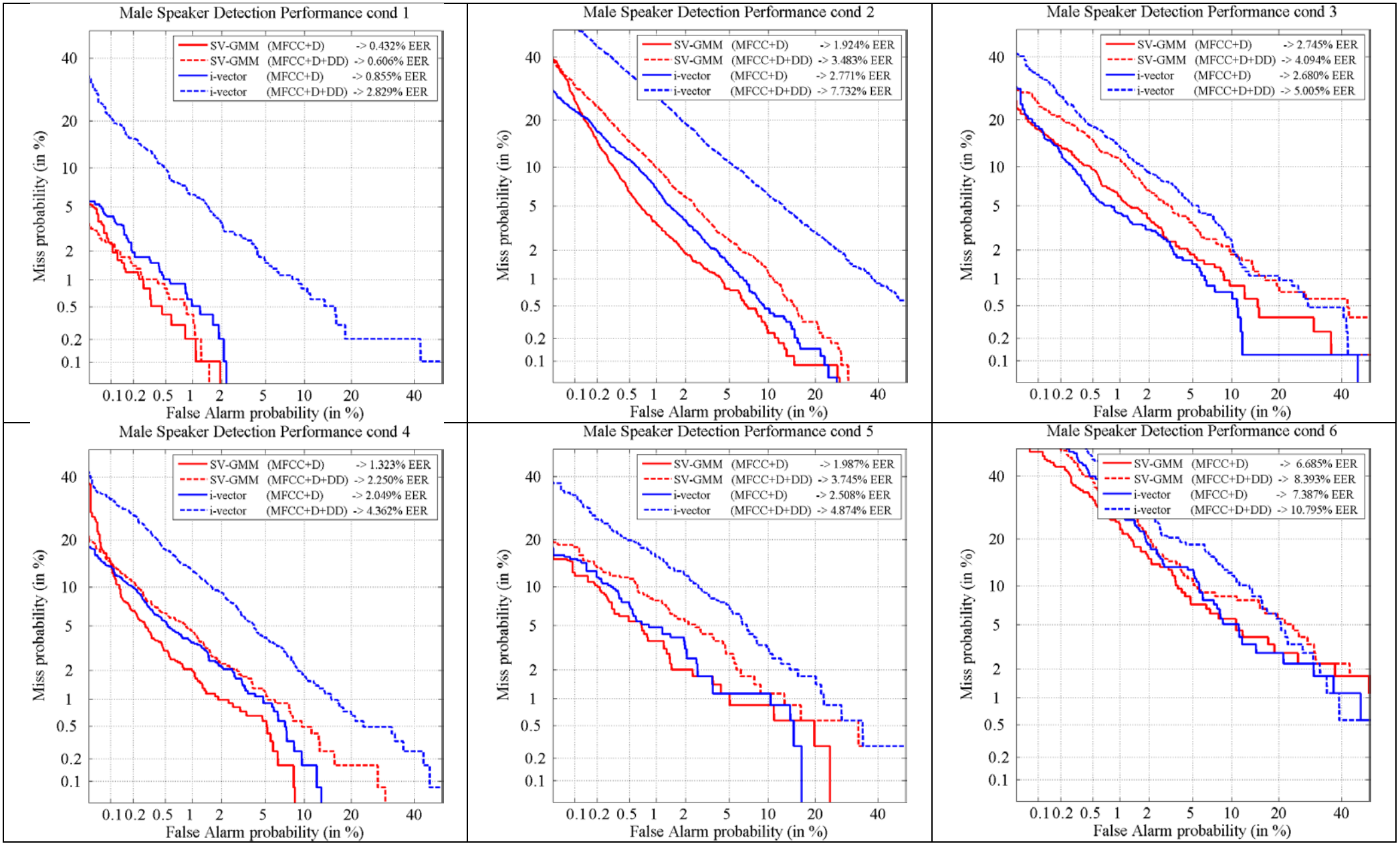
Regarding the process itself, without going into a detailed analysis, we have to discuss processing times affecting the performance of the system. In the case of applying the GMM-UBM approach, we find a bottleneck in the scoring stage. Since dimensionality reduction is applied neither to the speaker`s models nor to the test files, time devoted to scoring and normalization may be unaffordable and extremely high if compared with the other two approaches. Specifically, if the GMM-UBM approach is used in female condition 2, which presents the higher number of trials to be processed, the time spent on evaluating all the trials can reach 420 hours; while the same process using the same normalization data in the SV-GMM approach is completed in less than 20 minutes. In the case of the *i*-vector approach, the bottleneck may be found in the computation of the

total variability matrix, T. For instance, assuming the use of the same training data in the same computer, while the computation of the 800-rank PCA-matrix for the female Tel-Mic space may take no more than 20 minutes, the computation of the equivalent T matrix may last more than 72 hours (4320 minutes). Obviously, the computation of these matrices is performed once and off-line after the set of features is fixed. So regardless this point, both *i*-vector and SV-GMM approaches will show similar processing times.

Anyway, in this case not only the SV-GMM approach provides the most successful results, but also it is the fastest method providing scores, therefore making it more suitable for rapid prototyping. For this reason, in what follows, all the presented results are those obtained following the SV-GMM approach, exception made of the analysis on the use of ΔΔ coefficients, which includes both the SV-GMM and the *i*-vector approaches.

As discussed above, in the previous tests we ruled out the use of the ΔΔ coefficients, since in the set of experiments run on HESPERIA, ALBAYZIN and MOBIO datasets their use did not provide any improvement in terms of recognition rates. However, since in the case of NIST-based experiments we face a quite different situation both in quality and quantity of information, and the use of ΔΔ coefficients is spread among the evaluation participants, it seems necessary to evaluate their usefulness.

To this end, we proposed a test using the *Baseline* front-end, thus applying a gender-independent configuration, with a typical configuration regarding the number of MFCCs, and the number of filters in the filter bank. In this case, the selected configuration was: *MFCC*={19}, *F*={40}, Δ={true} and ΔΔ={true}; which will be used as the input to the SV-GMM and the *i*-vector classifiers previously described. The results obtained using this set of features is going to be compared with the ones obtained when no ΔΔ coefficients are used.


Male Speaker Detection Performance cond 1
SV-GMM (MFCC+D) -> 0.432% EER
SV-GMM (MFCC+D+DD) -> 0.606% EER
i-vector (MFCC+D) -> 0.855% EER
i-vector (MFCC+D+DD) -> 2.829% EER
Male Speaker Detection Performance cond 2
SV-GMM (MFCC+D) -> 1.924% EER
SV-GMM (MFCC+D+DD) -> 3.483% EER
i-vector (MFCC+D) -> 2.771% EER
i-vector (MFCC+D+DD) -> 7.732% EER
Male Speaker Detection Performance cond 3
SV-GMM (MFCC+D) -> 2.745% EER
SV-GMM (MFCC+D+DD) -> 4.094% EER
i-vector (MFCC+D) -> 2.680% EER
i-vector (MFCC+D+DD) -> 5.005% EER
Male Speaker Detection Performance cond 4
SV-GMM (MFCC+D) -> 1.323% EER
SV-GMM (MFCC+D+DD) -> 2.250% EER
i-vector (MFCC+D) -> 2.049% EER
i-vector (MFCC+D+DD) -> 4.362% EER
Male Speaker Detection Performance cond 5
SV-GMM (MFCC+D) -> 1.987% EER
SV-GMM (MFCC+D+DD) -> 3.745% EER
i-vector (MFCC+D) -> 2.508% EER
i-vector (MFCC+D+DD) -> 4.874% EER
Male Speaker Detection Performance cond 6
SV-GMM (MFCC+D) -> 6.685% EER
SV-GMM (MFCC+D+DD) -> 8.393% EER
i-vector (MFCC+D) -> 7.387% EER
i-vector (MFCC+D+DD) -> 10.795% EER
Miss probability (in %)
False Alarm probability (in %)
0.1 0.2 0.5 1 2 5 10 20 40

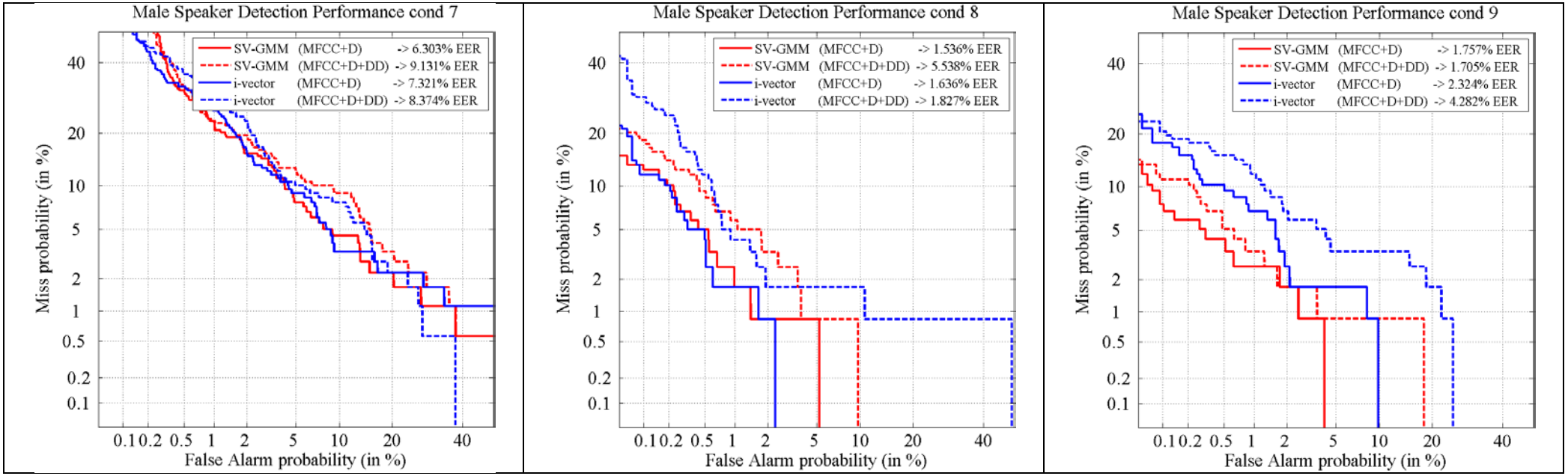
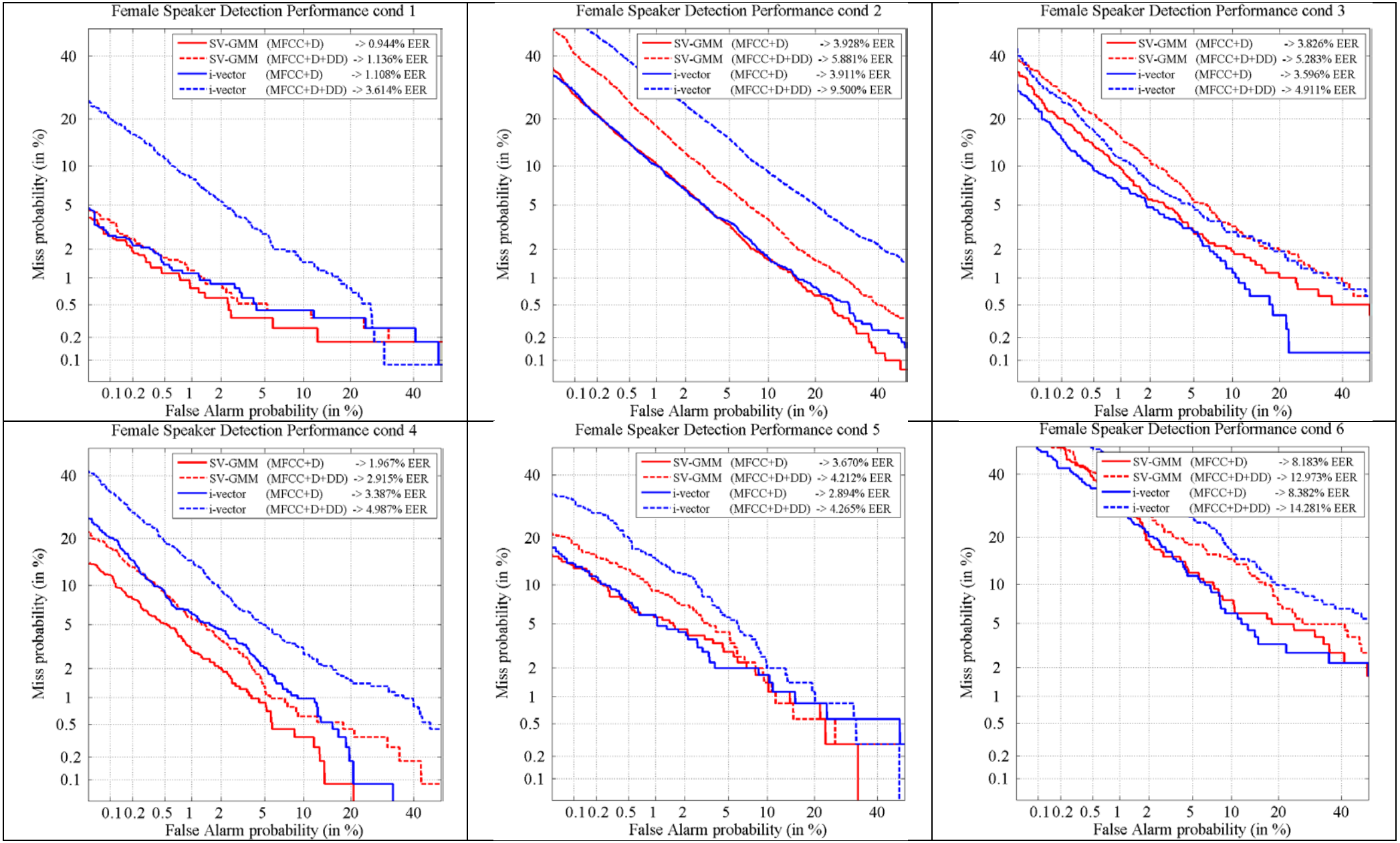



**Figure 5-106** DET curves for male speakers under different conditions comparing the use ΔΔ coefficients


Female Speaker Detection Performance cond 1
SV-GMM (MFCC+D) -> 0.944% EER
SV-GMM (MFCC+D+DD) -> 1.136% EER
i-vector (MFCC+D) -> 1.108% EER
i-vector (MFCC+D+DD) -> 3.614% EER
Female Speaker Detection Performance cond 2
SV-GMM (MFCC+D) -> 3.928% EER
SV-GMM (MFCC+D+DD) -> 5.881% EER
i-vector (MFCC+D) -> 3.911% EER
i-vector (MFCC+D+DD) -> 9.500% EER
Female Speaker Detection Performance cond 3
SV-GMM (MFCC+D) -> 3.826% EER
SV-GMM (MFCC+D+DD) -> 5.283% EER
i-vector (MFCC+D) -> 3.596% EER
i-vector (MFCC+D+DD) -> 4.911% EER
Female Speaker Detection Performance cond 4
SV-GMM (MFCC+D) -> 1.967% EER
SV-GMM (MFCC+D+DD) -> 2.915% EER
i-vector (MFCC+D) -> 3.387% EER
i-vector (MFCC+D+DD) -> 4.987% EER
Female Speaker Detection Performance cond 5
SV-GMM (MFCC+D) -> 3.670% EER
SV-GMM (MFCC+D+DD) -> 4.212% EER
i-vector (MFCC+D) -> 2.894% EER
i-vector (MFCC+D+DD) -> 4.265% EER
Female Speaker Detection Performance cond 6
SV-GMM (MFCC+D) -> 8.183% EER
SV-GMM (MFCC+D+DD) -> 12.973% EER
i-vector (MFCC+D) -> 8.382% EER
i-vector (MFCC+D+DD) -> 14.281% EER
Miss probability (in %)
False Alarm probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40

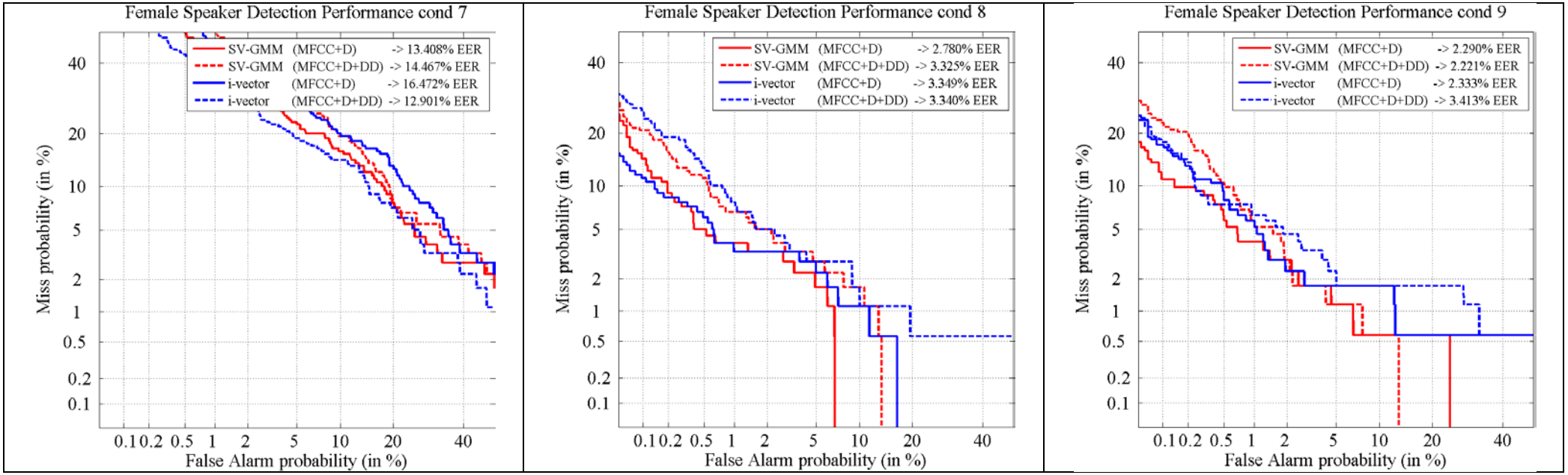



**Figure 5-107** DET curves for female speakers under different conditions comparing the use ΔΔ coefficients

Figure 5-106 (male speakers) and Figure 5-107 (female speakers) provide the DET curves corresponding to the different conditions that have been previously defined. Specifically, from left to right and from top to bottom DET curves represent the conditions 1 to 9. Solid line curves represent the results obtained by the systems when no ΔΔ coefficients are used, while dash line curves provide the recognition rates obtained when ΔΔ coefficients are included in the feature vectors. Additionally, blue lines refer to the system based on the *i*-vector approach, while red lines represent the results obtained when the SV-GMM classifier is used. From this set of features it is clear that the use of ΔΔ coefficients do not provide any benefit either for male or for female speakers in any of the conditions in which the development set have been divided, but for the condition 7 in the case of female speakers. Obviously, this result does not justify the incorporation of the ΔΔ coefficients into the feature vectors.

Once we have ruled out the use of ΔΔ coefficients, as we previously did in all the previously presented scenario, the next step consist in analyzing the effect of the different number of channels in the filter bank used to compute MFCCs as well as the number of MFCCs that will constitute the feature vectors. However, due to the amount of information that needs to be processed in each test, we have limited the values under evaluation, i.e. *F* and *MFCC*, even more than in the case of the scenario defined with the MOBIO database. And what is more, we are going to reuse the best configuration selected for the MOBIO scenario, in the current scenario, as some conditions exhibit certain similarities between both.

Therefore, in this case we will analyse the effect of using different number of filters in the filter bank, *F*= {30, 40, 50}, keeping fixed the number of *MFFC*=19. Additionally, we will test the settings providing best results in terms of *EER* on the MOBIO SRE. That is to say, {*F*=34, *MFCC*=24} for female speakers and {F=38, *MFCC*=28} for male speakers. It seems necessary to remind that this set of tests was carried out exclusively with the SV-GMM based recognition system.

After this limited analysis regarding the number of filters in the filter bank and the number of MFCCs, there are some conclusions that can be extracted. First of all, we can assert, as previously did with the experiments carried out on HESPERIA, ALBAYZIN and MOBIO databases, that a gender dependent parameterization provides better recognition rates in terms of *EER* than a gender independent parameterization. This is clear from the point of view of the value obtained for the *ACER* quality measure, since the most successful result for male speakers (*ACER*=2.539%) is obtained when *F*=30 and *MFCC*=19; whereas for female speakers, the best result in terms of *ACER* (3.966%) is obtained for the configuration in which *F*=50 and *MFCC*=19.

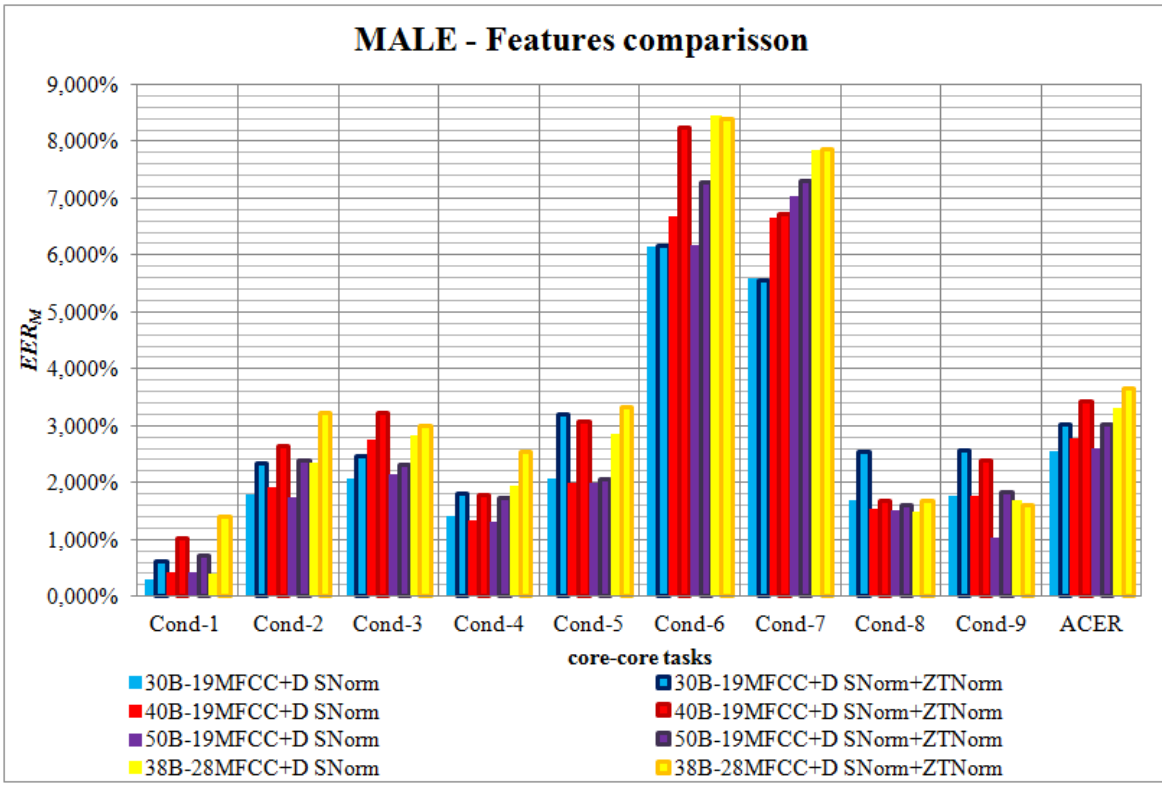



**Figure 5-108** *$EER_M$ and ACER* obtained depending on the *Baseline* front-end setup NIST SRE 2010

| | *F*=30 *MFCC*=19 | | *F*=40 *MFCC*=19 | | *F*=50 *MFCC*=19 | | *F*=38 *MFCC*=28 | |
|---|---|---|---|---|---|---|---|---|
| SV-GMM | SNorm | S+ZT Norm | SNorm | S+ZT Norm | SNorm | S+ZT Norm | SNorm | S+ZT Norm |
| *Cond 1* | 0.299% | 0.618% | 0.432% | 1.023% | 0.433% | 0.722% | 0.400% | 1.387% |
| *Cond 2* | 1.793% | 2.331% | 1.924% | 2.643% | 1.734% | 2.387% | 2.343% | 3.222% |
| *Cond 3* | 2.062% | 2.450% | 2.745% | 3.225% | 2.145% | 2.306% | 2.830% | 2.991% |
| *Cond 4* | 1.419% | 1.811% | 1.323% | 1.774% | 1.316% | 1.724% | 1.940% | 2.538% |
| *Cond 5* | 2.079% | 3.192% | 1.987% | 3.083% | 2.002% | 2.046% | 2.843% | 3.319% |
| *Cond 6* | 6.150% | 6.154% | 6.685% | 8.249% | 6.178% | 7.277% | 8.467% | 8.385% |
| *Cond 7* | 5.585% | 5.546% | 6.656% | 6.731% | 7.041% | 7.303% | 7.837% | 7.864% |
| *Cond 8* | 1.695% | 2.538% | 1.536% | 1.672% | 1.509% | 1.595% | 1.495% | 1.668% |
| *Cond 9* | 1.771% | 2.567% | 1.757% | 2.376% | 1.030% | 1.818% | 1.682% | 1.598% |
| *Acer* | 2.539% | 3.023% | 2.783% | 3.419% | 2.599% | 3.020% | 3.315% | 3.664% |

**Table 5-35** *$EER_M$ and ACER* obtained depending on the *Baseline* front-end setup NIST SRE 2010 (best results are highlighted in green)

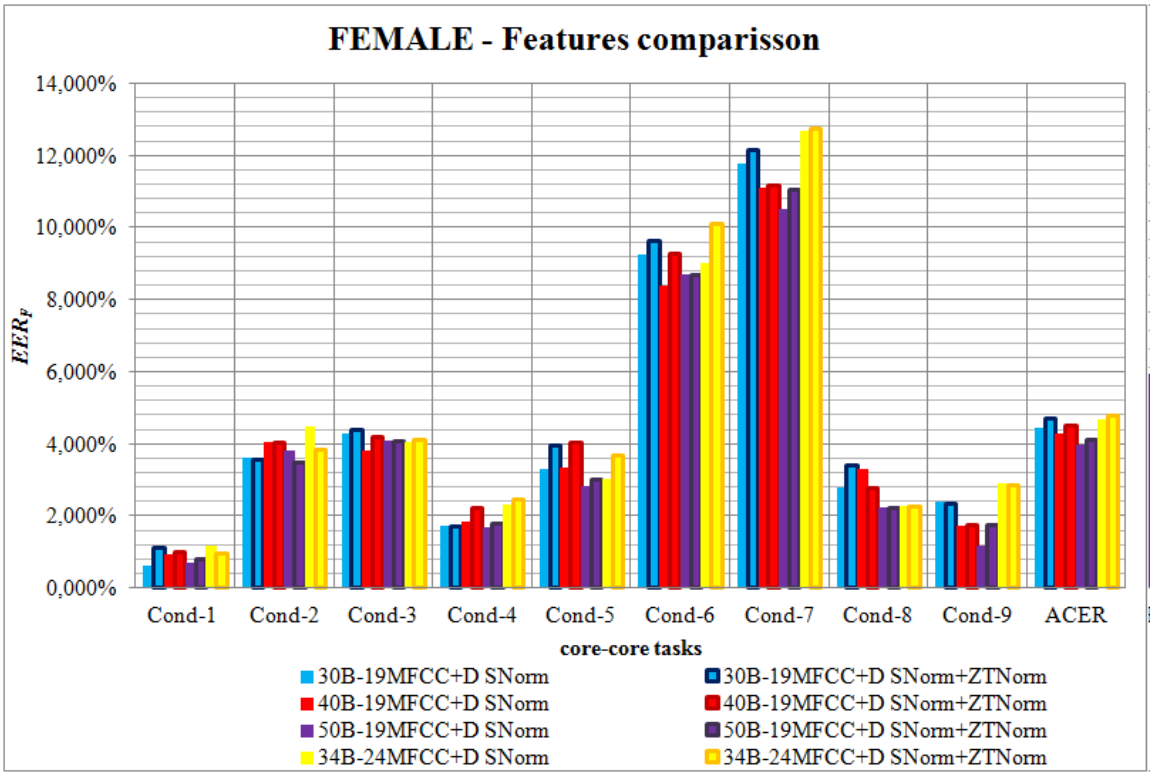



**Figure 5-109** *$EER_F$ and ACER* obtained depending on the *Baseline* front-end setup NIST SRE 2010

| | *F*=30 *MFCC*=19 | | *F*=40 *MFCC*=19 | | *F*=50 *MFCC*=19 | | *F*=38 *MFCC*=28 | |
|---|---|---|---|---|---|---|---|---|
| SV-GMM | SNorm | S+ZT Norm | SNorm | S+ZT Norm | SNorm | S+ZT Norm | SNorm | S+ZT Norm |
| *Cond 1* | 0.626% | 1.113% | 0.948% | 0.994% | 0.712% | 0.787% | 1.189% | 0.953% |
| *Cond 2* | 3.628% | 3.552% | 4.055% | 4.018% | 3.798% | 3.494% | 4.493% | 3.812% |
| *Cond 3* | 4.269% | 4.365% | 3.801% | 4.190% | 4.079% | 4.084% | 4.056% | 4.104% |
| *Cond 4* | 1.736% | 1.718% | 1.822% | 2.215% | 1.700% | 1.773% | 2.325% | 2.458% |
| *Cond 5* | 3.313% | 3.943% | 3.341% | 4.030% | 2.840% | 2.994% | 3.009% | 3.661% |
| *Cond 6* | 9.255% | 9.638% | 8.383% | 9.278% | 8.705% | 8.663% | 9.031% | 10.116% |
| *Cond 7* | 11.790% | 12.163% | 11.121% | 11.170% | 10.491% | 11.061% | 12.690% | 12.736% |
| *Cond 8* | 2.801% | 3.389% | 3.282% | 2.777% | 2.215% | 2.218% | 2.255% | 2.235% |
| *Cond 9* | 2.411% | 2.320% | 1.741% | 1.741% | 1.153% | 1.750% | 2.906% | 2.851% |
| *Acer* | 4.425% | 4.689% | 4.277% | 4.490% | 3.966% | 4.092% | 4.662% | 4.770% |

**Table 5-36** *$EER_F$ and ACER* obtained depending on the *Baseline* front-end setup NIST SRE 2010 (best results are highlighted in green)

Another aspect deserving consideration is an effect that we have already faced in the two scenarios defined for the HESPERIA database. Particularly, when we face different recording conditions, the configuration providing the most successful results is different in terms of filters in filter bank and number of MFCCs for each recording condition. This fact can be verified on Table 5-35 and Table 5-36, in which, regardless the speaker's gender, the better results in terms of *EER* for different conditions are obtained for different configurations. This is justified by the variability between the recordings belonging to different conditions, regarding type of transmission channel, type of

microphone, presence of noise and emotional state of speakers. In other words, for male speakers and conditions 1, 3, 6 and 7 the configuration providing better results is {*F*=30, *MFCC*=19}; for conditions 2, 4, 5 and 9 the preferred configuration is {*F*=50, *MFCC*=19}; and for condition 8 {*F*=38, *MFCC*=28} provides the most successful results. The same analysis can be performed for female speakers. But in this case, the configuration providing the most successful results for almost all conditions is {*F*=50, *MFCC*=19}. Base on these results we can also conclude that the use of *ACER* as quality measure, somewhat hides these singularities.

Simultaneously, we have evaluated the use of two types of score normalization techniques, namely SNorm and the combination of SNorm and ZTNorm. In this sense we can conclude that the use of ZTNorm in the case of male speakers do not provide an additional benefit either in terms of *ACER* or in terms of *EER* for all conditions, but for *Cond* 7. In the case of female speakers, applying ZTNorm combined with SNorm generates better results in terms of *EER* for condition 2, while for the remaining conditions better results are obtained when ZTNorm is not applied. Moreover, in terms of *ACER*, better results are obtained when SNorm is applied alone for all the defined configurations.

Finally, it must be noted that the configuration selected from the scenario defined for the MOBIO database does not improve the results obtained by the other tested configurations for any condition, but for condition 8 on male speakers. Therefore these results confirm the trend that we have seen in the different scenarios addressed so far. That is to say, the set of parameters selected for a specific scenario using a particular database is not straightforwardly transferable to another scenario using either the same or another database.

The next step, like in previous scenarios, consist in checking the usefulness of the extra parameters, namely E, ΔE, F0 and F3, that have been successfully used before. To this end we select, for each gender the configuration that provides the most successful results in terms of *ACER*. Therefore, we are following a gender dependent approach. For male speakers, we have reduced the set of tests so ZTNorm has not been used in combination with SNorm, since as previously concluded, its use did not provide any benefit. Specifically, we run a set of tests in which each of these extra parameters were included in the feature vector defined by the GDC {*F*=30, *MFCC*=19} either alone or combined with the others.

Figure 5-110 reflects the results obtained for each combination in terms of *ACER*, where horizontal lines represent the *ACER* obtained by the configurations shown in Table 5-35. In this case, it is clear that the inclusion of the extra parameters provides an additional improvement in terms of *ACER*, with respect to the case in which no extra parameters are used for multiple configurations. However, there is a winner configuration that provides a significant improvement, specifically the use of F0 and F3 allows for a relative reduction of 8.4% in terms of *ACER* (from $ACER_{GDC}$=2.539% to $ACER_{GDC+F0+F3}$=2.326%). As stated above the use of *ACER* as quality measure somehow masks the results, in terms of *EER*, obtained for the different conditions in which the development set is divided. To shed light on this fact, Figure 5-111 provides the results in terms of *EER* obtained by both configurations, namely GDC (light blue) and GDC+F0+F3 (dark blue), for each of the tested conditions. From the results provided in Table 5-37, it is clear that the inclusion of the F0 and F3 extra parameters in the feature vectors systematically helps in the reduction of *EER*, except for *Cond 5*.

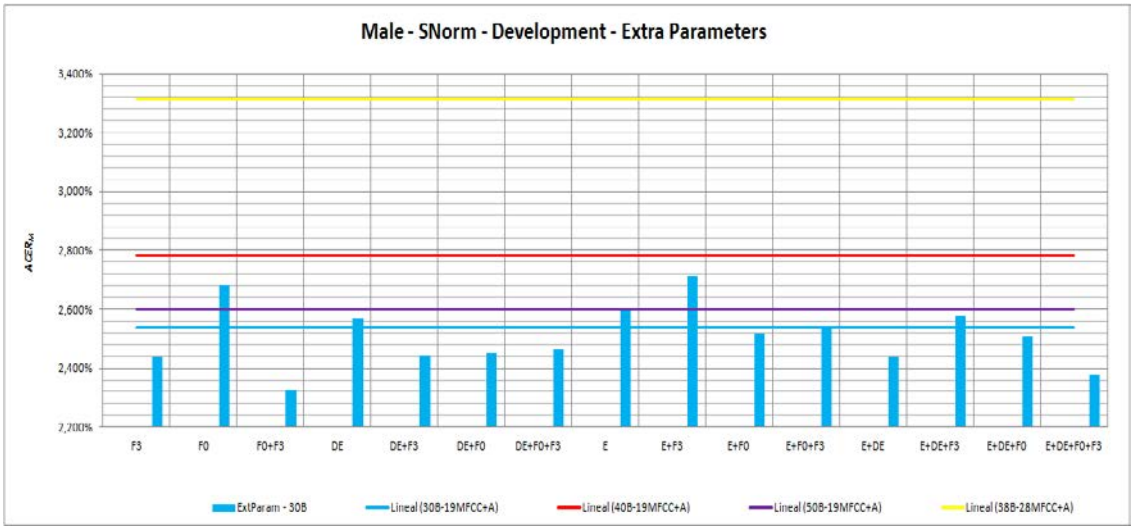



**Figure 5-110** *ACER* obtained for GDC when extra parameters are incorporated into the feature vector and SNorm is applied for male speakers

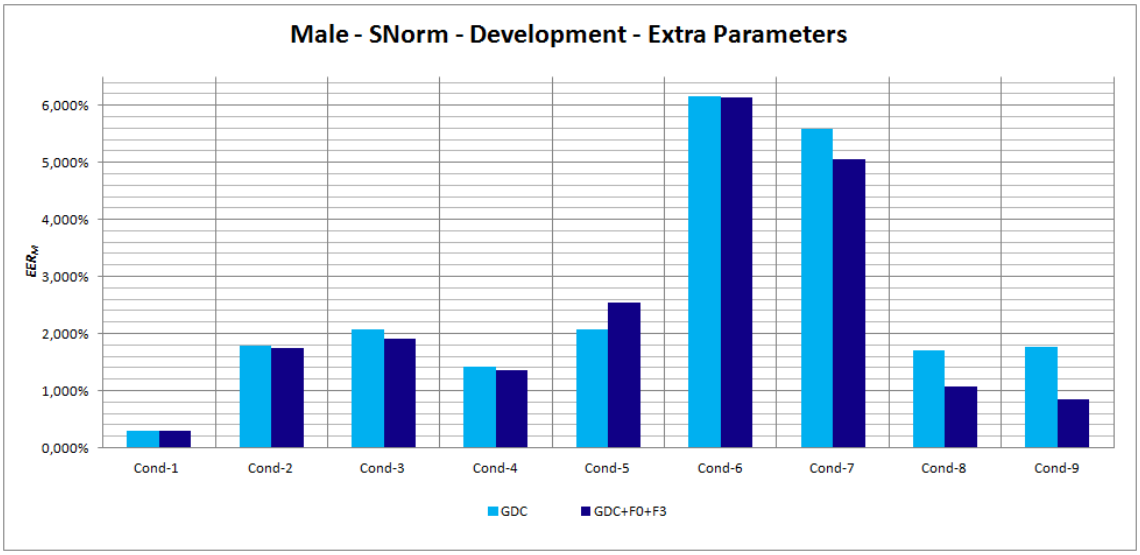



**Figure 5-111** $EER_M$ comparison for the different conditions in which development set is divided

| SV-GMM | *GDC* | *GDC*+F0+F3 | *Relative Reduction* |
|---|---|---|---|
| *Cond 1* | 0.299% | 0.290% | 2.97% |
| *Cond 2* | 1.793% | 1.741% | 2.93% |
| *Cond 3* | 2.062% | 1.901% | 7.81% |
| *Cond 4* | 1.419% | 1.349% | 4.94% |
| *Cond 5* | 2.079% | 2.537% | -22.05% |
| *Cond 6* | 6.150% | 6.139% | 0.19% |
| *Cond 7* | 5.585% | 5.060% | 9.41% |
| *Cond 8* | 1.695% | 1.061% | 37.39% |
| *Cond 9* | 1.771% | 0.853% | 51.85% |
| *Acer* | 2.539% | 2.326% | 8.42% |

**Table 5-37** $EER_M$ comparison for the different conditions in which development set is divided

In the case of female speakers, Figure 5-112 (SNorm) and Figure 5-113 (SNorm + ZTNorm) provide the results obtained for each combination in terms of *ACER*, where horizontal lines represent the *ACER* obtained by the configurations shown in Table 5-36. In this case, it is clear that the inclusion of the extra parameters does not systematically provide an additional improvement in terms of *ACER*, with respect to the case in which no extra parameters are used. However, there is still a configuration that provides a slight improvement, specifically the use of E, F0 and F3 allows for a relative reduction up to 6% in terms of *ACER* (from $ACER_{GDC}$=4.092% to $ACER_{GDC+E+F0+F3}$=3.840%) when ZTNorm and SNorm are combined. It must be noted as well, that this configuration also provides a relative reduction close to 2% in terms of *ACER*, (from $ACER_{GDC}$=3.966% to $ACER_{GDC+E+F0+F3}$=3.889%) when just SNorm is used.

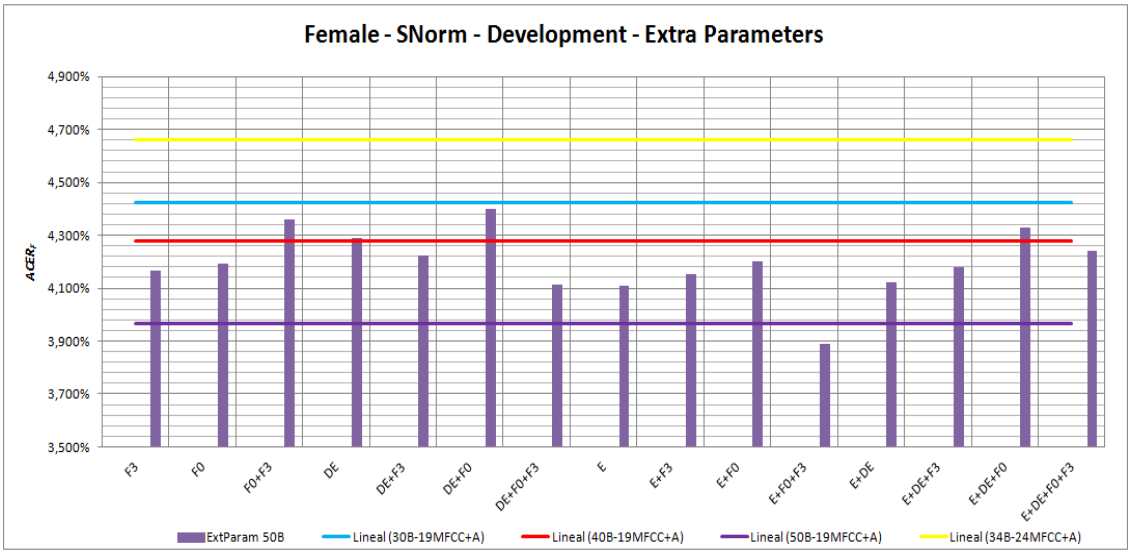



**Figure 5-112** *ACER* obtained for GDC when extra parameters are incorporated into the feature vector and SNorm is applied for female speakers

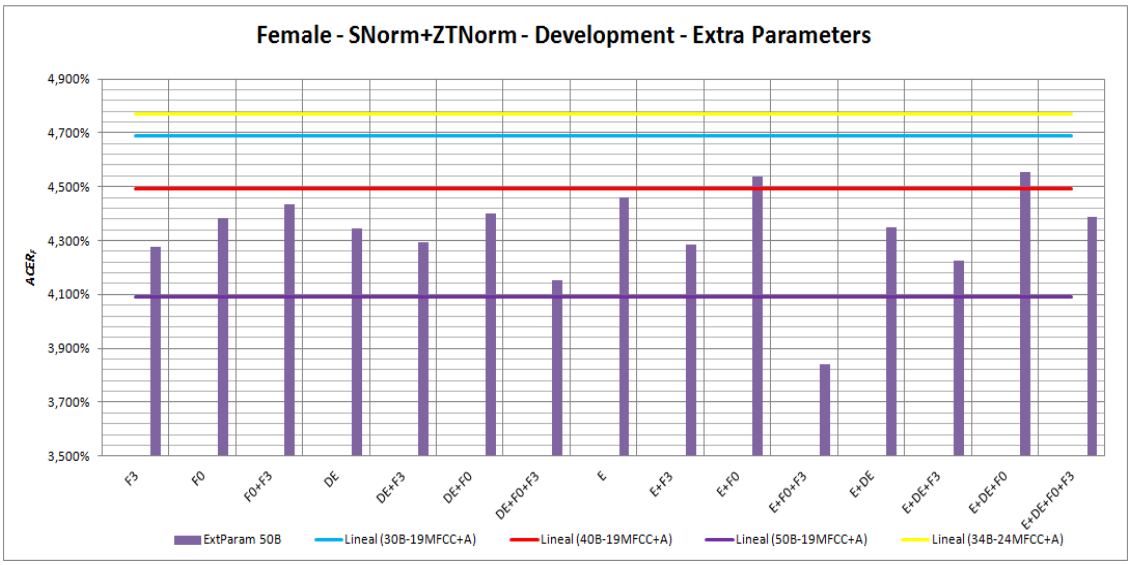



**Figure 5-113** *ACER* obtained for GDC when extra parameters are incorporated into the feature vector and SNorm is combined with ZTNorm for female speakers

| | SNorm | | | SNorm + ZTNorm | | |
|---|---|---|---|---|---|---|
| SV-GMM | GDC | GDC+E+F0+F3 | Relative Reduction | GDC | GDC+E+F0+F3 | Relative Reduction |
| *Cond 1* | 0.712% | 0.855% | -20.04% | 0.787% | 1.165% | -47.96% |
| *Cond 2* | 3.798% | 3.740% | 1.54% | 3.494% | 3.422% | 2.06% |
| *Cond 3* | 4.079% | 3.676% | 9.89% | 4.084% | 3.770% | 7.69% |
| *Cond 4* | 1.700% | 1.974% | -16.07% | 1.773% | 1.697% | 4.31% |
| *Cond 5* | 2.840% | 2.652% | 6.61% | 2.994% | 2.843% | 5.02% |
| *Cond 6* | 8.705% | 7.928% | 8.93% | 8.663% | 7.635% | 11.86% |
| *Cond 7* | 10.491% | 10.350% | 1.35% | 11.061% | 10.042% | 9.21% |
| *Cond 8* | 2.215% | 2.125% | 4.04% | 2.218% | 2.258% | -1.82% |
| *Cond 9* | 1.153% | 1.705% | -47.90% | 1.750% | 1.726% | 1.38% |
| *Acer* | 3.966% | 3.889% | 1.93% | 4.092% | 3.840% | 6.15% |

**Table 5-38** $EER_F$ comparison for the different conditions in which development set is divided

Table 5-38 provides the individual results obtained in terms of *EER* for all the conditions in which the development set is divided. Different columns are provided for the different score normalization techniques applied, namely, SNorm and SNorm combined with ZTNorm. The most successful results are highlighted in green for each condition and for each score normalization technique. In the case of applying SNorm, 6 conditions out of 9 show a reduction in terms of EER when selected extra parameters are used, whereas in the case of combining ZTNorm and SNorm, 7 out of 9 show an improvement in terms of *EER* when E+F0+F3 parameters are used. The fact that for female speakers, the use of ZTNorm combined with SNorm provides better results can be explained by the fact that the amount of data available for normalization purposes is higher for female speakers than for male speakers, therefore covering the impostor's space more precisely. However this improvement is quite small.

To conclude this set of tests, it is worth noting, despite the limited amount of tests that we have run on the development data, that once again a gender dependent approach provides better recognition rates than a gender independent one in order to precisely characterise speakers. Additionally, we have found again that the use of the proposed extra parameters as a complement to classical parameters also helps to increase recognition rates. Particularly, the use of Pitch (F0) and 3$^{rd}$ formant estimation (F3) helps in the reduction of *EER*, both for male and female speakers, in most of the tested conditions. Additionally, we see again how the use of E and ΔE combined, widely used in speaker recognition systems, is not the best option either for male or for female speakers.

The next step, like in the previous scenario, consists in introducing what we have called extended-biometric parameters extracted by the GDEB front-end. The approach that has been followed, consists in incorporating the set of parameters extracted from the glottal source estimate (labelled as GSE) into the most successful setup, i.e. GDC MFCCs+Δ+F0+F3 in the case of male speakers and GDC MFCCs+Δ+E+F0+F3 for female speakers. Where GDC={*F*=30, *MFCC*=19} for male speakers and GDC={*F*=50, *MFCC*=19} for female speakers. The use of parameters extracted from the vocal tract

estimate has been ruled out as their use has not provided additional improvements respect to the use of GSE parameters in previous scenarios.

It is worth noting the ambition of this set of tests. Ambitious, in the sense that the main goal is the reduction of the quality measure *ACER*, implying that somehow we are pursuing the simultaneous reduction of *EER* for all conditions in which development is divided. In previous experiments we have seen that the selected configuration, for GSE parameters, providing better results in terms of *EER* is different deepening on the type of recordings under analysis. Thus the search for a generic configuration for GSE parameters providing an improvement in terms of *ACER* and therefore in terms of *EER*, may be seen as a chimera. However, even in a scenario with such variability, the use of parameters extracted from the glottal source estimate allows for a more accurate characterization of speakers and therefore an improvement in the recognition rates can be obtained.

Although multiple configurations have been tested, Table 5-39 for male speakers and, Table 5-40 (SNorm) and Table 5-41 (SNorm+ZTNorm) for female speakers, show the final configurations chosen for each gender, as well as the recognition rates obtained in each case in terms of *EER* (best highlighted in green) and *ACER*. Additionally, the relative reduction (RR) in terms of *EER*, if compared to the GIC, is also presented in brackets for each configuration, provided that the comparison with GIC instead of GDC is based on the fact that GIC is considered the state-of-the-art to beat in the front-end subsystem.

| MALE | SNorm | | | |
|---|---|---|---|---|
| SV-GMM | GIC | GDC [RR] | GDC+F0+F3 [RR] | GSE [RR] |
| CONFIG | *F*=40 *MFCC*=19 | *F*=30 *MFCC*=19 | *F*=30 *MFCC*=19 | *F*=30 *MFCC*=19 **Source-Tract Sep. Alg:** Prediction Order: 32 Forgetting Factor: 0.995 **GSE:** 32-Channel Filter bank 6 MFCC |
| *Cond 1* | 0.432% | 0.299% [30.67%] | 0.290% [32.73%] | 0.298% [31.08%] |
| *Cond 2* | 1.924% | 1.793% [6.81%] | 1.741% [9.55%] | 1.662% [13.61%] |
| *Cond 3* | 2.745% | 2.062% [24.87%] | 1.901% [30.73%] | 2.270% [17.31%] |
| *Cond 4* | 1.323% | 1.419% [-7.23%] | 1.349% [-1.93%] | 1.094% [17.36%] |
| *Cond 5* | 1.987% | 2.079% [-4.59%] | 2.537% [-27.65%] | 2.242% [-12.82%] |
| *Cond 6* | 6.685% | 6.150% [7.99%] | 6.139% [8.17%] | 5.593% [16.34%] |
| *Cond 7* | 6.656% | 5.585% [16.09%] | 5.060% [23.99%] | 6.002% [9.84%] |
| *Cond 8* | 1.536% | 1.695% [-10.36%] | 1.061% [30.91%] | 0.852% [44.52%] |
| *Cond 9* | 1.757% | 1.771% [-0.80%] | 0.853% [51.46%] | 0.848% [51.73%] |
| *Acer* | 2.783% | 2.539% [8.75%] | 2.326% [16.43%] | 2.318% [16.71%] |

**Table 5-39** $EER_M$ obtained on the development set (SNorm), comparing classical parameters with extra parameters and extended biometric parameters

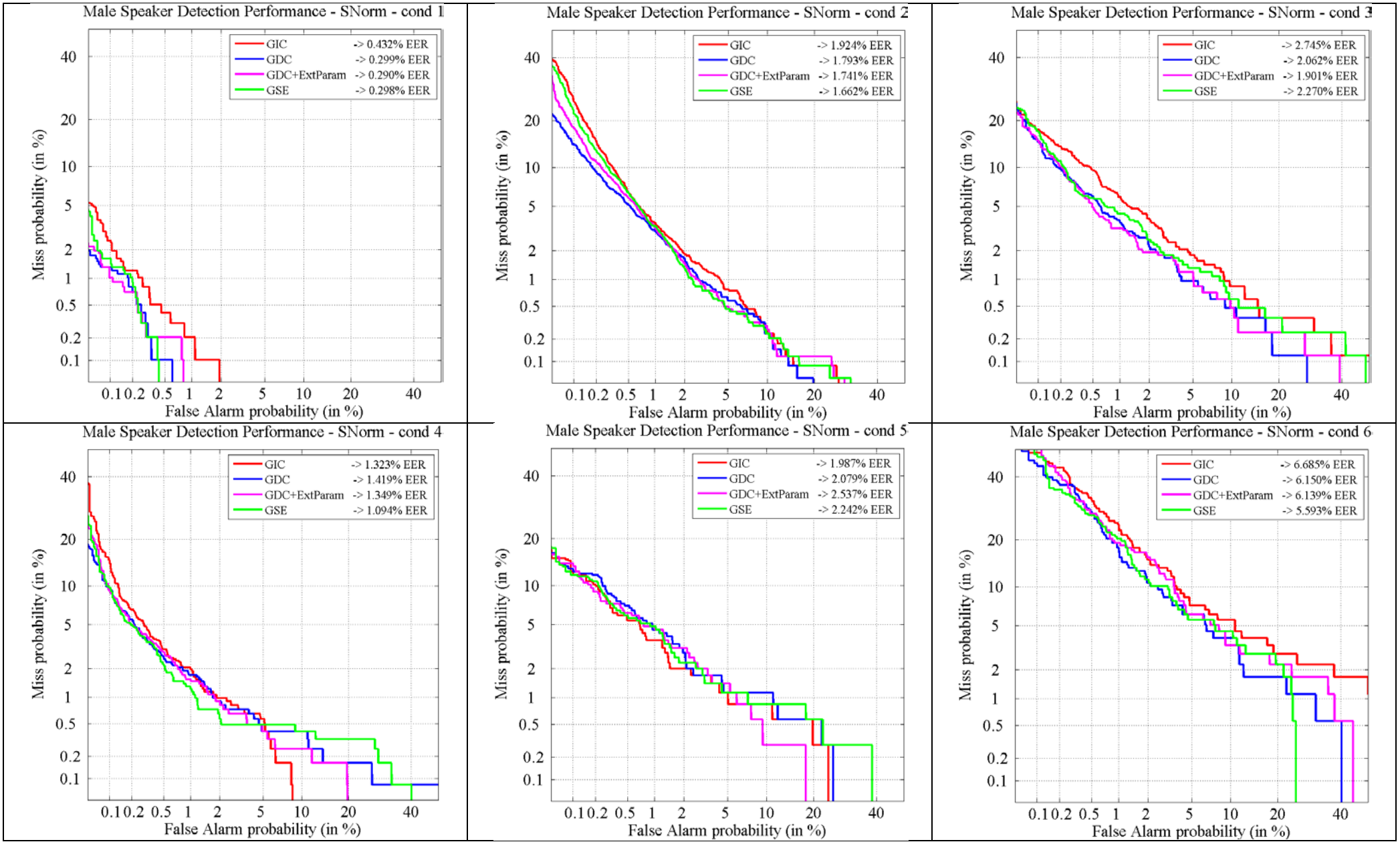
| FEMALE | SNorm | | | |
|---|---|---|---|---|
| SV-GMM | GIC | GDC [RR] | GDC+E+F0+F3 [RR] | GSE [RR] |
| CONFIG | *F*=40 *MFCC*=19 | *F*=50 *MFCC*=19 | *F*=50 *MFCC*=19 | *F*=50 *MFCC*=19 **Source-Tract Sep. Alg:** Prediction Order: 29 Forgetting Factor: 0.995 **GSE:** 36-Channel Filter bank 4 MFCC |
| *Cond 1* | 0,948% | 0,712% [24.92%] | 0,855% [9.87%] | 0,727% [23.30%] |
| *Cond 2* | 4,055% | 3,798% [6.34%] | 3,740% [7.77%] | 3,599% [11.25%] |
| *Cond 3* | 3,801% | 4,079% [-7.31%] | 3,676% [3.31%] | 3,307% [13.01%] |
| *Cond 4* | 1,822% | 1,700% [6.67%] | 1,974% [-8.34%] | 1,794% [1.55%] |
| *Cond 5* | 3,341% | 2,840% [14.99%] | 2,652% [20.62%] | 2,662% [20.33%] |
| *Cond 6* | 8,383% | 8,705% [-3.84%] | 7,928% [5.43%] | 7,635% [8.92%] |
| *Cond 7* | 11,121% | 10,491% [5.66%] | 10,350% [6.93%] | 10,261% [7.73%] |
| *Cond 8* | 3,282% | 2,215% [32.51%] | 2,125% [35.24%] | 2,131% [35.06%] |
| *Cond 9* | 1,741% | 1,153% [33.80%] | 1,705% [2.08%] | 1,165% [33.10%] |
| *Acer* | 4,277% | 3,966% [7.28%] | 3,889% [9.07%] | 3,698% [13.54%] |

**Table 5-40** $EER_F$ obtained on the development set (SNorm), comparing classical parameters with extra parameters and extended biometric parameters

| FEMALE | SNorm + ZTNorm | | | |
|---|---|---|---|---|
| SV-GMM | GIC | GDC [RR] | GDC+E+F0+F3 [RR] | GSE [RR] |
| CONFIG | *F*=40 *MFCC*=19 | *F*=50 *MFCC*=19 | *F*=50 *MFCC*=19 | *F*=50 *MFCC*=19 **Source-Tract Sep. Alg:** Prediction Order: 29 Forgetting Factor: 0.995 **GSE:** 17-Channel Filter bank 2 MFCC |
| *Cond 1* | 0,994% | 0,787% [20.83%] | 1,165% [-17.13%] | 1,044% [-4.94%] |
| *Cond 2* | 4,018% | 3,494% [13.03%] | 3,422% [14.82%] | 3,255% [18.99%] |
| *Cond 3* | 4,190% | 4,084% [2.53%] | 3,770% [10.02%] | 3,406% [18.71%] |
| *Cond 4* | 2,215% | 1,773% [19.92%] | 1,697% [23.37%] | 2,060% [6.96%] |
| *Cond 5* | 4,030% | 2,994% [25.72%] | 2,843% [29.45%] | 3,116% [22.69%] |
| *Cond 6* | 9,278% | 8,663% [6.63%] | 7,635% [17.71%] | 8,221% [11.39%] |
| *Cond 7* | 11,170% | 11,061% [0.97%] | 10,042% [10.09%] | 8,967% [19.72%] |
| *Cond 8* | 2,777% | 2,218% [20.15%] | 2,258% [18.69%] | 1,716% [38.21%] |
| *Cond 9* | 1,741% | 1,750% [-0.52%] | 1,726% [0.87%] | 1,738% [0.17%] |
| *Acer* | 4,490% | 4,092% [8.88%] | 3,840% [14.49%] | 3,725% [17.05%] |

**Table 5-41** $EER_F$ obtained on the development set (SNorm+ZTNorm), comparing classical parameters with extra parameters and extended biometric parameters

The DET curves representing the results obtained with each of the previously presented configurations (see Table 5-39, Table 5-40, and Table 5-41) are depicted in Figure 5-114 for male speakers and in Figure 5-115 and Figure 5-116 for female speakers.


Male Speaker Detection Performance - SNorm - cond 1
GIC -> 0.432% EER
GDC -> 0.299% EER
GDC+ExtParam -> 0.290% EER
GSE -> 0.298% EER
Male Speaker Detection Performance - SNorm - cond 2
GIC -> 1.924% EER
GDC -> 1.793% EER
GDC+ExtParam -> 1.741% EER
GSE -> 1.662% EER
Male Speaker Detection Performance - SNorm - cond 3
GIC -> 2.745% EER
GDC -> 2.062% EER
GDC+ExtParam -> 1.901% EER
GSE -> 2.270% EER
Male Speaker Detection Performance - SNorm - cond 4
GIC -> 1.323% EER
GDC -> 1.419% EER
GDC+ExtParam -> 1.349% EER
GSE -> 1.094% EER
Male Speaker Detection Performance - SNorm - cond 5
GIC -> 1.987% EER
GDC -> 2.079% EER
GDC+ExtParam -> 2.537% EER
GSE -> 2.242% EER
Male Speaker Detection Performance - SNorm - cond 6
GIC -> 6.685% EER
GDC -> 6.150% EER
GDC+ExtParam -> 6.139% EER
GSE -> 5.593% EER
Miss probability (in %)
False Alarm probability (in %)

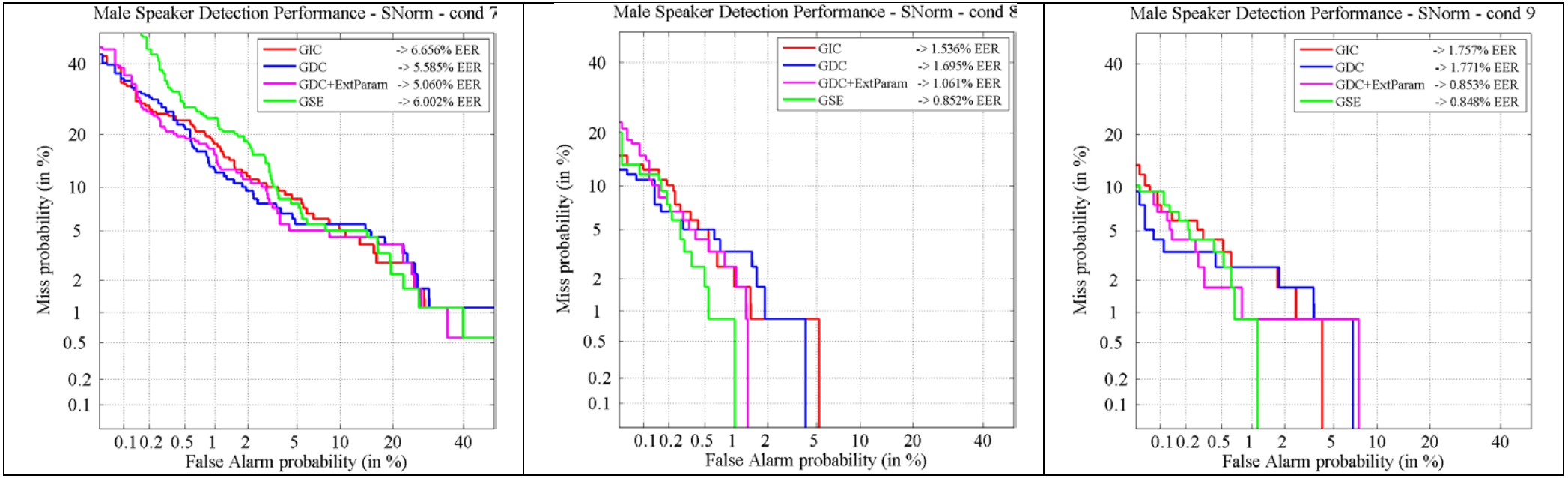
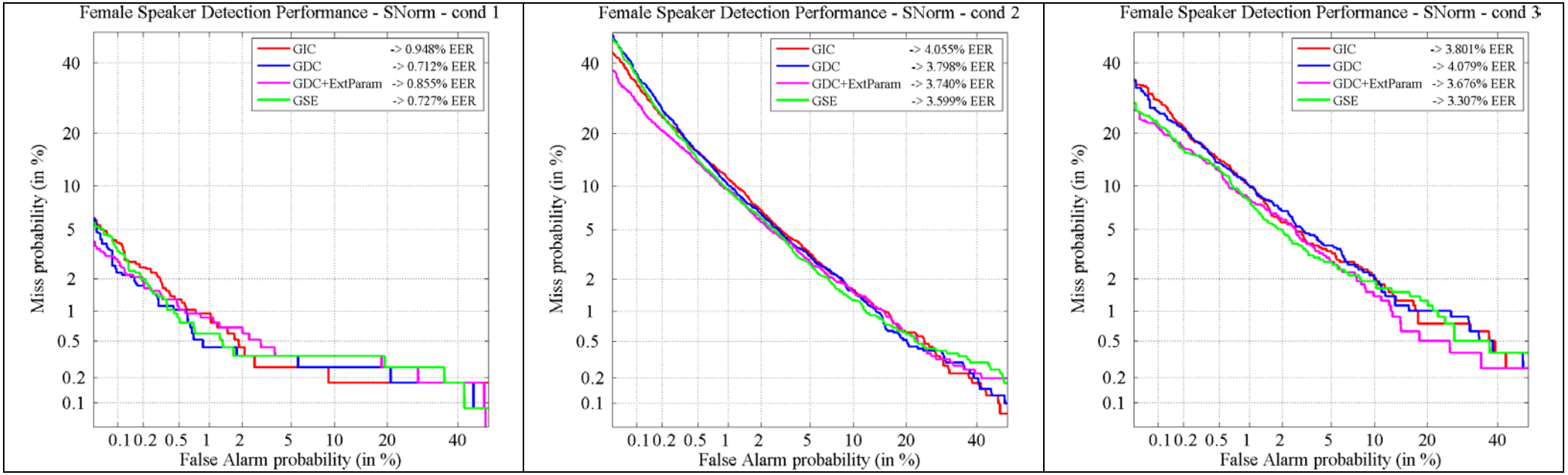



**Figure 5-114** DET curves comparing classical parameters and GDEB on the NIST SRE10 development set for male speakers and SNorm



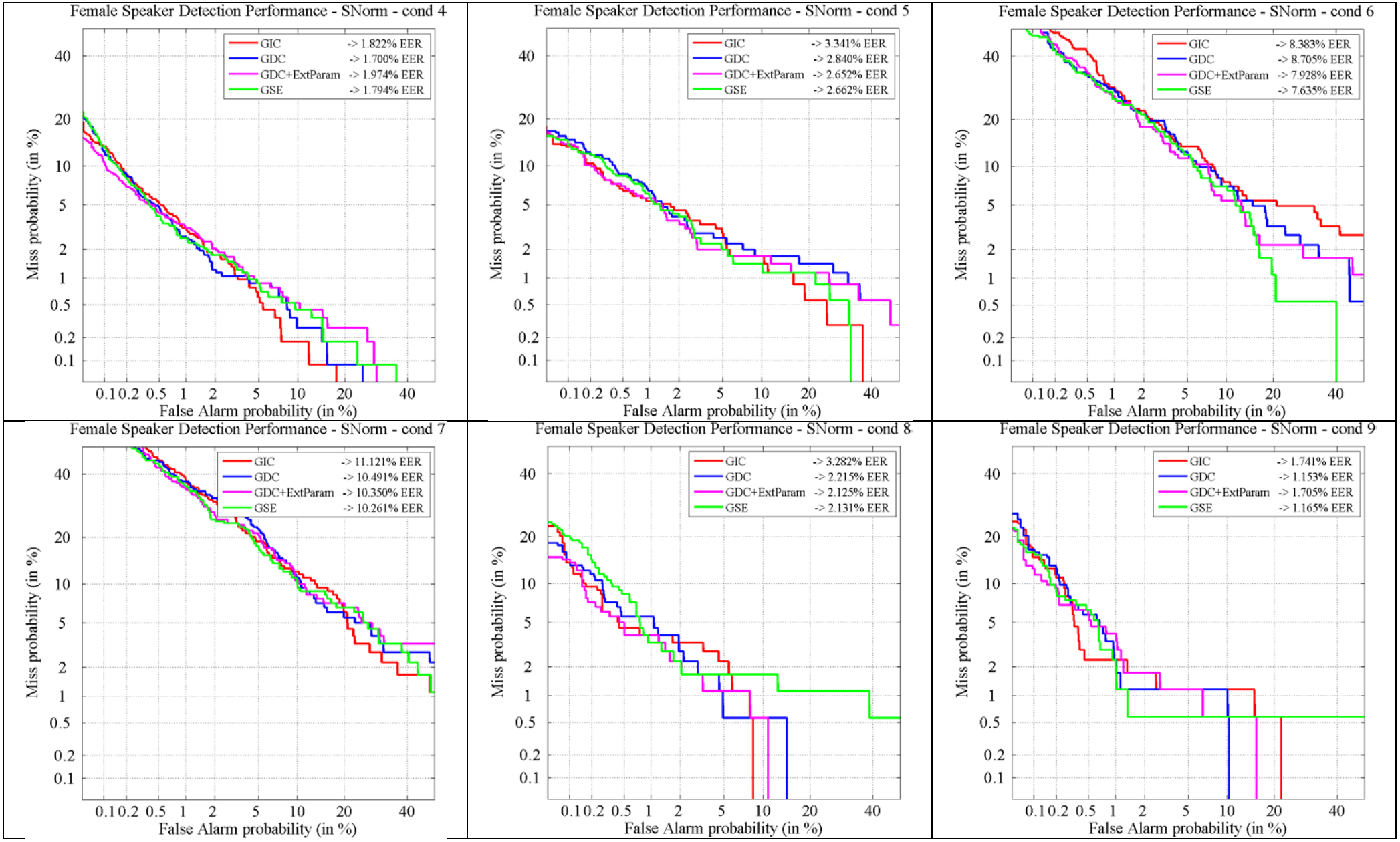
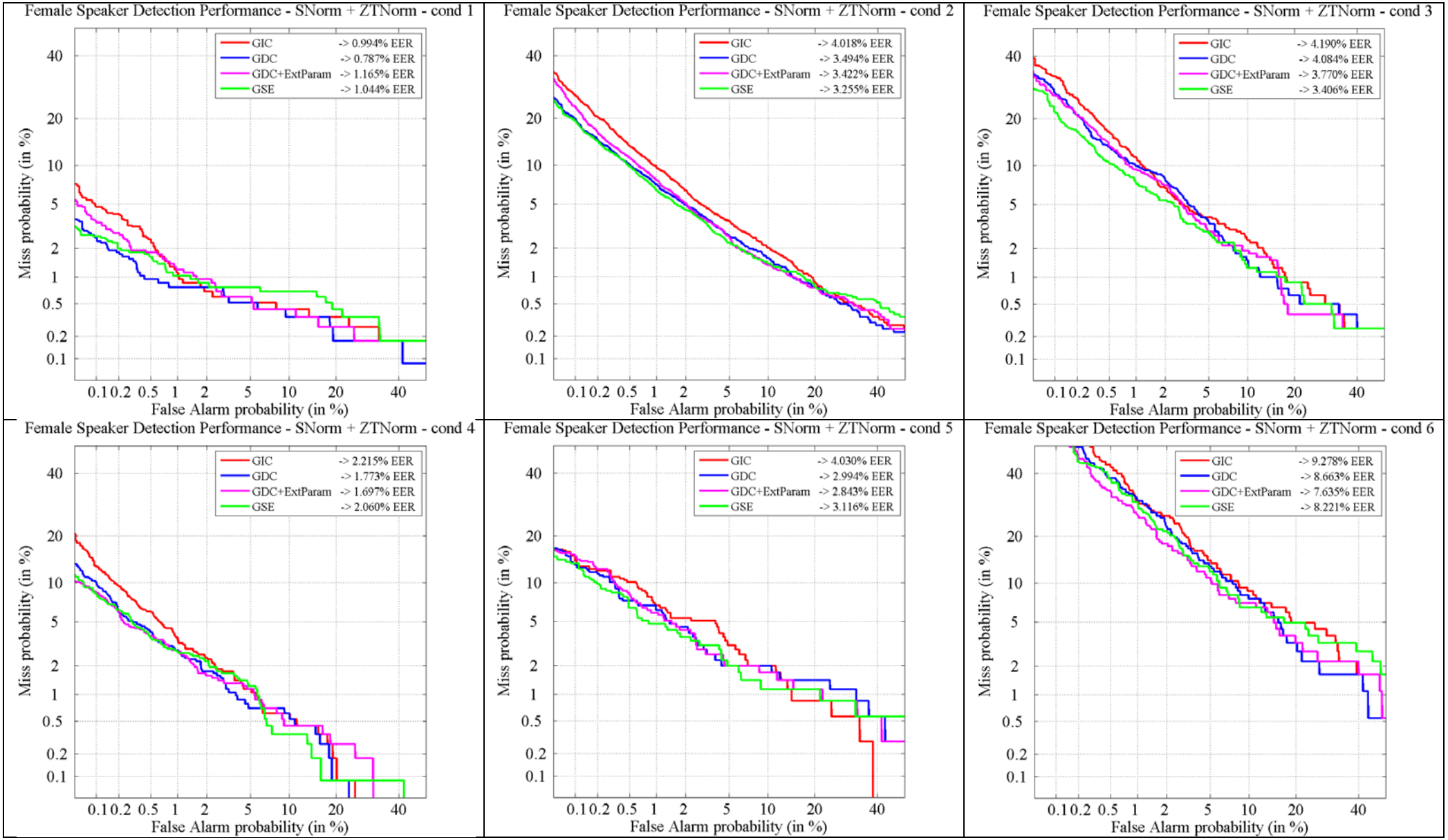



**Figure 5-115** DET curves comparing classical parameters and GDEB on the NIST SRE10 development set for female speakers and SNorm


Female Speaker Detection Performance - SNorm + ZTNorm - cond 1
GIC -> 0.994% EER
GDC -> 0.787% EER
GDC+ExtParam -> 1.165% EER
GSE -> 1.044% EER
Miss probability (in %)
False Alarm probability (in %)
Female Speaker Detection Performance - SNorm + ZTNorm - cond 2
GIC -> 4.018% EER
GDC -> 3.494% EER
GDC+ExtParam -> 3.422% EER
GSE -> 3.255% EER
Miss probability (in %)
False Alarm probability (in %)
Female Speaker Detection Performance - SNorm + ZTNorm - cond 3
GIC -> 4.190% EER
GDC -> 4.084% EER
GDC+ExtParam -> 3.770% EER
GSE -> 3.406% EER
Miss probability (in %)
False Alarm probability (in %)
Female Speaker Detection Performance - SNorm + ZTNorm - cond 4
GIC -> 2.215% EER
GDC -> 1.773% EER
GDC+ExtParam -> 1.697% EER
GSE -> 2.060% EER
Miss probability (in %)
False Alarm probability (in %)
Female Speaker Detection Performance - SNorm + ZTNorm - cond 5
GIC -> 4.030% EER
GDC -> 2.994% EER
GDC+ExtParam -> 2.843% EER
GSE -> 3.116% EER
Miss probability (in %)
False Alarm probability (in %)
Female Speaker Detection Performance - SNorm + ZTNorm - cond 6
GIC -> 9.278% EER
GDC -> 8.663% EER
GDC+ExtParam -> 7.635% EER
GSE -> 8.221% EER
Miss probability (in %)
False Alarm probability (in %)

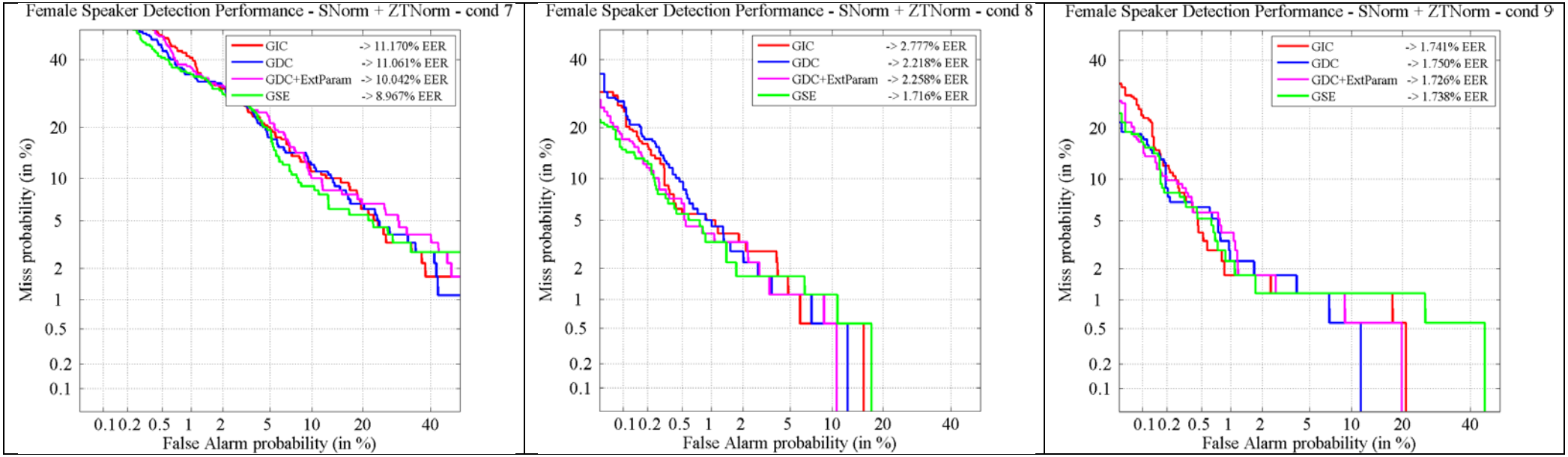



**Figure 5-116** DET curves comparing classical parameters and GDEB on the NIST SRE10 development set for female speakers and SNorm+ZTNorm

Despite the limits set to the experiments carried out on this scenario, mainly due to the volume of data to be processed, we can draw certain conclusion. According to the results, the use of a gender dependent configuration is essential in order to provide more accurate representation of speakers, and therefore to allow for a significant improvement in the performance of the recognition systems, in terms of reduction of *EER*. Additionally, the incorporation of what we have called extra parameters, in this case E, F0, and F3 for female speakers, and F0 and F3 for male speakers into the feature vectors, provide further advantage. And what is more important, the use of the parameter F3 proposed in this thesis has become a relevant option in all the analysed scenarios, proving its value for speaker characterization purposes.

Furthermore the proposed gender-dependent extended biometric parameterisation, in this case incorporating information just from the glottal source estimate in form of MFCCs, is the configuration that provides the most successful results in the development set for both male and female speakers, in terms of *ACER*. Specifically, for male speakers, the use of the GSE setup, thus a gender-dependent configuration incorporating extended biometric features, provides a relative reduction of 16.7% in terms of $ACER_M$, respect to the gender-independent configuration. Whereas in the case of female speakers, the use of the GSE setup allows for a relative reduction close to 17% in terms of $ACER_F$, with respect to the gender-independent configuration, in the case of combining ZTNorm and SNorm, and 13.5% in the case of applying SNorm alone. Although the relative reduction in terms of *ACER* is smaller in this last case, the most successful result for female speakers is obtained when ZTNorm is not used ($ACER_{F\text{-}SNORM}$=3.698% vs. $ACER_{F\text{-}SNORM\text{-}ZTNorm}$=3.725%). However, we can also verify an aspect which we have already mentioned, and that is none other than the difficulty of finding a generic configuration for GSE which systematically provides an improvement on the recognition rates for all the conditions of the development set individually.

Finally, it may be interesting to compare the results obtained by our system, using a gender-dependent extended biometric parameterisation, with the *EER* obtained by state-of-art systems submitted to the NIST SRE 2010 (see Table 5-42). It must be noted that NIST do not provide separate information depending on the gender, therefore straightforward comparison is not adequate. However, it must be noted that our system is able to provide better recognition rates in terms of *EER*, than state-of-art systems (highlighted in green) for some conditions, especially for male speakers.

| Train Condition | State-of-art *EER* range | $EER_M$ | $EER_F$ |
|---|---|---|---|
| *Cond 1* | 1.0% - 2.0% | 0.298% | 0,727% |
| *Cond 2* | 2.0% - 4.0% | 1.662% | 3,599% |
| *Cond 3* | 1.5% - 2.0% | 2.270% | 3,307% |
| *Cond 4* | 1.5% - 2.5% | 1.094% | 1,794% |
| *Cond 5* | 1.5% - 2.5% | 2.242% | 2,662% |
| *Cond 6* | 2.0% - 5.0% | 5.593% | 7,635% |
| *Cond 7* | 4.0% - 5.0% | 6.002% | 10,261% |
| *Cond 8* | 0.5% - 1.0% | 0.852% | 2,131% |
| *Cond 9* | 1.0% - 2.0% | 0.848% | 1,165% |

**Table 5-42** *EER* ranges for the different tasks on NIST SRE 2010

Once we have established the configurations that allow us to obtain the best results in terms of *ACER*, on the development set, the next step, consists in checking the behaviour of the system on the evaluation set. In this sense it is necessary to take into

account that, contrary to what happened in the scenarios defined for the HESPERIA, ALBAYZIN and even MOBIO database, where the evaluation conditions were similar to the development conditions; in this case the evaluation scenario (based on NIST SRE 2012) clearly differs form the scenario defined for development. It mainly differs in two important aspects. First of all, one or more samples of speech data (recordings) coming from multiple channels (microphone and telephone recordings), are available for training/creating the speaker's model, whereas in the development set, only one recording is available. Secondly, some of the test segments will have fake additive noise imposed. These two differences, specially the second one, may cause the selected configurations on the development set not to work properly.

Table 5-43 for male, while Table 5-44 and Table 5-45 for female speakers, show the results obtained on the evaluation set using the different configurations previously specify in Table 5-39, Table 5-40, and Table 5-41.

| MALE | SNORM | | |
|---|---|---|---|
| SV-GMM | GIC | GDC+F0+F3 [RR] | GSE [RR] |
| CONFIG | *F*=40 *MFCC*=19 | *F*=30 *MFCC*=19 | *F*=30 *MFCC*=19 **Source-Tract Sep. Alg:** Prediction Order: 32 Forgetting Factor: 0.995 **GSE:** 32-Channel Filter bank 6 MFCC |
| *CC – 0* | 26,184% | 26,117% | 24,691% |
| *CC – 1* | 9,241% | 9,488% | 9,600% |
| *CC – 2* | 6,622% | 6,564% | 6,137% |
| *CC – 3* | 7,931% | 7,420% | 8,178% |
| *CC – 4* | 8,352% | 7,601% | 7,714% |
| *CC - 5* | 9,403% | 9,500% | 9,340% |
| *ACER* | 8,310% | 8,115% | 8,194% |

**Table 5-43** $EER_M$ obtained on evaluation set (SNorm), comparing classical parameters with extra parameters and extended biometric parameters

| FEMALE | | SNorm | |
|---|---|---|---|
| SV-GMM | GIC | GDC+E+F0+F3 [RR] | GSE [RR] |
| CONFIG | *F*=40 *MFCC*=19 | *F*=50 *MFCC*=19 | *F*=50 *MFCC*=19 **Source-Tract Sep. Alg:** Prediction Order: 29 Forgetting Factor: 0.995 **GSE:** 36-Channel Filter bank 4 MFCC |
| *CC - 0* | 25,641% | 23,099% | 22,974% |
| *CC – 1* | 10,212% | 10,144% | 10,024% |
| *CC – 2* | 9,598% | 9,412% | 9,068% |
| *CC – 3* | 8,700% | 9,159% | 8,852% |
| *CC – 4* | 8,733% | 8,647% | 9,086% |
| *CC - 5* | 11,986% | 11,564% | 11,249% |
| *ACER* | 9,846% | 9,785% | 9,656% |

**Table 5-44** $EER_F$ obtained on evaluation set (SNorm), comparing classical parameters with extra parameters and extended biometric parameters

| FEMALE | | SNorm + ZTNorm | |
|---|---|---|---|
| SV-GMM | GIC | GDC+E+F0+F3 [RR] | GSE [RR] |
| CONFIG | *F*=40 *MFCC*=19 | *F*=50 *MFCC*=19 | *F*=50 *MFCC*=19 **Source-Tract Sep. Alg:** Prediction Order: 29 Forgetting Factor: 0.995 **GSE:** 17-Channel Filter bank 2 MFCC |
| *CC - 0* | 25,686% | 25,348% | 22,320% |
| *CC – 1* | 11,194% | 11,169% | 11,287% |
| *CC – 2* | 10,097% | 9,938% | 10,344% |
| *CC – 3* | 9,239% | 9,905% | 10,185% |
| *CC – 4* | 9,356% | 9,609% | 10,127% |
| *CC - 5* | 12,070% | 12,176% | 12,289% |
| *ACER* | 10,391% | 10,559% | 10,846% |

**Table 5-45** $EER_F$ obtained on evaluation set (SNorm+ZTNorm), comparing classical parameters with extra parameters and extended biometric parameters

From the results shown on the previous tables, we can draw the following conclusios. In first place, for both male and female speakers, we obtain unusual high error rates. This

is due to a key factor we have already presented. In the case of male speakers, there are up to 64 test files from which no information can be extracted, which affects almost 835 trials. Whereas in the case of female speaker this situation affects up to 171 files inviolving 2300 trials. All theses files have been modified to incorporate fake additive noise, resulting in recordings of the type represented in Figure 5-117, from which clearly no useful information can be extracted for speaker recognition purposes. Specifically, for almost 0.5% of male trials and 0.5% of female trials, no score can be produced and therefore no decision can be made.

In the case of male speakers, the results obtained by incorporating extended biometric parameters (no matter whether just F0 and F3 or F0, F3 and glottal source estimate MFCC) clearly outperformed the results obtained when only classical parameters are used. In the case of female speakers, combining the use of SNorm and ZTNorm provides lower recognition rates than in the case of just applying SNorm. In addition in this last case, the GDEB parameterization allows for a reduction in terms of EER for almost all the conditions.

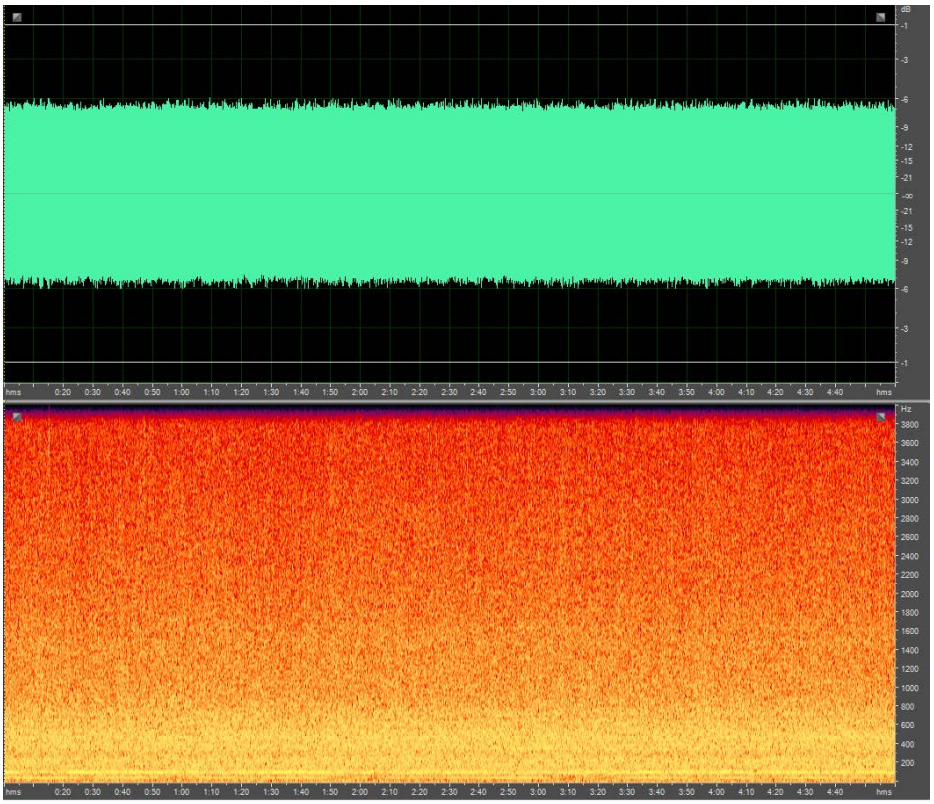


**Figure 5-117** NIST SRE12 files with fake additive noise (up - time domain, down frequency domain)

It must be noted that, despite facing an evaluation scenario that has little to do with the development scenario, the use of the gender dependent extended biometric parameterization allows for a reduction of the recognition errors in almost all of the conditions in which the evaluation set is divided.

The DET curves representing the results obtained with each of the previously presented configurations (see Table 5-43, Table 5-44 and Table 5-45) are depicted in Figure 5-118 for male speakers and in Figure 5-119 and Figure 5-120 for female speakers.

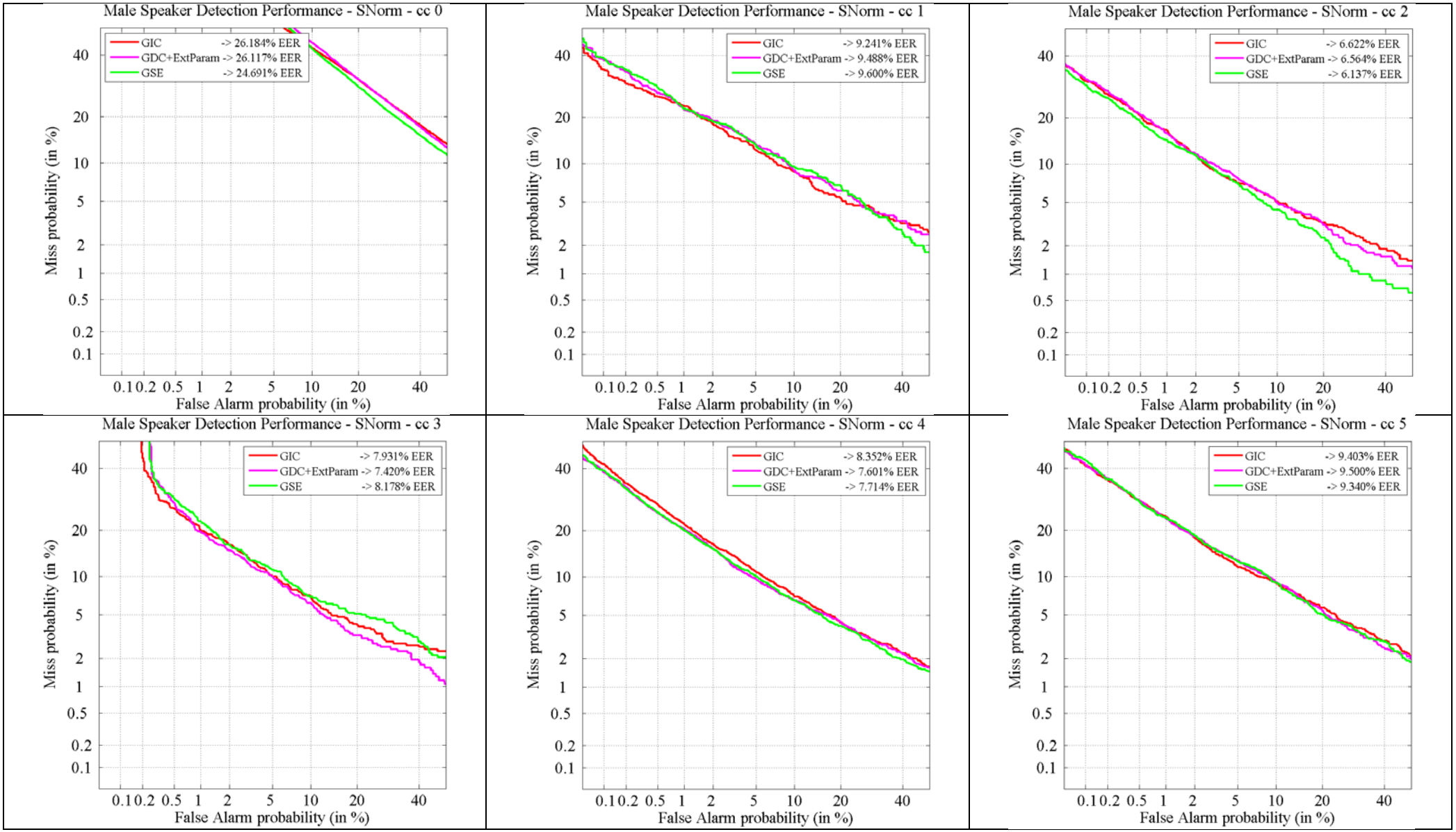



**Figure 5-118** DET curves comparing classical parameters and GDEB on the NIST SRE12 evaluation set for male speakers and SNorm

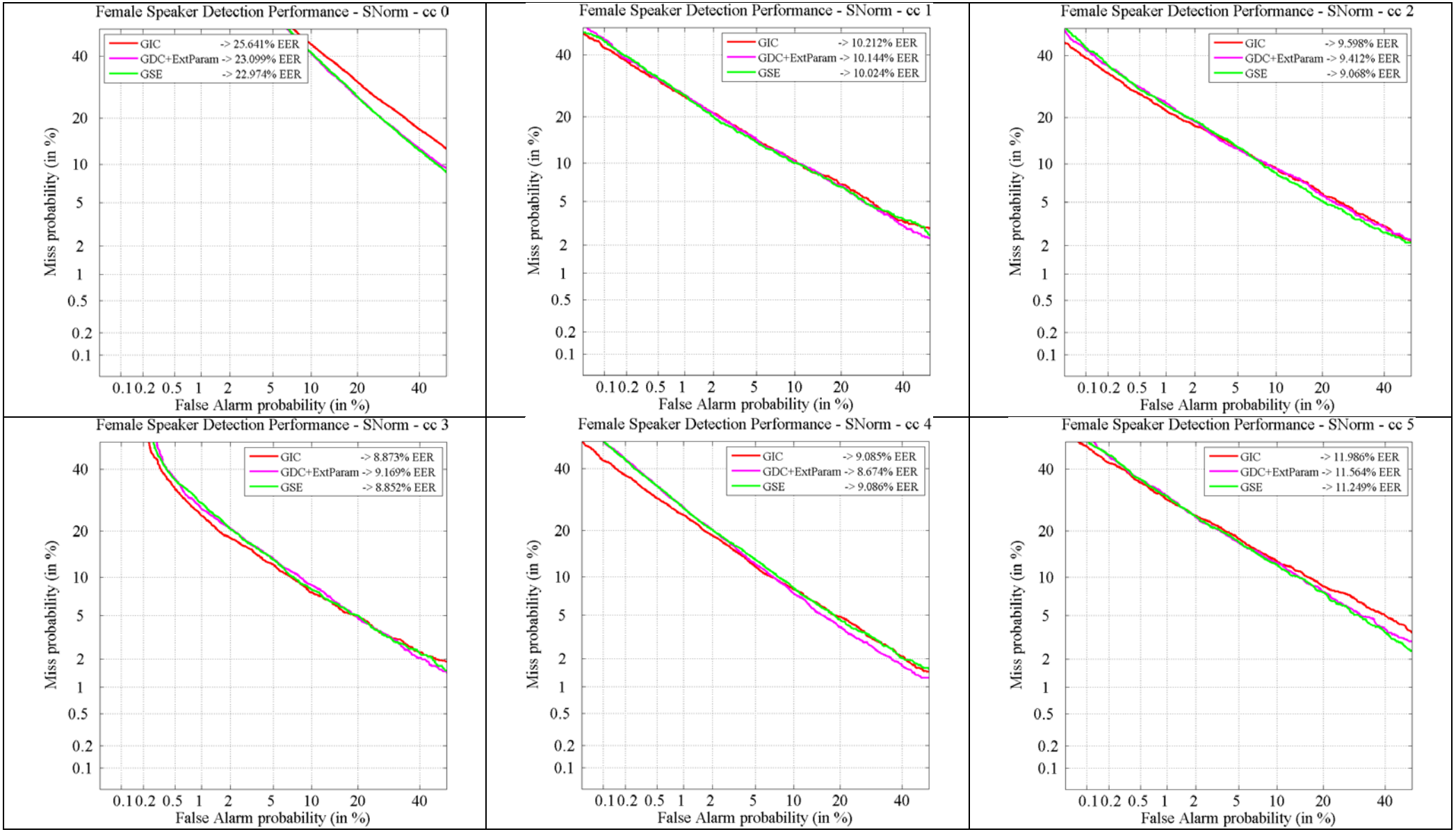



**Figure 5-119** DET curves comparing classical parameters and GDEB on the NIST SRE12 evaluation set for female speakers and SNorm

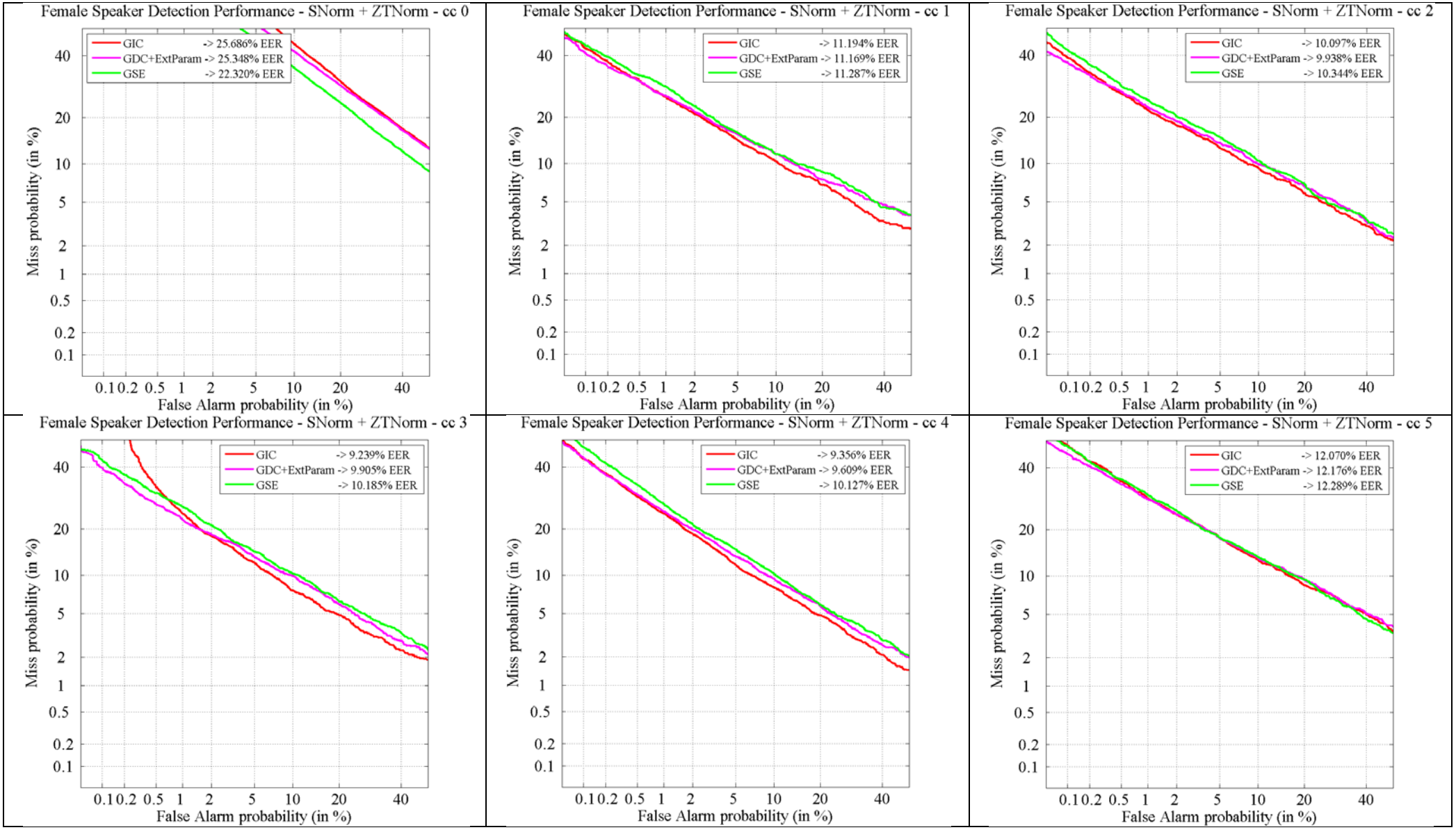



**Figure 5-120** DET curves comparing classical parameters and GDEB on the NIST SRE12 evaluation set for female speakers and SNorm+ZTNorm

#### *5.2.4.1 Brief Conclussions*

In this scenario, using the NIST data bases and somehow adjusting the scenario to the 2010 and 2012 evaluation plans, it is difficult to draw specific conclusions about the results. This is due to three key factors:

- Problem dimensionality: The amount of information to deal with both in the development set and in the evaluation set causes an unaffordable tuning process like the one proposed in the above scenarios.
- Inter- and intra-speaker variability: We are facing a problem with a large intra- and inter-speaker variability. In previous scenarios we have concluded that the optimal configuration will be different depending on the type of recordings that we must deal with (i.e. Microphone/Telephone). Therefore, it would be appropriate to perform a search for optimal configuration for each of the conditions in which the evaluation set and development set are divided. Otherwise we assume that a compromise must be reached in other to simultaneously minimise EER for all conditions. Therefore we will not obtain the best possible configuration for each condition.
- Development and Evaluation divergence: In previous scenarios, evaluation and development operation mode was the same, except that the system faces unknown data. However, in this scenario, evaluation plan is quite different from development plan. Therefore it is unlikely that the results obtained on the development set by a specific configuration may offer the same performance on the evaluation set.

Anyway, we can draw the following conclusions that confirm the results obtained in previous scenarios. The use of a gender-dependent extended biometric parameterisation in which GSE information has been incorporated provides a clear improvement in terms of recognition rates respect to the use of the classic gender-independent approach. The use of $\Delta\Delta$ coefficients is completely discouraged as the recognition rates obtained with configurations including these coefficients, are worse than in the case of using just MFCCs+$\Delta$. Additionally, it has been confirmed that the use of extra parameters under some specific combination helps to increase recognition rates. Specifically, F0 and F3 are useful for both male and female speakers.

Regarding the classification method, it is clear that in high variability scenarios like this one the use of the GMM-UBM approach it is not recommended both in terms of recognition rates and computational efficiency. Additionally, the SV-GMM approach provide better results than the *i*-vector approach with less computational load.

Last but not least, we have verified that even in such a variability scenario, including information from the GSE into the feature vectors, allows for an increase on the recognition rates that is consistent over the development and the evaluation set.

# 6 CONCLUSIONS

## 6.1 OVERVIEW

Since early 1970's Automatic Speaker Recognition has been an active research area, mainly due to the interest shown by different institutions in providing a common framework to both test the advances in the area as well as to recommend the fields to which major research efforts should be directed. Major advances in the area have been achieved thanks to improvements in classification methods. This is clearly observed in the range of classification methods that have been implemented to meet the challenges of these evaluations such as: UBM-GMM, total variability spaces, Gaussian supervectors, etc.; alone or combined with normalisation post-processing steps such as LDA, WCCN, or NAP.

However, despite been of great importance, the front-end that feeds those classification systems has been relegated to a second plane in terms of research interest. This is demonstrated by the fact that still most systems use gender-independent MFCC coefficients extracted from the power spectral density of speech as a whole. The present work defends the idea that an adequate front-end properly characterising the speakers, and not one that has its origin in the speech recognition area, is as important as the selection of the classification method.

The advances that have taken place in the area of digital signal processing, along with a better understanding of the phonation processes, have made it possible the development of software tools that allow us to perform a more detailed analysis of the voice signal. The presented methodology, developed through the use of inverse filtering, provides a way to extract from the voice signal the glottal source and vocal tract estimates, which conveniently parameterised allows a more accurate characterisation of speakers.

The thesis work that has been presented here will hopefully help to recover the interest in this subarea of biometric speaker recognition, i.e. speaker characterisation, as really promising recognition results are obtained regardless the multiple scenarios in which our proposal has been tested.

## 6.2 CONCLUSIONS

We must not forget that the main objective of this thesis is to improve the characterisation of speakers, so that by integrating biometric features, speaker recognition systems can be trustfully used in security environments, like other biometric systems such as those relying in fingerprints or iris biometrics. To achieve this goal, the thesis has been organised in five parts, the last two being experimental.

Chapter 1 provides a unique review on biometrics, as well as a unique synthesis of the current state-of-the-art of different biometric characteristics typically used. Besides, after the general overview on biometrics, we focus on the particularities of Speaker Recognition, from the feature extraction process, to decision making, through classification methods. After this analysis, some conclusions can be drawn.

- Although there are multiple applications that offer Speaker Recognition Solutions, this field still remains an open issue that allows for a wide margin for improvement.

- This is an active research area thanks to the interest shown by different institutions in providing a common framework to both test the advances in the area as well as to recommend the fields to which major research efforts should be directed.
- Main research efforts on the area have been directed to the improvement of classification methods.
- Two different types of parameters are usually used to characterise speakers on Speaker Recognition systems:
    - Parameters inherited from the speech processing area, typically MFCC and its derivatives.
    - High level parameters (suprasegmental), more related to behavioural aspects rather than physiological characteristics; and therefore easily replicable by impostors.

  Surprisingly all the approaches analysed face the problem with gender-independent parameters, despite being a well known fact that male and female voices show clear differences, due to non-negligible sexual dimorphism.
- As a result of this, we conclude that there is a need to improve speaker recognition systems, not from the classification point of view, but by providing a set of characteristics that accurately defines the speakers. The formulated hypothesis is twofold, first the proposed parameters must be gender-dependent, and second they must rely on biometric information, preferably standing on a physiological background.

Chapter 2 represents the core of the theoretical work on this thesis. In particular, a comprehensive review of speech production processes including its mathematical modelling has been presented. All this mathematical reasoning leads us to Gunnar Fant's voice production model, which in its simple form can be summarised in the existence of an excitation signal (glottal source) that is modelled by a filter (representing the vocal tract). Taking this model as starting point, a new glottal-source vs. vocal-tract separation algorithm based on the theory of inverse filtering via linear prediction, is proposed. This method is compared with other proposed solutions, showing differences not only in the separation process but in the way, the resulting components, are used.

Additionally, we present a complete parameterisation of the signals obtained. The resulting features, in contrast to classical MFCC coefficients estimated from the original voice signal, enclose biometric significance, as they establish a relationship between numerical values and physical characteristics of the speaker hardly forgeable. However, it was found that these features, despite their theoretical potential and proven validity in diverse areas (including speaker recognition in the NIST HASR[5] 2010 and 2012 evaluations) show an important limitation in their practical application to automatic speaker recognition. This limitation relies on the need to provide the algorithms with a high quality voice signal in order to obtain accurate estimations, which is not possible (without human supervision) in the various tests that have been proposed in Chapter 4 (mainly due to the reduced sampling frequency of the recordings, the noisy environment, the channel variability, etc.). For this reason, in Chapter 5, we present an

[5] NIST HASR (Human Assisted Speaker Recognition) focus on the way human experts effectively utilize automatic speaker recognition technology (http://www.nist.gov/itl/iad/mig/hasr.cfm)

alternative parameterisation of the glottal source and vocal tract estimates based on its frequency domain analysis, i.e. based on MFCC coefficients.

An important aspect of using the MFCC parameterisation of the glottal source and vocal tract estimates is the fact that the computational requirements (in processing time) to compute the new parameters are not drastically increased. The major effort relies on the component separation algorithm, since once the two new signals are obtained, the feature extraction process is somehow identical to the MFCC classical extraction, and what is more, it can be carried out simultaneously on all the involved signals (raw voice signal, glottal source estimate and vocal tract estimate).

Chapter 3 provides an extensive review of the classification methods widely used in speaker recognition, starting with those used at early stages of research in the area, such as Vector Quantisation and HMMs, moving into GMM-UBM, continuing with two-class classifiers (SVM), and finishing with new trends in classification methods applied to speaker recognition: supervectors and *i*-vectors. Additionally, some hints on fusion at different levels have been introduced. After analyzing the multiple options available to face the problem, some assumptions can be made:

- GMM-UBM approach continues to be the *de facto* reference method in text-independent speaker recognition.
- The use of more complex classification methods such as *i*-vectors or JFA requires for its proper operation a large amount of data for training the models, so the computational cost is drastically increased if compared with other approaches. Furthermore, it is not always possible to have such amount of training data for a particular system, since the data used must be as similar as possible to the actual training and test conditions.
- The latest trend in speaker recognition seems to be the fusion of multiple classifiers, at the scoring level, even though fused systems are generally based on the same front-end. In this sense we can conclude that it seems that it does not matter what the input to the system is, because at the end we will manage to get the correct score by combining scores, although lacking a solid biometric meaning. This approach sharply vary from our vision, as we think that it is more important to provide an accurate description of the speakers as the starting point in order that better recognition rates can be obtained regardless the classification method used.

Chapter 4 holds a dual purpose, on one hand we present in detail different databases widely used to test speaker recognition systems; on the other hand we describe a set of tests using the referred databases that allows us to verify that the speaker parameterisation proposed on this thesis provides better performance (in terms of recognition rates) than traditional parameterisation.

From the review of other works in this area, we have drawn some conclusions concerning testing and comparison.

- If we want to apply the speaker recognition system in security environments it seems inappropriate to pose text-dependent tests. As it is recommended to change any password code from time to time, also voice based keys must change. So text-dependent tests are beyond the scope of this thesis, and we have focused on text-constrained and text-independent environments.

- We have also verified that, despite using the same database, the performance of two systems cannot be compared straightforward, given that the selection of the training and test data can be different. Therefore, as we want to compare the recognition ability of the proposed parameterisation over classical approaches, we need to develop two recognition systems based on the same classifier but each fed by each front-end under study. Facing the same settings and conditions (enrolment and test) a straight comparison of systems in terms of recognition rates is possible.
- Last but not least, we face external/independent evaluations. It is important to let our system undergo external evaluations where the task definition and settings are beyond our control. This scenario which apparently seems to be useful when assessing our work, presents also a problem that substantially affects the recognition results. We refer to the fact that although the evaluations we have chosen (MOBIO and NIST) clearly establish the data to be used as well as the test conditions, they leave the possibility of using additional external data open to world modelling stages as well as to normalisation. This means that systems, actually, do not face the test under the same conditions. In the case of NIST SREs, we have checked that the normalisation stage is critical if competitive results are to be obtained, especially due to the high variability held by the databases. MOBIO does not remain unaware of this circumstance. Particularly, we have seen in the 2013 evaluation that an unfortunate selection of additional data for normalisation purposes can ruin systems' performance.

Regarding the ad-hoc tests that we have prepared on text-constrain and text-independent scenarios, they have been conceived to try to reflect real working environments, and it is with this premise that they have been designed and tested.

- In order to transfer this technology from a research workbench to real-life environments, final user acceptance is essential. In the case of speaker recognition systems, the time that must be lost in the enrolment phase seems to be the most annoying factor. Therefore, it is essential to limit the amount of training data (which actually means reduction of enrolment time) to build speaker models. In particular, for ALBAYZIN tests, enrolment has been limited to 12 seconds (3 files x 2 seconds/file) while in the case of HESPERIA this limit is set to 40 seconds (20 files x 2 seconds/file). In comparison, this enrolment times are significantly lower than the ones provided in NIST SRE10 core trials, set between 3 to 15 minutes.
- Regarding the number of different speakers, we face the limitation imposed by the used databases. Although it can be considered small and/or not representative (especially concerning research aspects), we can conclude that it is adequate for real-life scenarios. Specifically, statistics from 2011 show that 99% of companies in Spain have a number of employees below 50, which means that the proposed tests reflect the situation that will be present in most of the scenarios in which the technology is to be implemented.
- The division of the databases on different sets (development, evaluation and background), allows us to verify how the system behaves under new data/users, or whether we have over trained the system and therefore does not generalise in the expected way. However, this scenario is not entirely realistic, as for instance in security environments we usually configure the system to work on a closed-set configuration. So, during enrolment and test the same known sets of users are

available and what is new to the systems are the new claims of identity. This means that all the data available from all users can be used for training and normalisation.

Chapter 5 presents the results obtained in the set of tests presented in chapter 4. The most important conclusion that we can derive from the results is that the use of the proposed parameterisation, i.e. the gender-dependent parameterisation which complements classic MFCC features with extended biometric features obtained from the glottal source and vocal tract estimates, provides better results in term of recognition rates than the classical gender-independent approaches. Thus, confirming the hypothesis of the present work. Additionally, at the end of each of the presented scenarios a brief summary of the conclusions to be drawn from the results, especially concerning the set of parameters used, is performed. However these conclusions must be refined and qualified considering the different results we have obtained.

- Regarding the use of classical features for speaker characterisation, we have carried out multiple tests (especially where HESPERIA and ALBAYZIN databases are involved) and we can conclude that the use of $\Delta$ coefficients clearly improve the recognition rates whereas the use of the $\Delta\Delta$ coefficients seems not to be adequate to properly characterise the speakers. This behaviour has also been confirmed in the test carried out on NIST SRE10.

  Clearly, the use of $\Delta\Delta$ coefficients in the speaker recognition area is discouraged, as demonstrated by the results obtained.

- Regarding the gender parameterisation approach proposed as initial hypothesis of the present work, it has become clear through all the tests carried out, that when it comes to improve recognition rates of speaker recognition systems it is essential to process each gender independently. Moreover, according to some preliminary results on HESPERIA database, it seems that there is more phonation variability in the female speaker population than in male speakers, but at the same time it remains consistent among individuals within the same age range groups. This indicates that it may be appropriate not only to provide speaker recognition systems with a gender-dependent parameterisation, but also establish female subgroups according to age range.

  In this regard, it is worth noting that in most of the proposed scenarios, the optimal configuration usually entails a filter bank with higher number of channels for female speakers than for male speakers, on the filterbank used to compute the MFCCs.

- Concerning the use of extended biometric information which complements classical parameters to provide an accurate characterisation of speakers, it has been shown through the different tests proposed using multiple databases that it clearly improves systems performance in terms of higher recognition rates. Regarding the extended biometrics we can perform a separate analysis on what we have called alternative parameters, and on the set of parameters extracted from the glottal source and vocal tract estimates.

  - Alternative parameters: Under this name we have grouped the parameters related to energy of the frame (E), delta energy ($\Delta$E), pitch (F0) and formant 3 estimate (F3). Energy and $\Delta$E coefficients are typically used as they constitute a heritage from the speech recognition area, however we have seen that in most cases their used (when combined with classical

MFCCs) do not usually provide an improvement in terms of recognition rates. On the contrary, F0 and specially F3 (usually combined with some of the other alternative parameters) are extremely useful in most of the presented scenarios in other to precisely characterise speakers, and therefore to reduce recognition errors. Thus the use of F3 proposed in this thesis is confirmed as an important contribution, as this parameter seems to be very relevant for speaker recognition tasks.

- GSE and VTE: In all the tests carried out in this work, we have always proceeded by first incorporating features extracted from the glottal source estimate and once a valid configuration is selected, features from the isolated vocal tract estimate are added to the feature vector. This approach is based on two fundamental premises. First of all, we assume that information conveyed by the glottal source is not only closely related to physiological (and therefore biometric) characteristics of speakers, but also somewhat free from the influence of the message casted. For this reason it should be more effective in order to characterise speakers, especially in text-independent scenarios. This takes us to the second premise. As already said, the vocal tract estimate provides information dependent on the phonetic content of the message. Therefore the amount of training information needed to model the speaker is going to be larger than in the case of the glottal component as sufficient phonetic coverage will be required from the speaker. This requirement is not always possible to fulfil with the databases used in this thesis, and in many practical cases in the real world. However, although it is well known that the vocal tract component of voice is of great interest for the biometrical characterisation of the speaker, the results obtained for the different scenarios show that the use of parameters extracted from the vocal tract estimate, when included into the feature vectors, do not provide additional benefits to the recognition system compared to the case of using information from the glottal source estimate combined with MFCCs extracted from the raw voice signal.

Besides, the set of tests presented in this thesis allows us to verify that both the process and therefore the proposed extended biometric parameters are robust to channel variability. Specifically, the tests run on the HESPERIA database show how even when the training is performed on microphone recordings and test are run over telephone recordings the use of extended biometric parameters produced a reduction in error rates.

Additionally, it seems appropriate to provide a brief note on the results obtained in the NIST SRE10, as they are not as brilliant as the ones obtained in the other scenarios. Clearly, the results are influenced, among other factors, by the characteristics of the database and by the fact that no manual treatment or inspection of data is allowed according to the evaluation plan. As previously mentioned, databases provided by NIST enclose not high but huge inter- and intra-speaker variability due not only to changes in the speaker itself (recordings from the same speaker are made years apart) but also due to a really awful recording methodology in which for instance the speaker is at such a distance from the microphone that it hardly captures its voice, or recordings in which the environmental noise is so strong that it completely masks speaker's voice. Taking into account this scenario, it is difficult to tune a system that needs, for

accurate operation quality recordings, or at least recordings where just one speaker is present. Moreover, we have been quite ambitious trying to find just a single configuration which improves recognition-rate overall trial conditions. It would have been more appropriate to face each condition independently. Nonetheless, the presented results are very competitive and allow us to demonstrate even in this complex scenario that a gender-dependent parameterisation is essential, and that the use of features extracted from the glottal source estimate reduces misidentification.

- Regarding the computational requirements for computing the new extended biometric parameters, we can conclude that the front-end designed in this thesis does not lead to a computational overload much larger than a classical front-end incorporating $\Delta$ and $\Delta\Delta$ coefficients. Although a detailed analysis is beyond the scope of the present work, we can perform a brief comparison of both front-ends.

  We can assume that the feature extraction process proposed here consists of two different stages. The first one (which obviously is not needed in a classic front-end) is the source-tract separation. The time devoted to this stage will depend on the size of the voice input signal and the order of the filter used. Once the three signals (raw voice, glottal source estimate and vocal tract estimate) are available, the next stage consists in extracting the MFCC coefficients form each of the signals. This process is independent for each signal and therefore can be performed simultaneously. Thus at this stage, the additional computational cost is caused by larger storage requirements, specifically it is going to be three times larger as we are simultaneously working with 3 different signals instead of just one. Finally, it must be noted that in the case of working with a classic front-end a third stage must be added for the computation of the $\Delta\Delta$ coefficients, which are not required in our proposal, and evidently involves additional time costs.

  However, regarding computational costs it must be noted that the bottleneck of speaker recognition systems is found not in the front-end, but in the classification and score stage. These stages are influenced by the size of the feature vectors previously generated. In our proposal, despite introducing extended biometric parameters in the feature vector, the number of parameters is less than in the case of introducing $\Delta\Delta$ coefficients, so the total processing time will be reduced by using the gender-dependent extended biometric parameterisation with respect to using MFCC+$\Delta$+$\Delta\Delta$.

- Regarding classification methods, it is necessary to point out that the aim of this work was neither proposing any new classification method nor determining which one of the most used methods provides better results. Nevertheless, different classification methods have been used depending on the specific scenarios that have been proposed.

  From the presented results two main conclusions can be drawn. First of all, regardless the classification method used, the gender-dependent extended biometric parameterisation provides better recognition rates than classical parameterisation. Second, the selection of a specific classification and scoring method must be data-dependent. For instance, using the GMM/UBM approach in the NIST SRE is going to be more time consuming and will provide worse results than for instance using the Gaussian supervector approach. By contrast, using the Gaussian supervector or the *i*-vector approach in the MOBIO,

ALBAYZIN or HESPERIA scenario will not be adequate as not enough data will be available for normalisation or transformation matrix estimation so results achieved with the GMM-UBM paradigm will outperform those obtained using the alternative approaches.

- Concerning the use of normalisation methods, it is difficult to draw a clear conclusion on whether their use is going to be more efficient or not.

  In order for the normalisation techniques to be effective, it is necessary to have a set of impostors available, really covering the whole set of possible alternatives to target speakers, under the tested conditions. As previously mentioned, in the set of tests carried out using HESPERIA, ALBAYZIN and MOBIO databases, the number of speakers marked as impostors as well as the number of recordings available for each of them is relatively small, this not being enough for the normalisation technique to work properly. It must be noted that under certain conditions, the used of score normalisation techniques can improve recognition rates; and their use is essential for instance in NIST SREs. However, the application of normalisation methods will depend on the specific scenario and the information available for cohort modelling.

## 6.3 FUTURE WORK

As we have seen throughout the extension of this thesis, the use of gender-dependent biometric parameters, extracted from the deconstruction of voice in their estimates of glottal source and vocal tract has proved being useful in the area of speaker recognition. Additionally, the use of these estimates, conveniently parameterised, is also useful for gender detection and age classification. In parallel, it has also been shown that alternative parameterisations of these estimates are useful in the evaluation of disease [Gómez,2009] and the analysis of vocal quality.

However, this thesis is not the end; it is merely the end of the beginning. Some questions still remain open as well as several lines of research have emerged, that are described below:

- **Quality Measures**

  Throughout this work we have faced different scenarios involving different types of transmission channels, different emotional states of speakers, and different recording scenarios. This has led to the need to slightly modify the separation algorithm of the vocal tract and glottal source components of voice to deal with these circumstances. That is why we refer to these components as estimates of source and tract instead of just glottal source and vocal tract (see section 5.1.2). However, given that this variability is going to be present in one way or another in any real-life speaker recognition system, it seems necessary to define a quality measure that indicates the feasibility not only of the separation process but of the biometric feature extraction as well. This quality measure will constitute an objective criterion to determine whether the use of the proposed features is going to provide a benefit or not.

- **Fusion**

  At the beginning of this research the decision was made to incorporate in a single feature vector both classical and biometric parameters. However, it may be interesting to study other methods of integration of this information based on the fusion of complementary systems; one based on classical parameters and

another based on biometrics (see section 3.6). Furthermore, this fusion could be driven by the quality measures that result from the study raised in the previous point.

- **Robustness against speaker temporal variations**

  This is another key aspect affecting any speaker recognition system. Not only do we mean changes due to emotional states (joy, elation, sadness, anger, etc.) or temporal health issues, but also the changes that occur in the voice as a direct result of aging or neurological deterioration.

  Consider the case where we have a speaker model obtained in an instant $t_0$, and we want to determine whether a recording captured years later corresponds to the same speaker. It is easy to imagine that there is going to be a discrepancy between the speaker model and the actual biometric information. Despite being a scenario perhaps unusual in speaker recognition systems, it appears to have some interest [Lanitis,2010] as demonstrated by the fact that this type of case is cited in the NIST evaluations. For this reason it seems appropriate to carry out an analysis of the effect of aging in the biometric features proposed in the present study.

  The biggest issue that we have to face in this case is the difficulty to obtain data reflecting this scenario, since most of the available databases, except a few cases in the NIST database, do not contemplate it.

- **Monozygotic and dizygotic twins, siblings and unrelated speakers**

  The feature set presented in this work to improve recognition rates intend to somehow model physical characteristics of speakers. Usually two unrelated speakers will have different physical characteristics thus leading to different parameter's value. The question that arises now is: what happens when there is a relationship between two speakers? i.e., what happens in the case of siblings, monozygotic and dizygotic and twins? The latter case is attracting great interest as demonstrated by the research activity in this area [Van Lierde,2005], [San Segundo,2013]. Therefore, it seems interesting to evaluate the robustness of our proposal in scenarios like this, as some other biometric authentication methods such as face seem to fail. Again, the biggest problem that we face is the lack of appropriate databases to conduct this analysis.

- **Goats and Lambs**

  [Doddington,1998] classifies users of a biometric recognition system into four different groups: Sheep, Goats, Lambs and Wolves (see Chapter 1.4). In security environments, where the set of users with access is known, closed and usually fixed, it would make sense to devote an extra effort in the normalization of goat (the users who are particularly difficult to recognise due to its high intra-class variation) and lamb (users particularly easy to imitate ) type speakers.

- **Feature selection**

  We have seen how the integration of biometric features in speaker recognition systems allow for a substantial improvement of recognition rates. We have also seen that it is necessary to perform a calibration step of the system depending on the type of recordings (i.e., the database, text-dependent vs. text-independent scenarios, etc.), in order to establish the most appropriate combination of parameters. This step seems to be necessary even in the case of using

exclusively classical parameters. Therefore, the study of a method to find the adequate combination of parameters as independently as possible of these sources of variability remains an open question.

- **Feature extraction**

  In this work we have chosen to use a parameterisation of the vocal tract and glottal source estimates in the form of MFCCs. As discussed in section 2.4.1.3, we have ruled out the use of another set of parameters, which in theory should provide better results for different reasons.

  One reason is that the alternative set of parameters defined in section 2.4.1.3 requires high quality sound recordings to be estimated. However, high quality is neither usually present in any of the databases used (except HESPERIA), nor in many real application scenarios. Therefore, it could be of great interest to carry out a study to obtain such parameters from lower quality recordings and verify their validity for automatic speaker recognition.

  Another reason is the additional problem that may appear when using parameters which convey specific semantics. Specifically, the expected benefit may be neutralized by two key factors: estimation error and estimation at unsuitable instants. In this sense, an additional problem arises regarding the value that must be assigned to these semantic parameters at the specific instant when their estimation is not adequate. Should we assign the last value? Maybe zero? Or maybe the mean value over some specific previous interval? For this reason, it may be useful to conduct a study on the effect of non-periodic parameters on the whole speaker recognition system.

- **Calibration**

  Throughout the present work, we have used the EER as optimisation objective in order to compare performance of different systems. This choice is justified not only because this value constitutes a summary of the whole DET-curve and therefore of the system behaviour at different operating points, but because it also constitutes an upper limit of the system's capacity decision. However, in the NIST SRE (which is the frame reference for speaker recognition) the DCF and minDCF are used instead as optimisation parameters. Therefore, it seems appropriate to focus our efforts on this direction (maybe using the BOSARIS toolkit).

As we have said on different occasions, the front-end of a speaker recognition system seems to have been relegated to a secondary plane when compared to the research interest in classification and normalisation methods. Most speaker recognition systems keep on using MFCCs extracted from the power spectral density of speech as a whole; although these coefficients are inherited from the speech recognition area which although offering good performance are not as accurate as expected when characterizing the speaker. For this reason, once the high level of maturity in terms of classification techniques has been reached, it seems necessary to focus on the extraction process in order to improve the performance of these systems, selecting features that allow for an unambiguous way to characterise speakers. From this point of view, this work offers a new starting point for further research in the speaker recognition area.

## 6.4 CONTRIBUTIONS

Since in the speaker recognition area there are some common frameworks or evaluations in which the state-of-the-art of the technology can be measured, it seems necessary to participate in them so that we can assess what is the real value of our contribution to this area. From this point of view, we can cite as one of the contributions of this thesis, the participation in international Speaker Recognition Evaluations, such as NIST SRE 2010, NIST SRE 2012, and MOBIO SRE 2013. The results obtained in these evaluations are dissimilar. In the case of the NIST SRE 2010, it was our first attempt to build a complete SR system which incorporates a gender-dependent biometric parameterisation and the results were somehow disappointing. However after a post-processing step we can conclude that the results were influenced not by a bad choice of the parameters but by the particularities and peculiarities of the NIST databases which we were unable to cope with, on due time. As presented in Chapter 5, once we learned how to deal with the multiple particularities of these databases, we were able to reach state-of-the-art recognition rates in terms of EER. As a counter example, we can cite the MOBIO SRE 2013, in which we managed to get the best recognition rates as reported by organisers and presented also in Chapter 5.

From a less practical, but more scientific point of view, several papers have been published and many congress contributions have been made in the course of the research leading to this thesis. These contributions are listed by date from most to less recent, except for the first three which deserve special mention for being published in journals with impact factor.

- **JCR Papers:**
  - Gómez, P., Rodellar, V., Nieto, V., Muñoz, C., **Mazaira, L.M.**, et al; *"Characterizing Neurological Disease from Voice Quality Biomechanical Analysis"*, Cognitive Computing, Vol.5, Issue 4, pp.399-425, 2013 (ISSN: 1866-9956) – ***JCR: 0.867***
  - Gómez, P., Ferrández, J.M., Rodellar, V., Álvarez, A., **Mazaira, L.M.**, Martínez, R., Muñoz, C.; "Neuromorphic Detection of Speech Dynamics", Neurocomputing, Vol. 74, Issue 8, pp. 1191-1201, 2011 (ISBN: 978-989-8425-42-3) – ***JCR: 1.44***
  - Gómez, P., Fernández-Baíllo, R., Rodellar, V., Nieto, V., Álvarez, A., **Mazaira, L. M.**, Martínez, R, Godino, J. I.; *"Glottal Source Biometrical Signature for Voice Pathology Detection"*, Speech Communication, Vol. 51, pp.759-781,2009 (ISSN: 0167-6393) – ***JCR: 1.229***
  - Gómez, P., Ferrández, J. M., Rodellar, V., Álvarez, A., **Mazaira, L.M**.; *"A Bio-inspired Architecture for Cognitive Audio"*, Lecture Notes on Computer Science, LNCS: IWINAC 07 (Springer Verlag), Vol. 4527, pp.132-142, Jun. 2007. (ISBN: 3-540-73052-4) - ***JCR: 0,415***
- **Papers**:
  - Gómez, P., Belmonte, E., Rodellar, V., Nieto, V., Álvarez, A., **Mazaira, L. M.**; "Biomechanical Evaluation of the Singing Voice", Proc. of MAVEBA 2013, pp. 137-140, Firenze University Press, Florence, Italy, 2013 (ISBN: 978-88-6655-469-1)
  - **Mazaira, L.M.**, Álvarez, A., Gómez,P., Martinez, R., Muñoz, C.; "The GIAPSI System for the 2013 Speaker Recognition Evaluation in Mobile Environments", In Actas de las VII Jornadas de Reconocimiento Biométrico de Personas, Zamora, Spain, Sep. 2013.(ISBN: 84-616-5690-3)

- Gómez, P, Nieto, V., Rodellar, V., Martinez, R., Muñoz, C., Álvarez, A., **Mazaira, L.M**, Scola, B., Poletti, D., "Wavelet Description of the Glottal Gap", In Procs. of the 18th Int. Conf. on Digital Signal Processing, Santorini, Greece, Jul. 2013. (ISBN: 978-1-4673-5807-1/13)
- Gómez, P, Nieto, V., Rodellar, V., Álvarez, A., **Mazaira, L.M.**, Martinez, R., Muñoz, C., Fernández, M., Ramirez, C., *"Estimating Tremor in Vocal Fold Biomechanics for Neurological Disease Characterisation"*, In The Procs. of the 18th Int. Conf. on Digital Signal Processing, Santorini, Greece, Jul. 2013, (ISBN: 978-1-4673-5807-1/13)
- Khoury, E., **Mazaira, L.M.**, et al. *"The 2013 Speaker Recognition Evaluation in Mobile environments"*, In the Procs. of the 6th IAPR International Conference on Biometrics, Madrid, Spain, Jun. 2013. (ISBN: 978-1-4799-0310-8)
- Gómez, P, Martinez de Arellano, A., Nieto, V., Rodellar, V., Álvarez, A., **Mazaira, L.M.**; *"Monitoring Treatment of Vocal FoldParalysis byBiomechanical Analysis of Voice"*, In Proc of I Jornadas Multidisciplinares de Usuarios de la Voz, el Habla y el Canto (JVHC2013), Las Palmas de Gran Canaria, Spain, Jun. 2013. (ISBN: 84-695-8101-5)
- Gómez, P, Belmonte, E., Nieto, V., Rodellar, V., Álvarez, A., **Mazaira, L.M.**;*"Vocal Fold Biomechanical Analysis for the Singing Voice"* In Proc of I Jornadas Multidisciplinares de Usuarios de la Voz, el Habla y el Canto (JVHC2013), Las Palmas de Gran Canaria, Spain, Jun. 2013 (ISBN: 84-695-8101-5)
- **Mazaira, L.M.**, Álvarez, A., Gómez, P., Martinez, R., Muñoz, C., *"Classical vs. Biometric Features in the 2013 Speaker Recognition Evaluation in Mobile Environments"* In Proc of I Jornadas Multidisciplinares de Usuarios de la Voz, el Habla y el Canto (JVHC2013), Las Palmas de Gran Canaria, Spain, Jun. 2013. (ISBN: 84-695-8101-5)
- Muñoz, C., Martinez, R., Gómez, P., Álvarez, A., **Mazaira, L.M.**, *"Gender Detection in Running Speech from Glottal and Vocal Tract Correlates"*, In Lecture Notes on Computer Science, LNCS: Advances in Nonlinear Speech Processing (Springer Verlag), pp.25-32, vol. 7911, 2011 (ISBN: 978-3-642-38846-7)
- **Mazaira L.M.**, Álvarez, A., Gómez, P., Martinez, R., Muñoz, C.; *"Improving Speaker Recognition Rates using alternative gender-dependent MFCC"*, In Proceedings of the VI Jornadas de Reconocimiento Biométrico de Personas JRBP12, pp. 207-216, Las Palmas de Gran Canaria, Spain, Jan. 2012. (ISBN: 978-84-695-0695-0)
- Gómez, P., **Mazaira, L.M.**, Muñoz, C., Martínez, R., Álvarez, A., Rodellar, V.; *"The vocal Passport: its nature and role in forensic sciences"*, In Proceedings of the VI Jornadas de Reconocimiento Biométrico de Personas JRBP12, pp. 3-16, Las Palmas de Gran Canaria, Spain, Jan. 2012. (ISBN: 978-84-695-0695-0)
- Gómez, P., Martínez, R., **Mazaira, L.M.**, Rodellar, V., Muñoz, C., Álvarez, A., Hierro, J.A., Nieto, R.; *"Distance Metric in Forensic Voice Evidence Evaluation using Dysphonia-relevant Features"*, In Proceedings of the VI Jornadas de Reconocimiento Biométrico de

Personas JRBP12, pp. 169-178, Las Palmas de Gran Canaria, Spain, Jan. 2012. (ISBN: 978-84-695-0695-0)

- Muñoz, C., Martínez, R., Gómez, P., Álvarez, A., **Mazaira, L.M.**, Nieto, V.; *"Análisis de parámetros para clasificación de locutores adultos según su género y edad"*, In Proceedings of the VI Jornadas de Reconocimiento Biométrico de Personas JRBP12, pp. 197-206, Las Palmas de Gran Canaria, Spain, Jan. 2012. (ISBN: 978-84-695-0695-0)
- Gómez, P., Ferrández, J.M., Rodellar, V., Muñoz, C., Martínez, R., Álvarez, A., **Mazaira; L.M.**; *"Bio-inspired Phonologic Processing: From Vowel Representation Spaces to Categories"*, In Lecture Notes on Computer Science: NOLISP 2011(Springer Verlag), pp. 119-126, Nov. 2011. (ISBN: 978-3-642-25019-4)
- Gómez, P., Rodellar, V., Nieto, V., Muñoz, C., **Mazaira, L.M.**, Ramirez, C., Fernández, M., Toribio, E.; *"Neurological Disease Detection and Monitoring from Voice Production"*, In Lecture Notes on Computer Science: NOLISP 2011(Springer Verlag), pp. 1-8, Nov. 2011. (ISBN: 978-3-642-25019-4)
- Gómez, P., Rodellar, V, Muñoz, C., Martínez, R., **Mazaira, L.M.**, Álvarez, A.; *"Glottal parameter estimation by Wavelet transform for voice biometrics"*, In Proc. Of the 2011 IEEE International Carnahan Conference on Security Technology (ICCST), IEEE, pp.1-8, Oct. 2011. (ISBN: 978-1-4577-0902-9)
- Gómez, P., Rodellar, V, Nieto, V., **Mazaira, L.M.**, Muñoz, C., Fernández, M., Toribio, E.; *"Voice Quality Analisys to Detect Neurological Diseases"*, In Proceedings of the 7th Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2011), pp. 79-82, Firenze, Italy, Aug. 2011. (ISBN: 978-88-6655-009-9)
- Gómez, P., Fernández-Baíllo, R, Ferrández, J.M., Rodellar, V., Álvarez, A., **Mazaira, L.M.**, Martínez, R., Muñoz, C.; *"Monitoring Neurological Disease in Phonation"*, In Lecture Notes on Computer Science, LNCS: IWINAC2011 (Springer Verlag), pp. 136-149, May. 2011. (ISBN: 978-3-642-21325-0)
- Gómez, P., Ferrandez, J.M., Rodellar, V., Álvarez, A., **Mazaira, L.M.,** Martinez, R., Muñoz, C., *"Neuromorphic Detection of Vowel Representation Spaces"*, In Lecture Notes on Computer Science, LNCS: IWINAC2011 (Springer Verlag), pp. 1-11, May. 2011. (ISBN: 978-3-642-21325-0)
- Muñoz, C., Martinez, R., Gómez, P., Lang, E.W., Álvarez, A., **Mazaira, L.M.**, Nieto, V., "KPCA vs PCA Study for an Age Classification of Speakers", In Lecture Notes on Computer Science, LNCS: Advances in Nonlinear Speech Processing (Springer Verlag), pp.190-198, vol. 7015, 2011 (ISBN: 978-3-642-25019-4)
- Gómez, P., Fernández-Baíllo, R., Nieto, V., Rodellar, V., **Mazaria, L.M.**, Muñoz, C., Álvarez, A., Martínez, R *"Vowel-Consonant Speech Segmentation by Neuromorphic Units"*, Biology, Computation and Linguistics (Frontiers in Artificial Intelligence and Applications), Vol. 228, pp. 180-199, IOS PRESS, Jan. 2011. (ISBN: 978-1-60750-761-1)
- Gómez, P., Fernández-Baíllo, R., Nieto, V., Rodellar, V., **Mazaria, L.M.**, Muñoz, C., Álvarez, A., Martínez, R.; *"Dispersion of Vocal-Fold*

*Biomechanical-Parameter Estimates"*, In Proc. Of FALA 2010, pp. 209-212, Vigo, Spain, Nov. 2010. (ISBN: 978-84-8158-510-0)

- **Mazaira, L.M.**, Álvarez, A., Gómez, P., Martinez, R., Muñoz, C.; *"Glottal Source Cepstrum Coefficients Applied to NIST SRE 2010"*, In Proceedings of the V Jornadas de Reconocimiento Biométrico de Personas JRBP10, pp. 223-232, Zaragoza, Spain, Sep. 2010.
- **Mazaira, L.M**., Álvarez, A., Gómez, P., Martinez, R., Muñoz, C.; *"The GIAPSI NIST 2010 Speaker Recognition Evaluation System"*, In 1er WTM-IP Workshop en Multibiométricas para la Identificación de Personas, pp.40-46, Las Palmas de Gran Canaria, Spain,Jun. 2010. (ISBN: 978-84-693-3389-1)
- Gómez, P., Álvarez, A., **Mazaira, L.M**., Fernández-Baillo, R., Rodellar, V., Nieto, V.; *"Glottal Biometric Features: Are Pathological Voice Studies Applicable to Voice Biometry?"*, In 1er WTM-IP Workshop en Multibiométricas para la Identificación de Personas, pp.34-39, Las Palmas de Gran Canaria, Spain,Jun. 2010. (ISBN: 978-84-693-3389-1)
- Gómez, P., Ferrández, J.M., Rodellar, V., **Mazaira, L.M**., Muñoz, C.; *"Modelling Short-Time Parsing of Speech Features in Neocortical Structures"* In Lecture Notes on Artificial Intelligence (Springer Verlag), Vol. 6098, pp.159-168, 2010. (ISBN: 978-3-642-13032-8)
- Gómez, P., Ferrández, J.M., Rodellar, V., Fernández, R., Álvarez, A., Martínez, R., Nieto, V., **Mazaira, L.M.**, Muñoz, C.; *"Neuromorphic Speech Processing: Objectives and Methods"*, In Machine Audition: Principles, Algorithms and Systems, pp.447-473, Hershey, New York, USA, 2010, (ISBN: 978-1-61520-919-4)
- Gómez, P., Ferrández, J.M., Rodellar, V., Álvarez, A., **Mazaira, L.M.**, Martínez, R., Muñoz, C.; *"Detection of Speech Dynamics by Neuromorphic Units"*, In Lecture Notes in Computer Science (Springer Verlag), Vol. 5602, pp.67-78, Berlin. Jun. 2009 (ISBN: 3-642-02266-9)
- Gómez, P., Ferrández, J. M., Rodellar, V., Martínez, R., Muñoz, C., Álvarez, A., **Mazaira, L. M**.; *"Bio-inspired Dynamic Formant Tracking for Phonetic Labelling"*, In Actas de las V Jornadas en Tecnología del Habla, pp. 33-36, Bilbao,Spain, 2008 (ISBN: 978-84-9860-169-5)
- Gómez, P., Fernández, R., .Álvarez, A., **Mazaira, L.M.**, Martínez, R., Rodellar, V.; *"Speaker's Gender Detection from Glottal Biometry"*, In Proceedings of the IV Jornadas de Reconocimiento Biométrico de Personas JRBP08, pp. 113-122, Valladolid, Spain., 2008. (ISBN: 978-84-691-5008-5)
- Gómez, P., Fernández, R., Álvarez, A., **Mazaira, L.M.**, Rodellar, V., Martínez, R., Muñoz, C.; *"A Hybrid Parameterisation Technique for Speaker Identification"*, In Proceedings of the EUSIPCO 2008 – paper 1569104632, Laussane, Switzerland, 2008. (ISSN 2219-5491)
- Gómez, P., Fernández, R., Álvarez, A., **Mazaira, L.M.**, Rodellar, V., Martínez, R., Muñoz, C.; *"Glottal-Source Spectral Biometry for Voice Characterisation"*, In Proceedings of the EUSIPCO 2008 – paper 1569104624, Laussane, Switzerland, 2008**.** (ISSN 2219-5491)
- Gómez, P., Álvarez, A., **Mazaira, L.M**., Fernández, R., Nieto, V., Martínez, R., Muñoz, C., Rodellar, V.; *"Bio-inspired broad-class phonetic labelling"*, In Proceedings of the 1st IAPR Workshop on

Cognitive Information Processing (The International Association for Pattern Recognition), pp. 221-226, Athens, Greece, 2008.

- Gómez, P., Ferrández, J. M., Rodellar, V., Álvarez, A., **Mazaira, L.M.**; *"Articulatory Feature Detection based on Cognitive Speech Perception"*, Language Design, Vol.1, pp.119-126, 2008 (ISSN: 1139-4218)
- Gómez, P., Álvarez, A., **Mazaira, L.M.,** Fernández, R., Nieto, V., Martínez, R., Muñoz, C., Rodellar, V.; *"Decoupling Vocal Tract from Glottal Source Estimates in Speaker's Identification"*, Language Design, Vol.1, pp.111-118, 2008 (ISSN: 1139-4218)
- Gómez, P., Fernández-Baíllo, R., Martínez,R., Muñoz, C., **Mazaira, L.M.**, Álvarez, A., Godino, J.I.; *"Detecting Pathology in the Glottal Spectral Signature of Female Voice"*, In Proceedings of the 5th Int. Workshop on Models and Analysis of Vocal Fold Emissions for Biomedical Applications (MAVEBA07), pp.183-186, Firenze, Italy, Dec. 2007. (ISBN: 978-88-8453-673-3)
- Gómez, P., Álvarez, A., **Mazaira, L.M**., Fernández, R., Rodellar, V., Martínez, R., Muñoz, C.; *"Estimating the Dispersion of the Biometric Glottal Signature in Continuous Speech"*, Lecture Notes in Artificial Intelligence (Springer-Verlag), Vol. 4885, pp-255-262, 2007 (ISBN: 978-3-540-77346-7)
- Gómez, P., Fernández-Baillo, R., Álvarez, A., **Mazaira, L.M**., Rodellar, V., Martínez, R., Muñoz, C., Godino, J. I.; *"Glottal-Source Spectral Biometry for Voice Characterisation"*, 1er WTAC-ASAF Workshop en Tecnologías de Audio Cognitivo para Aplicaciones en Seguridad y Acústica Forense, pp.11-18, Las Palmas de Gran Canaria, Spain, Jun. 2007 (ISBN: 978-84-690-9175-3)
- Gómez, P., Álvarez, A., **Mazaira, L.M**., Fernández, R., Nieto, V., Martínez, R., Muñoz, C., Rodellar, V.; *"A Hybrid Parameterisation Technique for Speaker Identification"*, 1er WTAC-ASAF Workshop en Tecnologías de Audio Cognitivo para Aplicaciones en Seguridad y Acústica Forense, pp.19-22, Las Palmas de Gran Canaria, Spain, Jun. 2007 (ISBN: 978-84-690-9175-3)
- Gómez, P., Fernández, R., Rodellar, V. **Mazaira, L. M**., Martínez, R., Álvarez, A., Godino, J. I.; *"Biometry of Voice based on the Glottal-Source Spectral Profile"*, Proc. of the SAFE07: Workshop on Signal Processing Applications for public Security and Forensics, pp.363-366, Washington D.C., USA, Apr. 2007 (ISBN: 1-4244-1226-9).
- Gómez, P., Fernández, R., Rodellar, V., **Mazaira, L.M**., Martínez, R., Álvarez, A., Godino, J. I.; *"Biometría de la Voz basada en la huella espectral de la Onda Glótica";* Proc. de las III Jornadas Nacionales de Reconocimiento Biométrico de las Personas, pp. 163-175, Sevilla, Spain, Nov 2006 (84-934865-1-5).
- Gómez, P., Fernández, R., Álvarez, A., **Mazaira, L.M.**, Rodellar, V., Godino, J. I.; *"Descripción Biométrica de la Voz a partir de la Estructura Biomecánica de la Cuerda Vocal"*; In Proc. de las Jornadas en Tecnologías del Habla 06, pp.337-342, Zaragoza, Spain, Nov 2006 (ISBN: 84-96214-82-6)
- Gómez, P., Fernández, R., Godino, J. I., **Mazaira, L.M**.; *"Estudio de la Patología Vocal basado en la estimación del Correlato de la Onda*

*Mucosa de los Pliegues Vocales"*; In Proc. del Congreso Anual de la Sociedad Española de Ingeniería Biomédica CASEIB'06, pp523-526, Pamplona, Spain, Nov 2006. (ISBN: 84-9769-160-1).

- **Conferences**
  - **Mazaira, L.M.,** Álvarez, A., Gómez, P.; *"The GIAPSI NIST 2012 Speaker Recognition Evaluation System"*, NIST 2012 Speaker Recognition Evaluation Workshop, Orlando, USA, Dec. 2012.
  - **Mazaira, L.M.,** Álvarez, A., Gómez, P., Martínez. R., Muñoz, C.; *"The GIAPSI NIST 2010 Speaker Recognition Evaluation System"*, NIST 2010 Speaker Recognition Evaluation Workshop – Odyssey's Satellite, **Brno, Czech Republic, Jun 2010**
  - Gómez, P., Fernández, R., Martínez, R., Muñoz, C., Álvarez, A., **Mazaira, L.M**., *"Speaker's Gender Identification based on a Biometric Study of Glottal Dynamics"*, VII Pan European Voice Conference, Groningen, The Netherlands, Aug.2007
  - Fernández, R., Gómez, P., Fernández, F.J., **Mazaira, L.M.**; *"Clinical Description of the Voice Based on the Mucosal Wave Correlate"*, 27th World Congress of the International Association of Logopedics and Phoniatrics, Dinmark, Aug. 2007.
  - Gómez, P., Fernández-Baillo, R., Álvarez, A., **Mazaira, L.M.,** Rodellar, V., Martínez, R., Muñoz, C., Godino, J. I.; *"Glottal-Source Spectral Biometry for Voice Characterisation*", Workshop en Audio Cognitivo para Aplicaciones en Seguridad y Acústica Forense, Proyecto HESPERIA-CENIT, Maspalomas, Gran Canaria, Spain. Jun 2007
  - Gómez, P., Álvarez, A., Mazaira, L. M., Fernández, R., Nieto, V., Martínez, R., Muñoz, C., Rodellar, V.; *"A Hybrid Parameterisation Technique for Speaker Identification"*, Workshop en Audio Cognitivo para Aplicaciones en Seguridad y Acústica Forense, Proyecto HESPERIA-CENIT, Maspalomas, Gran Canaria, Spain. Jun 2007
  - Gómez, P. Álvarez, A., **Mazaira, L. M.**, Fernández, R., Rodellar, V.; "*Estimating the Stability and Dispersion of the Biometric Glottal Fingerprint in Continuous Speech*", ISCA Tutorial and Research Workshop NOLISP'07, Paris, France, May. 2007.

# I. BIBLIOGRAPHY

**[Abdulla,2001]** Abdulla, W.H. and Kasabov, N.K. *"Improving speech recognition performance through gender separation"*, Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES), pp.218-222, 2001, Dunedin, New Zealand. [ISBN:1-877139-40-8]

**[Abu El-Yazeed,2004]** Abu El-Yazeed, M.F.; El Gamal, M.A. and El Ayadi, M.M.H. *"On the determination of optimal model order for GMM-based text-independent speaker identification"*, EURASIP Journal on Advances in Signal Processing, vol.2004, issue 1, pp.1078-1087, 2004, Springer. [ISSN:1110-8657]

**[Adami,2003]** Adami, A.G.; Mihaescu, R. et al. *"Modeling prosodic dynamics for speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.4, pp.788-791, 2003, Hong Kong, China. [ISSN:1520-6149 | ISBN:0-7803-7663-3]

**[Adhami,2001]** Adhami, R. and Meenen, P. *"Fingerprinting for security"*, IEEE Potentials, vol.20, issue 3, pp.33-38, 2001, IEEE Computer Society. [ISSN:0278-6648]

**[Akaike,1974]** Akaike, H. *"A new look at the statistical model identification"*, IEEE Transactions on Automatic Control, vol.19, issue 6, pp.716-723, 1974, IEEE Computer Society. [ISSN:0018-9286]

**[Akande,2005]** Akande, O.O. and Murphy, P.J. *"Estimation of the vocal tract transfer function with application to glottal wave analysis"*, Speech Communication, vol.46, issue 1, pp.15-36, 2005, Elsevier.

**[Alipour,2000]** Alipour, F.; Berry, D.A. and Titze, I.R. *"A finite-element model of vocal-fold vibration"*, The Journal of the Acoustical Society of America, vol.108, issue 6, pp.3003-3012, 2000, Acoustic Society of America. [ISSN:0001-4966]

**[Alku,1992]** Alku,P. *"An automatic method to estimate the time-based parameters of the glottal pulseform"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.29-32, 1992, San Francisco, CA, USA. [ISSN:1520-6149 | ISBN:0-7803-0532-9]

**[Alku,1994]** Alku, P. and Vilkman, E. *"Estimation of the glottal pulseform based on Discrete All Pole modelling"*, Proceedings of the 3rd International Conference on Spoken Language Processing, pp.1619-1622, 1994, Yokohama, Japan.

**[Alku,2003]** Alku, P. *"Parameterisation Methods of the Glottal Flow Estimated by Inverse Filtering"*, Voice Quality: Functions, Analysis and Synthesis, VOQUAL, pp.81-88, 2003, Geneva, Switzerland.

**[Alonso-Fernandez,2005]** Alonso-Fernández, F.; Fierrez-Aguilar, J. et al. *"On-line signature verification using Tablet PC"*, Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, ISPA, pp.245-250, 2005, Zagreb, Croatia. [ISSN:1845-5921 | ISBN:953-184-089-X]

**[Al-Raisi,2008]** Al-Raisi, A.N. and Al-Khouri, A.M. *"Iris recognition and the challenge of homeland and border control security in UAE"*, Telematics and Informatics, vol.25, issue 2, pp.117-132, 2008, Tarrytown, NY, USA. [ISSN:0736-5853]

**[Amira,2005]** Amira, A. and Farrell, P. *"An automatic face recognition system based on wavelet transforms"*, IEEE International Symposium on Circuits and Systems, ISCAS, vol.6, pp.6252-6255, 2005, Kobe, Japan. [ISBN:0-7803-8834-8]

**[Ang,1997]** Ang, K.K. and Kot, A.C. *"Speaker verification for home security system"*, Proceedings of the IEEE International Symposium on Consumer Electronics, ISCE, pp.27-30, 1997, Singapore. [ISBN:0-7803-4371-9]

**[Araujo,2005]** Araujo, L.C.F.; Sucupira, L.H.R.Jr. et al. *"User authentication through typing biometrics features"*, IEEE Transactions on Signal Processing, vol.53, issue 2, pp.851-855, 2005, IEEE Computer Society. [ISSN:1053-587X]

**[Arroabarren,2003]** Arroabarren, I. and Carlosena, A. *"Glottal Source Parameterization: A comparative study"*, Voice Quality: Functions, Analysis and Synthesis, VOQUAL, pp.29-34, 2003, Geneva, Switzerland.

**[Atal,1976]** Atal, B.S. *"Automatic recognition of speakers from their voices"*, Proceedings of the IEEE, vol.64, issue 4, pp.460-475, 1976, IEEE Computer Society. [ISSN:0018-9219]

**[Auckenthaler,2000]** Auckenthaler, R.; Carey, M. and Lloyd-Thomas, H. *"Score Normalization for Text-Independent Speaker Verification Systems"*, Digital Signal Processing, vol.10, issue 1-3, pp.42-54, 2000, Elsevier. [ISSN:1051-2004]

**[Backstrom,2002]** Backstrom, T.; Alku, P. and Vilkman, E. *"Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range"*, IEEE Transactions on Speech and Audio Processing, vol.10, issue 3, pp.186-192, 2002, IEEE Computer Society. [ISSN:1063-6676]

**[Badran,2000]** Badran, E.F.M.F. and Selim, H. *"Speaker recognition using artificial neural networks based on vowel phonemes"*, 5th International Conference on Signal Processing Proceedings, WCCC-ICSP, vol.2, pp.796-802, 2000, Beijing, China. [ISBN:0-7803-5747-7]

**[Bahl,1983]** Bahl, L.R.; Jelinek,F. and Mercer,R. *"A Maximum Likelihood Approach to Continuous Speech Recognition"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.5, issue 2, pp.179-190, 1983, IEEE Computer Society. [ISSN:0162-8828]

**[Bahl,1986]** Bahl, L.; Brown, P. et al. *"Maximum mutual information estimation of hidden Markov model parameters for speech recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.11, pp.49-52, 1986, Tokio, Japan.

**[Bakis,1976]** Bakis, R. *"Continuous speech recognition via centisecond acoustic states"*, The Journal of the Acoustical Society of America, vol.59, issue 1, pp.97, 1976, Acoustic Society of America. [ISSN:0001-4966]

**[Barra,2006]** Barra, R.; Montero, J.M. et al. *"Prosodic and Segmental Rubrics in Emotion Identification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.1085-1088, 2006, Toulouse, France. [ISSN:1520-6149 | ISBN:1-4244-0469-X]

**[Barras,2003]** Barras, C. and Gauvain, J. *"Feature and score normalization for speaker verification of cellular data"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.49-52, 2003, Hong Kong, China. [ISSN:1520-6149 | ISBN:0-7803-7663-3]

**[Baum,1966-A]** Baum, L.E. and Eagon, J.A. *"An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology"*, Bulletin of the American Mathematical Society, vol.73, issue 3, pp.360-363, 1966, American Mathematical Society. [ISSN:0273-0979]

**[Baum,1966-B]** Baum, L.E. and Petrie, T. *"Statistical Inference for Probabilistic Functions of Finite State Markov Chains"*, Annals of Mathematical Statistics, vol.37, issue 6, pp.1554-1563, 1966, The Institute of Mathematical Statistics. [ISSN:0003-4851]

**[Baum,1968]** Baum, L.E. and Sell, G.R. *"Growth transformations for funtions on manifolds"*, Pacific Journal of Mathematics, vol.27, issue 2, pp.211-227, 1968, Pacific Journal of Mathematics. [ISSN:0030-8730]

**[Baum,1970]** Baum, L.E.; Petrie, T. et al. *"A maximization technique occurring in the statictical analysis of probabilistic functions of Markov chains"*, Annals of Mathematical Statistics, vol.41, issue 1, pp.164-171, 1970, The Institute of Mathematical Statistics. [ISSN:0003-4851]

**[Baum,1972]** Baum, L.E. *"An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process"*, Inequalities, III: Proceedings of the Third Symposium on Inequalities, vol.3, pp.1-8, 1972, New York, Academic Press. [ISBN-10:0126403031 | ISBN-13:9780126403039]

**[Bellegarda,1990]** Bellegarda, J.R. and Nahamoo, D. *"Tied mixture continuous parameter modeling for speech recognition"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.38, issue 12, pp.2033-2045, 1990, IEEE Computer Society. [ISSN:0096-3518]

**[Ben,2002]** Ben, M.; Blouet, R. and Bimbot, F. *"A Monte Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.689-692, 2002, Orlando, FL, USA. [ISSN:1520-6149 | ISBN:0-7803-7402-9]

**[Bengio,1995]** Bengio, Y. *"Neural Networks for Speech and Sequence Recognition"*, 1st Edition, pp.167, 1995, International Thomson Computer Press. [ISBN-10:1850321701 | ISBN-13:978-1850321705]

**[Bengio,2005]** Bengio, S.; Mariethoz, J. and Keller, M. *"The Expected Performance Curve"*, International Conference on Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning, 2005, Bonn, Germany.

**[Bergadano,2002]** Bergadano, F.; Gunetti, D. and Picardi, C. *"User authentication through keystroke dynamics"*, ACM Transactions on Information and System Security (TISSEC), vol.5, issue 4, pp.367-397, 2002, New York, NY, USA. [ISSN:1094-9224]

**[Berry,2001-A]** Berry, D.A. "Mechanisms of modal and nonmodal phonation", Journal of Phonetics, vol.29, issue 4, pp.431-450, 2001, Elsevier. [ISSN:0095-4470]

**[Berry,2001-B]** Berry, D.A.; Montequin, D.W. and Tayama, N. *"High-speed digital imaging of the medial surface of the vocal folds"*, The Journal of the Acoustical Society of America, vol.110, issue 5, pp.2539-2547, 2001, Acoustic Society of America. [ISSN:0001-4966]

**[Berry,2002]** Berry, D.A. *"Examination of models of mucosal wave propagation"*, The Journal of the Acoustical Society of America, vol.112, issue 5, pp.2446-2446, 2002, Acoustic Society of America. [ISSN:0001-4966]

**[Bilmes,1999]** Bilmes, J.A. *"Buried Markov models for speech recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.713-716, 1999, Phoenix,Arizona,USA. [ISSN:1520-6149 | ISBN:0-7803-5041-3]

**[Bimbot,2004]** Bimbot, F.; Bonastre, J. F. et al. *"A tutorial on text-independent speaker verification"*, EURASIP Journal on Advances in Signal Processing, vol.2004, issue 1, pp.430-451, 2004, Springer. [ISSN:1110-8657]

**[Bishop,2006]** Bishop, C.M. *"Pattern Recognition and Machine Learning"*, 1st Edition, pp.740, 2006, Springer. [ISBN-10:0387310738 | ISBN-13:978-0387310732]

**[Bleha,1990]** Bleha, S.; Slivinsky, C. and Hussien, B. *"Computer-access security systems using keystroke dynamics"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.12, issue 12, pp.1217-1222, 1990, IEEE Computer Society. [ISSN:0162-8828]

**[Boakye,2004]** Boakye, K. and Peskin, B. *"Text-Constrained Speaker Recognition on a Text-Independent Task"*, Odyssey 2004: The speaker and Language Recognition Workshop, pp.129-134, 2004, Toledo, Spain.

**[Boakye,2005]** Boakye, K. *"Speaker Recognition in the Text-Independent Domain Using Keyword Hidden Markov Models"*, M.S. Thesis, University of California at Berkeley, pp.1-35, 2005, USA.

**[Bocklet,2009]** Bocklet, T. and Shriberg, E. *"Speaker recognition using syllable-based constraints for cepstral frame selection"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.4525-4528, 2009, Taipei, Taiwan. [ISSN:1520-6149 | ISBN:978-1-4244-2353-8]

**[Boechat,2007]** Boechat, G.C.; Ferreira, J.C. and Filho, E.C.B.C. *"Authentication personal"*, International Conference on Intelligent and Advanced Systems, ICIAS, pp.254-256, 2007, Kuala Lupmur,Malaysia. [ISBN:978-1-4244-1355-3]

**[Bonastre,2005]** Bonastre, J.F.; Wils, F. and Meignier, S. *"ALIZE, a free toolkit for speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.737-740, 2005, Philadelphia, PA, USA. [ISSN:1520-6149 | ISBN:0-7803-8874-7]

**[Bonastre,2008]** Bonastre, J. F.; Scheffer, N. et al. *"ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition"*, Odyssey 2008: The Speaker and Language Recognition Workshop, paper 20, 2008, Stellenbosch, South Africa.

**[Boreki,2005]** Boreki, G. and Zimmer, A. *"Hand geometry: a new approach for feature extraction"*, IEEE Workshop on Automatic Identification Advanced Technologies, pp.149-154, 2005, Buffalo, NY, USA. [ISBN:0-7695-2475-3]

**[Boser,1992]** Boser, B.E.; Guyon, I.M. and Vapnik, V.N. *"A training algorithm for optimal margin classifiers"*, Proceedings of the fifth annual workshop on Computational learning theory, COLT, pp.144-152, 1992, Pittsburgh, PA, USA. [ISBN:0-89791-497-X]

**[Bowyer,2004]** Bowyer, K.W. *"An elective course in biometrics and privacy"*, 34th Annual Frontiers in Education, FIE, vol.3, pp.12-17, 2004, Savannah, GA, USA. [ISSN:0190-5848 | ISBN:0-7803-8552-7]

**[Brault,1993]** Brault, J. and Plamondon, R. *"Segmenting handwritten signatures at their perceptually important points"*, IEEE Transactions on Pattern Analysis and

Machine Intelligence, vol.15, issue 9, pp.953-957, 1993, IEEE Computer Society. [ISSN:0162-8828]

**[Brookes,1994]** Brookes, D.M. and Chan, D.S.F. *"Speaker characteristics from a glottal airflow model using robust inverse filtering"*, Proceedings of the Institute of Acoustics, vol.16, issue 5, pp.501-508, 1994 . [ISSN:0309-8117]

**[Brummer,2006]** Brummer, N. and du Preez, J. *"Application-independent evaluation of speaker detection"*, Computer Speech and Language - Odyssey 2004: The speaker and Language Recognition Workshop, vol.20, issue 2-3, pp.230-275, 2006, Toledo, Spain. [ISSN:0885-2308]

**[Brummer,2007]** Brummer, N.; Burget, L. et al. *"Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.2072-2084, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Burge,2000]** Burge, M. and Burger, W. *"Ear biometrics in computer vision"*, Proceedings of the 15th International Conference on Pattern Recognition, vol.2, pp.822-826, 2000, Barcelona, Spain. [ISSN:1051-4651 | ISBN:0-7695-0750-6]

**[Burges,1998]** Burges, C.J.C. *"A Tutorial on Support Vector Machines for Pattern Recognition"*, Data Mining and Knowledge Discovery, vol.2, issue 2, pp.121-167, 1998, Kluwer Academic Publishers. [ISSN:384-5810]

**[Burget,2007]** Burget, L.; Matejka, P. et al. *"Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.1979-1986, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Butler,2005]** Butler, J.M. *"Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers"*, 2nd Edition, pp.688, 2005, Academic Press. [ISBN-10:0121479528 | ISBN-13:978-0121479527]

**[Butt,2008]** Butt, M.A.A.; Masood, H. et al. *"Palmprint Identification Using Contourlet Transform"*, 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS, pp.1-5, 2008, Arlington, VA, USA. [ISBN:978-1-4244-2729-1]

**[Cadavid,2008]** Cadavid, S. and Abdel-Mottaleb, M. *"3D ear modeling and recognition from video sequences using shape from shading"*, Proceedings of the 19th International Conference on Pattern Recognition, pp.1-4, 2008, Tampa, FL, USA. [ISSN:1051-4651 | ISBN:978-1-4244-2174-9]

**[Campbell,1997]** Campbell, J.P.Jr. *"Speaker recognition: a tutorial"*, Proceedings of the IEEE, vol.85, issue 9, pp.1347-1462, 1997, IEEE Computer Society. [ISSN:0018-9219]

**[Campbell,2002-A]** Campbell, W.M. *"Generalized linear discriminant sequence kernels for speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.161-164, 2002, Orlando, FL, USA. [ISSN:1520-6149 | ISBN:0-7803-7402-9]

**[Campbell,2002-B]** Campbell, W.M.; Assaleh, K.T. and Broun, C.C. *"Speaker recognition with polynomial classifiers"*, IEEE Transactions on Speech and Audio Processing, vol.10, issue 4, pp.205-212, 2002, IEEE Computer Society. [ISSN:1063-6676]

**[Campbell,2003-A]** Campbell, J.P.Jr.; Reynolds, D.A. and Dunn, R.B. *"Fusing High- and Low-Level Features for Speaker Recognition"*, Proceedings of the 8th European Conference on Speech Communication and Technology, pp.2665-2668, 2003, Geneva, Switzerland.

**[Campbell,2003-B]** Campbell, W.M.; Campbell, J.P. et al. *"Phonetic speaker recognition with support vector machines"*, Advances in Neural Information Processing Systems (NIPS), vol.16, pp.1377-1384, 2003, Vancouver, Canada. [ISBN:0-262-20152-6]

**[Campbell,2004-A]** Campbell, W.; Campbell,J. R. et al. *"High-level speaker verification with support vector machines"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.73-76, 2004, Montreal, Canada. [ISSN:1520-6149 | ISBN:0-7803-8484-9]

**[Campbell,2004-B]** Campbell, W.M.; Reynolds, D.A. and Campbell, J.P. *"Fusing discriminative and generative methods for speaker recognition experiments on switchboard and nfi/tno field data"*, Odyssey 2004: The speaker and Language Recognition Workshop, pp.41-44, 2004, Toledo, Spain.

**[Campbell,2006]** Campbell, W.M.; Campbell, J.P. et al. *"Support vector machines for speaker and language recognition"*, Computer Speech and Language - Odyssey 2004: The speaker and Language Recognition Workshop, vol.20, issue 2-3, pp.210-229, 2006, Toledo, Spain. [ISSN:0885-2308]

**[Campbell,2006-A]** Campbell, W.M.; Sturim, D.E. and Reynolds, D.A. *"Support vector machines using GMM supervectors for speaker verification"*, IEEE Signal Processing Letters, vol.13, issue 5, pp.308-311, 2006, IEEE Computer Society. [ISSN:1070-9908]

**[Campbell,2006-B]** Campbell, W.M.; Sturim, D.E. et al. *"SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.97-100, 2006, Toulouse, France. [ISSN:1520-6149 | ISBN:1-4244-0469-X]

**[Campbell,2007]** Campbell, W.M.; Campbell, J.P. et al. *"Speaker Verification Using Support Vector Machines and High-Level Features"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.2085-2094, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Carey,1991]** Carey, M.J.; Parris, E.S. and Bridle, J.S. *"A speaker verification system using alpha-nets"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.397-400, 1991, Toronto, Ontario, Canada. [ISSN:1520-6149 | ISBN:0-7803-0003-3]

**[CASIA,2014]** CASIA *"CASIA Databases"*, Center for Biometrics and Security Research, Institute of Automation Chinese Academy of Sciences, Beijing, China. [on-line acc. 2014 http://www.cbsr.ia.ac.cn/english/Databases.asp]

**[Castaldo,2007]** Castaldo, F.; Colibro, D. et al. *"Compensation of Nuisance Factors for Speaker and Language Recognition"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.1969-1978, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Chai,2006]** Chai, Y.; Ren, J. et al. *"Automatic Gait Recognition using Dynamic Variance Features"*, Proceedings of the 7th International Conference on Automatic

Face and Gesture Recognition, FGR, pp.475-480, 2006, Southampton, UK. [ISBN:0-7695-2503-2]

**[Chang,2005]** Chang, W. *"Improving hidden Markov models with a similarity histogram for typing pattern biometrics"*, IEEE International Conference on Information Reuse and Integration, IRI, pp.487-493, 2005, Las Vegas, NV, USA. [ISBN:0-7803-9093-8]

**[Charbuillet,2007]** Charbuillet, C.; Gas, B. et al. *"Multi Filter Bank Approach for Speaker Verification Based on Genetic Algorithm"*, Advances in Nonlinear Speech Processing (LNCS), vol.4885, pp.105-113, 2007, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-77346-7]

**[Chaudhari,2003]** Chaudhari, U.V.; Navratil, J. and Maes, S.H. *"Multigrained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition"*, IEEE Transactions on Speech and Audio Processing, vol.11, issue 1, pp.61-69, 2003, IEEE Computer Society. [ISSN:1063-6676]

**[Chen,2001]** Chen, J.; Zhang, C. and Rong, G. *"Palmprint recognition using crease"*, Proceedings of the International Conference on Image Processing, vol.3, pp.234-237, 2001, Thessaloniki, Greece. [ISBN:0-7803-6725-1]

**[Chen,2004]** Chen, W. and Chang, W. *"Applying hidden Markov models to keystroke pattern analysis for password verification"*, IEEE International Conference on Information Reuse and Integration, IRI, pp.467-474, 2004, Las Vegas, NV, USA. [ISBN:0-7803-8819-4]

**[Chen,2005]** Chen, P.H.; Lin,C.J. and Schölkopf, B. *"A tutorial on v-support vector machines"*, Applied Stochastic Models in Business and Industry, vol.21, issue 2, pp.111-136, 2005 . [ISSN:1526-4025]

**[Chen,2007]** Chen, H. and Bhanu, B. *"Human Ear Recognition in 3D"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.29, issue 4, pp.718-737, 2007, IEEE Computer Society. [ISSN:0162-8828]

**[Cheng,1989]** Cheng, Y.M. and O'Shaughnessy, D. *"Automatic and reliable estimation of glottal closure instant and period"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.37, issue 12, pp.1805-1815, 1989, IEEE Computer Society. [ISSN:0096-3518]

**[Cheng,2004]** Cheng, S.; Wang, H. and Fu,H. *"A model-selection-based self-splitting Gaussian mixture learning with application to speaker identification"*, EURASIP Journal on Advances in Signal Processing, vol.2004, issue 1, pp.2626-2639, 2004, Springer. [ISSN:1110-8657]

**[Chetouani,2009]** Chetouani, M.; Faundez-Zanuy, M. et al. *"Investigation on LP-residual representations for speaker identification"*, Pattern Recognition, vol.42, issue 3, pp.487-494, 2009, Elsevier. [ISSN:0031-3203]

**[Cho,2006]** Cho, T.H. *"Pattern Classification Methods for Keystroke Analysis"*, International Joint Conference SICE-ICASE, pp.3812-3815, 2006, Busan,South Korea. [ISBN:89-950038-4-7]

**[Choras,2006]** Choras, M.; Choras, R.S. *"Geometrical Algorithms of Ear Contour Shape Representation and Feature Extraction"*, Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, ISDA, vol.2, pp.451-456, 2006, Jinan, China. [ISBN:0-7695-2528-8]

**[Choras,2007-A]** Choras, M. *"Image Feature Extraction Methods for Ear Biometrics--A Survey"*, 6th International Conference on Computer Information Systems and Industrial Management Applications, CISIM, pp.261-265, 2007, Minneapolis, MN, USA. [ISBN:0-7695-2894-5]

**[Choras,2007-B]** Choras, M. *"Emerging Methods of Biometrics Human Identification"*, Second International Conference on Innovative Computing, Information and Control, ICICIC, pp.365-368, 2007, Kumamoto, Japan. [ISBN:0-7695-2882-1]

**[Chu,2001]** Chu, S.C and Roddick, J.F. *"Pattern Clustering Using Incremental Splitting for Non-Uniformly Distributed Data"*, Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies, KES, pp.1037-1041, 2001, IOS Press. [ISBN:1-58603-192-9]

**[Chu,2005]** Chu, C.T. and Chen, C.H. *"High performance iris recognition based on LDA and LPCC"*, 17th IEEE International Conference on Tools with Artificial Intelligence, ICTAI, pp.421-425, 2005, Hong Kong, China. [ISSN:1082-3409 | ISBN:0-7695-2488-5]

**[Cieri,2007]** Cieri, C.; Corson, L. et al. *"Resources for New Research Directions in Speaker Recognition:The Mixer 3, 4 and 5 Corpora"*, Proceedings of the 8th Annual Conference of the International Speech Communication Association, pp.950-953, 2007, Antwerp, Belgium.

**[Clarke,1994]** Clarke, R. *"Human Identification in Information Systems: Management Challenges and Public Policy Issues"*, Information Technology & People, vol.7, issue 4, pp.6-37, 1994, MCB UP Ltd. [ISSN:0959-3845]

**[Colombi,1996]** Colombi, J.M.; Ruck, D.W. et al. *"Cohort selection and word grammar effects for speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.85-88, 1996, Atlanta, GA, USA. [ISSN:1520-6149 | ISBN:0-7803-3192-3]

**[Cooley,1965]** Cooley, J.W. and Tukey, J.W. *"An Algorithm for the Machine Calculation of Complex Fourier Series"*, Mathematics of Computation, vol.19, issue 90, pp.297-301, 1965, American Mathematical Society. [ISSN:0025-5718]

**[Cooley,1969]** Cooley, J.W.; Lewis, P. and Welch, P. *"The finite Fourier transform"*, IEEE Transactions on Audio and Electroacoustics, vol.17, issue 2, pp.77-85, 1969, IEEE Computer Society. [ISSN:0018-9278]

**[Covavisaruch,2005]** Covavisaruch, N.; Prateepamornkul, P. et al. *"Personal Verification and Identification Using Hand Geometry"*, ECTI Transactions on Computer and Information Technology, vol.1, issue 2, pp.134-140, 2005 . [ISSN:1905 - 050X]

**[Cristianini,2000]** Cristianini, N. and Shawe-Taylor, J. *"An Introduction to Support Vector Machines and Other Kernel-based Learning Methods"*, 1st Edition, pp.189, 2000, Cambridge University Press. [ISBN-10:0521780195 | ISBN-13:978-0521780193]

**[Cunado,2003]** Cunado, D.; Nixon, M.S. and Carter, J.N. *"Automatic extraction and description of human gait models for recognition purposes"*, Computer Vision and Image Understanding, vol.90, issue 1, pp.1-41, 2003, Elsevier. [ISSN:1077-3142]

**[Cutzu,2003]** Cutzu, F. *"Polychotomous Classification with Pairwise Classifiers: A New Voting Principle"*, Multiple Classifier Systems (LNCS), vol.2709, pp.115-124, 2003, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-44938-6]

**[D'Alessandro,2007]** D'Alessandro, C.; Bozkurt, B. et al. *"Phase-Based Methods for Voice Source Analysis"*, Advances in Nonlinear Speech Processing (LNCS), vol.4885, pp.1-27, 2007, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-77346-7]

**[Daugman,1993]** Daugman, J.G. *"High confidence visual recognition of persons by a test of statistical independence"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.15, issue 11, pp.1148-1161, 1993, IEEE Computer Society. [ISSN:0162-8828]

**[Daugman,2004]** Daugman, J. *"How iris recognition works"*, IEEE Transactions on Circuits and Systems for Video Technology, vol.14, issue 1, pp.21-30, 2004, IEEE Computer Society. [ISSN:1051-8215]

**[Davis,1980]** Davis, S. and Mermelstein,P. *"Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.28, issue 4, pp.357-366, 1980, IEEE Computer Society. [ISSN:0096-3518]

**[de Oliveira Rosa,2000]** de Oliveira Rosa, M.; Pereira, J.C. and Grellet, M. *"Adaptive estimation of residue signal for voice pathology diagnosis"*, IEEE Transactions on Biomedical Engineering, vol.47, issue 1, pp.96-104, 2000, IEEE Computer Society. [ISSN:0018-9294]

**[Dehak,2006]** Dehak, N. and Chollet,G. *"Support Vector Gmms for Speaker Verification"*, IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, pp.1-4, 2006, San Juan, Puerto Rico. [ISBN:1-424400471-1]

**[Dehak,2008]** Dehak, N.; Dehak, R. et al. *"Comparison between factor analysis and GMM support vector machines for speaker verification"*, Odyssey 2008: The Speaker and Language Recognition Workshop, paper 9, 2008, Stellenbosch, South Africa.

**[Dehak,2009-A]** Dehak, N. *"Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification"*, M.S. Thesis, École de technologie supérieure, Universite du Quebec, pp.1-183, 2009, Montreal, CANADA. [ISBN:978-0-494-50490-1]

**[Dehak,2009-B]** Dehak, N.; Dehak, R. et al. *"Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification"*, Proceedings of the 10th Annual Conference of the International Speech Communication Association, pp.1559-1562, 2009, Brighton, UK.

**[Dehak,2009-C]** Dehak, N.; Kenny, P. et al. *"Support vector machines and Joint Factor Analysis for speaker verification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.4237-4240, 2009, Taipei, Taiwan. [ISSN:1520-6149 | ISBN:978-1-4244-2353-8]

**[Dehak,2010]** Dehak, N.; Kenny, P. et al. *"Front-End Factor Analysis For Speaker Verification"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.19, issue 4, pp.788-798, 2010, IEEE Computer Society. [ISSN:1558-7916]

**[Delac,2004]** Delac, K. and Grgic, M. *"A survey of biometric recognition methods"*, Proceedings of the 46th International Symposium on Electronics in Marine, Elmar, pp.184-193, 2004, Zadar, Croatia. [ISSN:1334-2630 |ISBN:953-7044-02-5]

**[Deller,1999]** Deller, J.R.; Hansen, J.H.L. and Proakis, J.G. *"Discrete-Time Processing of Speech Signals"*, 1st Edition, pp.936, 1999, Wiley-IEEE Press. [ISBN-10:0780353862 | ISBN-13:978-0780353862]

**[Dempster,1977]** Dempster, A.P.; Laird, N.M. and Rubin, D.B. *"Maximum Likelihood from Incomplete Data via the EM Algorithm"*, Journal of the Royal Statistical Society, Series B, vol.39, issue 1, pp.1-38, 1977 . [ISSN:1467-9868]

**[Denes,2012]** Denes, P.B. and Pinson, E.N. *"The speech chain: The physics and biology of spoken language"*, pp.174, 2012, Literary Licensing, LLC. [ISBN-10:1258407752 | ISBN-13:978-1258407759]

**[Deriche,2008]** Deriche, M. *"Trends and Challenges in Mono and Multi Biometrics"*, First Workshops on Image Processing Theory, Tools and Applications, IPTA, pp.1-9, 2008, Sousse,Tunisia. [ISBN:978-1-4244-3321-6]

**[Deshpande,2008]** Deshpande, M.S. and Holambe, R.S. *"Text-Independent Speaker Identification Using Hidden Markov Models"*, First International Conference on Emerging Trends in Engineering and Technology, ICETET, pp.641-644, 2008, Nagpur, Maharashtra, India. [ISBN:978-0-7695-3267-7]

**[Digalakis,1996]** Digalakis, V.V.; Monaco, P. and Murveit, H. *"Genones: generalized mixture tying in continuous hidden Markov model-based speech recognizers"*, IEEE Transactions on Speech and Audio Processing, vol.4, issue 4, pp.281-289, 1996, IEEE Computer Society. [ISSN:1063-6676]

**[Ding,2005]** Ding, Y.; Zhuang, D. and Wang, K. *"A study of hand vein recognition method"*, IEEE International Conference Mechatronics and Automation, vol.4, pp.2106-2110, 2005, Niagara Falls, Ontario, Canada. [ISBN:0-7803-9044-X]

**[Dobeš,2014]** Dobeš, M. and Machala, L. *"Iris Database"*, Department of Computer Science, Palacký University, Olomouc, Czech Republic. [on-line acc. 2014 http://phoenix.inf.upol.cz/iris/]

**[Doddington,1985]** Doddington, G.R. *"Speaker recognition: Identifying people by their voices"*, Proceedings of the IEEE, vol.73, issue 11, pp.1651-1664, 1985, IEEE Computer Society. [ISSN:0018-9219]

**[Doddington,1998]** Doddington, G.; Liggett, W. et al. *"Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation"*, Proceedings of the 5th International Conference on Spoken Language Processing, pp.1-5, 1998, Sydney, Australia.

**[Doddington,2001]** Doddington, G.R. *"Speaker recognition based on idiolectal differences between speakers"*, Proceedings of the 7th International Conference on Spoken Language Processing, pp.2521-2524, 2001, Aalborg, Denmark.

**[Duda,2000]** Duda, R.O.; Hart, P.E. and Stork, D.G. *"Pattern Classification"*, 2nd Edition, pp.680, 2000, Wiley-Interscience. [ISBN-10:0471056693 | ISBN-13:978-0471056690]

**[Dugelay,2002]** Dugelay, J.; Junqua, J.C. et al. *"Recent advances in biometric person authentication"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.4, pp.4060-4063, 2002, Orlando, FL, USA. [ISSN:1520-6149 | ISBN:0-7803-7402-9]

**[Dumas,2005]** Dumas, B.; Pugin,C. et al. *"MyIdea - Multimodal Biometrics Database, Description of Acquisition Protocols"*, Proceedings of Third COST 275 Workshop Biometrics on the Internet, pp.59-62, 2005, Hatfield, UK.

**[Dumitru,2006]** Dumitru, C.O.; Gavat, I. and Vieru, R. *"Speaker Verification using HMM for Romanian Language"*, 48th International Symposium ELMAR-2006 focused on Multimedia Signal Processing and Communications, pp.131-134, 2006, Zadar, Croatia. [ISSN:1334-2630 |ISBN:953-7044-03-3]

**[Durbin,1960]** Durbin, J. *"The Fitting of Time-Series Models"*, Revue de l'Institut International de Statistique / Review of the International Statistical Institute, vol.28, issue 3, pp.233-244, 1960, International Statistical Institute (ISI). [ISSN:0373-1138]

**[El Hannani,2007]** El Hannani, A. and Petrovska-Delacrétaz, D. *"Data-Driven High-Level Information for Text-Independent Speaker Verification"*, IEEE Workshop on Automatic Identification Advanced Technologies, pp.209-213, 2007, Alghero, Italy. [ISBN:1-4244-1300-1]

**[Elliott,2007]** Elliott, S.J.; Massie, S.A. and Sutton, M.J. *"The Perception of Biometric Technology: A Survey"*, IEEE Workshop on Automatic Identification Advanced Technologies, pp.259-264, 2007, Alghero, Italy. [ISBN:1-4244-1300-1]

**[Elsherief,2006]** Elsherief, S.M.; Allam, M.E. and Fakhr, M.W. *"Biometric Personal Identification Based on Iris Recognition"*, The 2006 International Conference on Computer Engineering and Systems, pp.208-213, 2006, Cairo, Egypt. [ISBN:1-4244-0271-9]

**[Ephraim,1985]** Ephraim, Y. and Malah, D. *"Speech enhancement using a minimum mean-square error log-spectral amplitude estimator"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.33, issue 2, pp.443-445, 1985, IEEE Computer Society. [ISSN:0096-3518]

**[Ephraim,1987]** Ephraim, Y.; Dembo, A. and Rabiner, L. *"A minimum discrimination information approach for hidden Markov modeling"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.12, pp.25-28, 1987, Dallas, TX, USA.

**[Faltlhauser,2001]** Faltlhauser, R. and Ruske, G. *"Improving Speaker Recognition Performance Using Phonetically Structured Gaussian Mixture Models"*, Proceedings of the 7th International Conference on Spoken Language Processing, pp.751-754, 2001, Aalborg, Denmark.

**[Fant,1971]** Fant, G. *"Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations"*, Revised edition, pp.328, 1971, Walter de Gruyter. [ISBN-10:9027916004 | ISBN-13:978-9027916006]

**[Fant,1985]** Fant, G.; Liljencrants, J. and Lin, Q. *"A four-parameter model of glottal flow"*, Quarterly Progress and Status Report QPSR 4 (STL-QPSR), vol.26, issue 4, pp.1-13, 1985, Deparment of Speech, Music and Hearing, KTH CSC- The School of Computer Science and Communication, Stockholm, Sweden.

**[Farrús,2007]** Farrús, M.; Hernando, J. and Ejarque, P. *"Jitter and Shimmer Measurements for Speaker Recognition"*, Proceedings of the 8th Annual Conference of the International Speech Communication Association, pp.778-781, 2007, Antwerp, Belgium.

**[Farzin,2008]** Farzin, H.; Abrishami-Moghaddam, H. and Mohammad-Shahram, M. *"A Novel Retinal Identification System"*, EURASIP Journal on Advances in Signal Processing, vol.2008, pp.1-10, 2008, Springer. [ISSN:1110-8657]

**[Fatima,2004]** Fatima, N.; Aftab, S. et al. *"Speaker recognition using lower formants"*, Proceedings of the 8th International Multitopic Conference, INMIC, pp.125-130, 2004, Lahore, Pakistan. [ISBN:0-7803-8680-9]

**[Faundez-Zanuy,2005]** Faundez-Zanuy, M.; Ferrer, M.A. et al. *"Hand Geometry Based Recognition with a MLP Classifier"*, Advances in Biometrics (LNCS), vol.3832, pp.721-727, 2005, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-31111-9]

**[Faundez-Zanuy,2006]** Faundez-Zanuy, M.; Fierrez-Aguilar, J. et al. *"Multimodal biometric databases: an overview"*, IEEE Aerospace and Electronic Systems Magazine, vol.21, issue 8, pp.29-37, 2006, IEEE Computer Society. [ISSN:0885-8985]

**[Faundez-Zanuy,2007]** Faundez-Zanuy, M.; Fabregas, J. et al. *"Evaluation of Supervised vs. Non Supervised Databases for Hand Geometry Verification"*, Computational and Ambient Intelligence (LNCS), vol.4507, pp.1122-1129, 2007, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-73006-4]

**[Fauve,2007]** Fauve, B.G.B.; Matrouf, D. et al. *"State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.1960-1968, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Fawcett,2006]** Fawcett, T. *"An introduction to ROC analysis"*, Pattern Recognition Letters, vol.27, issue 8, pp.861-874, 2006, Elsevier. [ISSN:0167-8655]

**[Ferras,2007]** Ferras, M.; Leung, C. et al. *"Constrained MLLR for Speaker Recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.4, pp.53-56, 2007, Honolulu, HI, USA. [ISSN:1520-6149 | ISBN:1-4244-0727-3]

**[Ferras,2009]** Ferras, M.; Leung, C.C. et al. *"Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.18, issue 6, pp.1366-1378, 2009, IEEE Computer Society. [ISSN:1558-7916]

**[Ferreira,2005]** Ferreira, A.A.; Ludermir, T.B. and de Aquino, R.R.B. *"A comparative study of neural network to artificial noses"*, Proceedingsof the IEEE International Joint Conference on Neural Networks, IJCNN, vol.4, pp.2081-2086, 2005, Montreal, Quebec, Canada. [ISBN:0-7803-9048-2]

**[Ferrer,2007-A]** Ferrer, L.; Shriberg, E. et al. *"Parameterization of Prosodic Feature Distributions for SVM Modeling in Speaker Recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.4, pp.233-236, 2007, Honolulu, HI, USA. [ISSN:1520-6149 | ISBN:1-4244-0727-3]

**[Ferrer,2007-B]** Ferrer, M.A.; Morales, A. et al. *"Low Cost Multimodal Biometric identification System Based on Hand Geometry, Palm and Finger Print Texture"*, 41st Annual IEEE International Carnahan Conference on Security Technology, pp.52-58, 2007, Ontario, Canada. [ISBN:978-1-4244-1129-0]

**[Ferrer,2008]** Ferrer, L.; Graciarena, M. et al. *"System combination using auxiliary information for speaker verification"*, IEEE International Conference on Acoustics,

Speech and Signal Processing, ICASSP, pp.4853-4856, 2008, Las Vegas, NV, USA. [ISSN:1520-6149 | ISBN:978-1-4244-1483-3]

**[Fierrez,2007]** Fierrez, J.; Ortega-Garcia, J. et al. *"Biosec baseline corpus: A multimodal biometric database"*, Pattern Recognition, vol.40, issue 4, pp.1389-1392, 2007, Elsevier. [ISSN:0031-3203]

**[Fierrez,2010]** Fierrez, J.; Galbally, J. et al. *"BiosecurID: a multimodal biometric database"*, Pattern Analysis & Applications, vol.13, issue 2, pp.235-246, 2010, Springer-Verlag. [ISSN:1433-7541]

**[Fine,2001]** Fine, S.; Navratil, J. and Gopinath, R.A. *"A hybrid GMM/SVM approach to speaker identification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.417-420, 2001, Salt Lake City, UT, USA. [ISSN:1520-6149 | ISBN:0-7803-7041-4]

**[Forney,1973]** Forney, G.D.Jr. *"The viterbi algorithm"*, Proceedings of the IEEE, vol.61, issue 3, pp.268-278, 1973, IEEE Computer Society. [ISSN:0018-9219]

**[Fouquier,2007]** Fouquier, G.; Likforman, L. et al. *"The Biosecure Geometry-Based System for Hand Modality"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.801-804, 2007, Honolulu, HI, USA. [ISSN:1520-6149 | ISBN:1-4244-0727-3]

**[Fowler,2000]** Fowler, J.E. *"QccPack: An Open-Source Software Library for Quantization, Compression, and Coding"*, Proceedings of the Data Compression Conference, DCC, 2000, Snowbird, UT, USA. [ISSN:1068-0314 | ISBN:0-7695-0592-9]

**[Fraley,1998]** Fraley, C. and Raftery, A.E. *"How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis"*, The Computer Journal, vol.41, issue 8, pp.578-588, 1998, British Computer Society. [ISSN:0010-4620]

**[Friedman,1996]** Friedman, J.H. *"Another Approach to Polychotomous Classification"*, Technical Report, pp.1-14, 1996, Department of Statistics, Stanford University, Stanford, CA, USA.

**[Friedman,1999]** Friedman, M. and Kandel, A. *"Introduction to Pattern Recognition : Statistical, Structural, Neural and Fuzzy Logic Approaches"*, pp.329, 1999, World Scientific Publishing Company Pvt. Ltd., In. [ISBN-10:9810233124 | ISBN-13:978-9810233129]

**[Furui,1981]** Furui, S. *"Cepstral analysis technique for automatic speaker verification"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.29, issue 2, pp.254-272, 1981, IEEE Computer Society. [ISSN:0096-3518]

**[Furui,1986]** Furui, S. *"Speaker-independent isolated word recognition using dynamic features of speech spectrum"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.34, issue 1, pp.52-59, 1986, IEEE Computer Society. [ISSN:0096-3518]

**[Furui,2000]** Furui, S. *"Digital Speech Processing, Synthesis, and Recognition"*, 2nd Edition, pp.476, 2000, CRC Press. [ISBN-10:0824704525 | ISBN-13:978-0824704520]

**[Gafurov,2007]** Gafurov, D.; Snekkenes, E. and Bours, P. *"Spoof Attacks on Gait Authentication System"*, IEEE Transactions on Information Forensics and Security, vol.2, issue 3, pp.491-502, 2007, IEEE Computer Society. [ISSN:1556-6013]

**[Gales,1996]** Gales, M.J.F. and Woodland, P.C. *"Mean and Variance Adaptation within the MLLR Framework"*, Computer Speech and Language, vol.10, issue 4, pp.249-264, 1996, Elsevier. [ISSN:0885-2308]

**[Galliano,2006]** Galliano, S.; Geoffrois, E. et al. *"Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News"*, 5th International Conference on Language Resources and Evaluation, LREC, pp.139-142, 2006, Genoa, Italy.

**[Garcia-Romero,2003]** Garcia-Romero, D.; Fierrez-Aguilar, J. et al. *"Support vector machine fusion of idiolectal and acoustic speaker information in Spanish conversational speech"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.229-232, 2003, Hong Kong, China. [ISSN:1520-6149 | ISBN:0-7803-7663-3]

**[Garcia-Romero,2006]** Garcia-Romero, D.; Fierrez-Aguilar, J. et al. *"Using quality measures for multilevel speaker recognition"*, Computer Speech and Language - Odyssey 2004: The speaker and Language Recognition Workshop, vol.20, issue 2-3, pp.192-209, 2006, Toledo, Spain. [ISSN:0885-2308]

**[Garcia-Salicetti,2003]** Garcia-Salicetti, S.; Beumier, C. et al. *"BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities"*, Audio- and Video-Based Biometric Person Authentication (LNCS), vol.2688, pp.845-853, 2003, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-40302-9]

**[Garofolo,2004]** Garofolo, J.S.; Laprun, C.D. et al. *"The NIST Meeting Room Pilot Corpus"*, 4th International Conference on Language Resources and Evaluation, LREC, pp.1411-1414, 2004, Lisbon, Portugal. [ISBN:2-9517408-1-6]

**[Gersho,1992]** Gersho, A. and Gray, R.M. *"Vector Quantization and Signal Compression"*, pp.732, 1992, Springer. [ISBN-10:0792391810 | ISBN-13:978-0792391814]

**[Glembek,2009]** Glembek, O.; Burget, L. et al. *"Comparison of scoring methods used in speaker recognition with Joint Factor Analysis"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.4057-4060, 2009, Taipei, Taiwan. [ISSN:1520-6149 | ISBN:978-1-4244-2353-8]

**[Glembek,2011]** Glembek, O.; Burget, L. et al. *"Simplification and optimization of i-vector extraction"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.4516-4519, 2011, Prague, Czech Repùblic. [ISSN:1520-6149 | ISBN:978-1-4577-0538-0]

**[Gómez,2004]** Gómez, P.; Godino, J.I.; et al. *"Biomechanical Parameter Fingerprint in the Mucosal Wave Power Spectral Density"*, Proceedings of the 8th International Conference on Spoken Language Processing, pp.842-845, 2004, Jeju, Korea.

**[Gómez,2004-A]** Gómez, P.; Díaz, F. et al. *"Precise reconstruction of the mucosal wave for voice pathology detection and characterization"*, The 12th European Signal Processing Conference (EUSIPCO), pp.2263-2266, 2004, Wien, AUTRICHE. [ISSN 2219-5491]

**[Gómez,2004-B]** Gómez, P.; Godino, J.I. et al. *"Evidence of vocal cord pathology from the mucosal wave cepstral contents"*, IEEE International Conference on Acoustics,

Speech and Signal Processing, ICASSP, vol.5, pp.437-440, 2004, Montreal, Canada. [ISSN:1520-6149 | ISBN:0-7803-8484-9]

**[Gómez,2005]** Gómez, P.; Godino, J.I. et al. *"Evidence of glottal source spectral features found in vocal fold dynamics"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.5, pp.441-444, 2005, Philadelphia, PA, USA. [ISSN:1520-6149 | ISBN:0-7803-8874-7]

**[Gómez,2005-A]** Gómez, P.; Martínez, R. et al. *"Estimation of Vocal Cord Biomechanical Parameters by Non-Linear Inverse Filtering of Voice"*, ITRW on Non-Linear Speech Processing (NOLISP), pp.174-183, 2005, Barcelona, Spain.

**[Gómez,2005-B]** Gómez, P.; Martínez, R. et al. *"Voice Pathology Detection by Vocal Cord Biomechanical Parameter Estimation"*, Nonlinear Analyses and Algorithms for Speech Processing (LNCS), vol.3817, pp.242-256, 2005, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-31257-4]

**[Gómez,2007-A]** Gómez, P.; Fernández, R. et al *"Detecting Pathology in the Glottal Spectral Signature of Female Voice"*, Models and Analysis of Vocal Emissions for Biomedical Applications : 5th International Workshop, MAVEBA, pp.183-186, 2007, Firenze, Italy. [ISBN:978-88-8453-673-3]

**[Gómez,2007-B]** Gómez, P.; Fernández, R. et al. *"Evaluation of Voice Pathology Based on the Estimation of Vocal Fold Biomechanical Parameters"*, Journal of Voice, vol.21, issue 4, pp.450-476, 2007, Elsevier. [ISSN:0892-1997]

**[Gómez,2008]** Gómez, P.; Álvarez, A. et al. *"A hybrid parameterization technique for speaker identification"*, The 16th European Signal Processing Conference (EUSIPCO), 2008, Laussanne, Switzerland. [ISSN 2219-5491]

**[Gómez,2009]** Gómez, P.; Fernández, R. et al. *"Glottal Source biometrical signature for voice pathology detection"*, Speech Communication, vol.51, issue 9, pp.759-781, 2009, Elsevier. [ISSN:0167-6393]

**[Gómez,2010]** Gómez, P.; Ferrández-Vicente, J. et al. *"Neuromorphic Speech Processing: Objectives and Methods"*, Machine Audition: Principles, Algorithms and Systems, pp.447-473, 2010, IGI GLobal. [ISBN-10:16-152-0919-0 | ISBN-13:978-16-152-0919-4]

**[Gonzalez-Rodriguez,2007]** Gonzalez-Rodriguez, J.; Ramos, D. et al. *"Speaker Recognition The A TVS-UAM System at NIST SRE 05"*, IEEE Aerospace and Electronic Systems Magazine, vol.22, issue 1, pp.15-21, 2007, IEEE Computer Society. [ISSN:0885-8985]

**[Grabowski,2006]** Grabowski, K.; Sankowski, W. et al. *"Reliable Iris Localization Method with Application to Iris Recognition in Near Infrared Light"*, Proceedings of the International Conference on Mixed Design of Integrated Circuits and Systems, MIXDES, pp.684 - 687, 2006, Gdynia, Poland. [ISBN:83-922632-2-7]

**[Gray,1984]** Gray, R.M. *"Vector Quantization"*, IEEE ASSP Magazine, vol.1, issue 2, pp.4-29, 1984, IEEE Computer Society. [ISSN:0740-7467]

**[Griffiths,1977]** Griffiths, L.J. *"A continuously-adaptive filter implemented as a lattice structure"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.683-686, 1977, Hartford, CT, USA.

**[Gross,2001]** Gross, R. and Shi, J. *"The CMU Motion of Body (MoBo) Database"*, Technical Report, CMU-RI-TR-01-18, pp.1-11, 2001, Carnegie Mellon University, Robotics Institute, Pittsburgh, PA,USA.

**[Gross,2005]** Gross, R. *"Face Databases"*, Handbook of Face Recognition, pp.301-328, 2005, Springer New York. [ISBN:978-0-387-40595-7]

**[Gu,2005]** Gu, H.Y.; Zhuang, Y.T. and Pan, Y.H. *"An iris recognition method based on multi-orientation features and Non-symmetrical SVM"*, Journal of Zhejiang University - Science A, vol.6, issue 5, pp.428-432, 2005, Zhejiang University Press. [ISSN:1009-3095]

**[Gudnason,2008]** Gudnason, J. and Brookes, M. *"Voice source cepstrum coefficients for speaker identification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.4821-4824, 2008, Las Vegas, NV, USA. [ISSN:1520-6149 | ISBN:978-1-4244-1483-3]

**[Guo,2008]** Guo, Y. and Tian, G. *"Gait recognition based on anatomical knowledge"*, 7th World Congress on Intelligent Control and Automation, WCICA, pp.6803-6806, 2008, Chongqing, China. [ISBN:978-1-4244-2113-8]

**[Guru,2009]** Guru, D.S.; Prakash, H.N. and Manjunath, S. *"On-line Signature Verification: An Approach Based on Cluster Representations of Global Features"*, Proceedings of the Seventh International Conference on Advances in Pattern Recognition, ICAPR, pp.209-212, 2009, Kolkata, India. [ISBN:978-1-4244-3335-3]

**[Han,1995]** Han, K. and Sethi, I.K. *"Signature identification via local association of features"*, Proceedings of the Third International Conference on Document Analysis and Recognition, vol.1, pp.187-190, 1995, Montreal, Canada. [ISBN:0-8186-7128-9]

**[Harris,1978]** Harris, F.J. *"On the use of windows for harmonic analysis with the discrete Fourier transform"*, Proceedings of the IEEE, vol.66, issue 1, pp.51-83, 1978, IEEE Computer Society. [ISSN:0018-9219]

**[Hashimoto,2006]** Hashimoto, J. *"Finger Vein Authentication Technology and Its Future"*, Symposium on VLSI Circuits, Digest of Technical Papers., pp.5-8, 2006, Honolulu, HI, USA. [ISBN:1-4244-0006-6]

**[Hatch,2005]** Hatch, A.O.; Peskin, B. and Stolcke, A. *"Improved Phonetic Speaker Recognition Using Lattice Decoding"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.169-172, 2005, Philadelphia, PA, USA. [ISSN:1520-6149 | ISBN:0-7803-8874-7]

**[Hatch,2006-A]** Hatch, A.O.; Kajarekar, S. and Stolche, A. *"Within-Class Covariance Normalization for SVM-based Speaker Recognition"*, Proceedings of the 9th International Conference on Spoken Language Processing, pp.1471-1474, 2006, Pittsburgh, PA, USA.

**[Hatch,2006-B]** Hatch, A.O. and Stolcke, A. *"Generalized Linear Kernels for One-Versus-All Classification: Application to Speaker Recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.5, pp.585-588, 2006, Toulouse, France. [ISSN:1520-6149 | ISBN:1-4244-0469-X]

**[Hebert,2005]** Hebert, M.; Boies, D. and Nuance Communications *"T-Norm for Text-Dependent Commercial Speaker Verification Applications: Effect of Lexical Mismatch"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP,

vol.1, pp.729-732, 2005, Philadelphia, PA, USA. [ISSN:1520-6149 | ISBN:0-7803-8874-7]

**[Heck,1997]** Heck, L.P. and Weintraub, M. *"Handset-dependent background models for robust text-independent speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.1071-1074, 1997, Munich, Bavaria, Germany. [ISSN:1520-6149 | ISBN:0-8186-7919-0]

**[Hennebert,2000]** Hennebert, J.; Melin, H. et al. *"POLYCOST: A telephone-speech database for speaker recognition"*, Speech Communication, vol.31, issue 2-3, pp.265-270, 2000, Elsevier. [ISSN:0167-6393]

**[Herbst,2008]** Herbst, G. and Bocklisch, S.R. *"Classification of keystroke dynamics - a case study of fuzzified discrete event handling"*, 9th International Workshop on Discrete Event Systems, WODES, pp.394-399, 2008, Gothenburg, Sweden. [ISBN:978-1-4244-2592-1]

**[Hermansky,1994]** Hermansky, H. and Morgan, N. *"RASTA processing of speech"*, IEEE Transactions on Speech and Audio Processing, vol.2, issue 4, pp.578-589, 1994, IEEE Computer Society. [ISSN:1063-6676]

**[Hill,1998]** Hill, R. *"Retina Identification"*, Biometrics, Personal Identification in Networked Society, issue 6, pp.123-142, 1998, Springer-Verlag. [ISBN-10:0387285393 | ISBN-13:978-0387285399]

**[Hoi,2004]** Hoi, C.H. and Lyu, M.R. *"Robust face recognition using minimax probability machine"*, IEEE International Conference on Multimedia and Expo, ICME, vol.2, pp.1175-1178, 2004, Taipei, China. [ISBN:0-7803-8603-5]

**[Hong,2007]** Hong, S.; Lee, H. et al. *"Human identification based on gait analysis"*, International Conference on Control, Automation and Systems, ICCAS, pp.2234-2237, 2007, Seoul, South Korea. [ISBN:978-89-950038-6-2]

**[Hong,2008]** Hong, S.; Lee, H. et al. *"Fusion of multiple gait features for human identification"*, International Conference on Control, Automation and Systems, ICCAS, pp.2121-2125, 2008, Seoul, South Korea. [ISBN:978-89-950038-9-3]

**[Hosseinzadeh,2008]** Hosseinzadeh, D. and Krishnan,S. *"Gaussian Mixture Modeling of Keystroke Patterns for Biometric Applications"*, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.38, issue 6, pp.816-826, 2008, IEEE Computer Society. [ISSN:1094-6977]

**[Hotelling,1933]** Hotelling, H. *"Analysis of a complex of statistical variables into principal components"*, Journal of Educational Psychology, vol.24, issue 6, pp.417-441, 1933, Warwick & York.

**[Hou,2003]** Hou, F. and Wang, B. *"Text-independent speaker recognition using probabilistic SVM with GMM adjustment"*, Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, pp.305-308, 2003, Beijing, China. [ISBN:0-7803-7902-0]

**[Hu,2008]** Hu, J.; Gingrich, D. and Sentosa, A. *"A k-Nearest Neighbor Approach for User Authentication through Biometric Keystroke Dynamics"*, IEEE International Conference on Communications, ICC, pp.1556-1560, 2008, Beijing, China. [ISBN:978-1-4244-2075-9]

**[Huang,1989]** Huang, X.D. and Jack, M.A. *"Semi-continuous hidden Markov models for speech signals"*, Computer Speech and Language, vol.3, issue 3, pp.239-251, 1989, Elsevier. [ISSN:0885-2308]

**[Huang,2002]** Huang, Y.P.; Luo, S.W. and Chen, E.Y. *"An efficient iris recognition system"*, Proceedings of 2002 International Conference on Machine Learning and Cybernetics, vol.1, pp.450-454, 2002, Beijing, China. [ISBN:0-7803-7508-4]

**[Huang,2007]** Huang, D.; Wang, Y.H. and Wang, Y.D. *"A robust infrared face recognition method based on adaboost gabor features"*, International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR, vol.3, pp.1114-1118, 2007, Beijing, China. [ISBN:78-1-4244-1065-1]

**[Hurley,2000]** Hurley, D.J.; Nixon, M.S. and Carter, J.N. *"Automatic ear recognition by force field transformations"*, IEE Colloquium on Visual Biometrics, vol.7, pp.1-5, 2000, London, UK.

**[Hurley,2005]** Hurley, D.J.; Nixon, M.S. and Carter, J.N. *"Force field feature extraction for ear biometrics"*, Computer Vision and Image Understanding, vol.98, issue 3, pp.491-512, 2005, Elsevier. [ISSN:1077-3142]

**[Husken,2005]** Husken, M.; Brauckmann, M. et al. *"Strategies and Benefits of Fusion of 2D and 3D Face Recognition"*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, pp.174-181, 2005, San Diego, CA, USA. [ISSN:1063-6919 | ISBN:0-7695-2372-2]

**[Iannarelli,1989]** Iannarelli, A.V. *"Ear Identification"*, pp.213, 1989, Paramont Publishing Company. [ISSBN-10:0962317802 | ISBN-13:978-096-2317-80-4]

**[Ichino,2006]** Ichino, M.; Sakano, H. and Komatsu, N. *"Multimodal Biometrics of Lip Movements and Voice using Kernel Fisher Discriminant Analysis"*, 9th International Conference on Control, Automation, Robotics and Vision, ICARCV, pp.1-6, 2006, Singapore. [ISBN:1-4244-0341-3]

**[Igarza,2003]** Igarza, J.; Goirizelaia, I. et al. *"Online Handwritten Signature Verification Using Hidden Markov Models"*, Progress in Pattern Recognition, Speech and Image Analysis (LNCS), vol.2905, pp.391-399, 2003, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-20590-6]

**[Ilyas,2007]** Ilyas, M.Z.; Samad, S.A. et al. *"Speaker Verification using Vector Quantization and Hidden Markov Model"*, 5th Student Conference on Research and Development, SCOReD, pp.1-5, 2007, Selangor, Malaysia. [ISBN:978-1-4244-1469-7]

**[Im,2001]** Im,S.; Park,H. et al. *"An Biometric Identification System by Extracting Hand Vein Patterns"*, Journal of the Korean Physical Society, vol.38, issue 3, pp.268-272, 2001, Korean Physical Society.

**[Ishizaka,1972]** Ishizaka, K. and Flanagan, J.L. *"Synthesis of voiced sounds from a two-mass model of the vocal cords"*, Bell System Technical Journal, vol.51, issue 6, pp.1233-1268, 1972, Blackwell Publishing Ltd. [ISSN:538-7305]

**[Itakura,1970]** Itakura, F. and Saito, S. *"A statistical method for estimation of speech spectral density and formant frequencies"*, Electronics and Communications in Japan (ECJ), vol.53-A, issue 1, pp.36-43, 1970 .

**[Itakura,1975]** Itakura, F. *"Line spectrum representation of linear predictor coefficients of speech signals"*, The Journal of the Acoustical Society of America, vol.57, issue S1, pp.S35, 1975, Acoustic Society of America. [ISSN:0001-4966]

**[Jaakkola,1998]** Jaakkola, T.S. and Haussler,D. *"Exploiting generative models in discriminative classifiers"*, Advances in Neural Information Processing Systems (NIPS), vol.11, pp.487-493, 1998, Denver, CO, USA. [ISBN:0-262-11245-0]

**[Jain,1998]** Jain, A.K.; Bolle, R. and Pankanti, S. *"Introduction to biometrics"*, Biometrics, Personal Identification in Networked Society, issue 1, pp.1-43, 1998, Springer-Verlag. [ISBN-10:0387285393 | ISBN-13:978-0387285399]

**[Jain,1999-A]** Jain, A.K.; Prabhakar, S. et al. *"FingerCode: a filterbank for fingerprint representation and matching"*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, vol.2, pp.187-193, 1999, Ft. Collins, CO, USA. [ISSN:1063-6919 | ISBN:0-7695-0149-4]

**[Jain,1999-B]** Jain, A.K.; Ross, A. and Pankanti, S. *"A Prototype Hand Geometry-based Verification System"*, Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA), pp.166-171, 1999, Washington, DC, USA.

**[Jain,2004-A]** Jain, A.K. *"Biometric recognition: how do I know who you are?"*, Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, pp.3-5, 2004, Kusadasi, Turkey. [ISBN:0-7803-8318-4]

**[Jain,2004-B]** Jain, A.K.; Ross, A. and Prabhakar, S. *"An introduction to biometric recognition"*, IEEE Transactions on Circuits and Systems for Video Technology, vol.14, issue 1, pp.4-20, 2004, IEEE Computer Society. [ISSN:1051-8215]

**[Jain,2008]** Jain, A.K.; Flynn, P. and Ross, A.A. *"Handbook on biometrics"*, pp.568, 2008, Springer. [ISBN-10:0387222960 | ISBN-13:978-0387222967]

**[Jankowski,1995]** Jankowski, C.R.,Jr.; Quatieri, T.F. and Reynolds, D.A. *"Measuring fine structure in speech: application to speaker identification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.325-328, 1995, Detroit, MI, USA. [ISSN:1520-6149 | ISBN:0-7803-2431-5]

**[Jeff Wu,1983]** Jeff Wu, C.F. *"On the Convergence Properties of the EM Algortihm"*, The Annals of Statistics, vol.11, issue 1, pp.95-103, 1983, The Institute of Mathematical Statistics. [ISSN:0090-5364]

**[Jin,2008]** Jin, Z; Teoh, A.B.J. et al. *"Typing dynamics biometric authentication through fuzzy logic"*, International Symposium on Information Technology, ITSim, vol.3, pp.1-6, 2008, Kuala Lupmur,Malaysia. [ISBN:978-1-4244-2327-9]

**[Joachims,1998]** Joachims, T. *"Making large-scale SVM learning practical"*, Advances in Kernel Methods - Support Vector Learning, issue 11, pp.164-184, 1998, The MIT Press, Cambridge, MA, USA. [ISBN-10:0-262-19416-3 | ISBN-13:978-0262194167]

**[Johansson,1975]** Johansson, G. *"Visual motion perception"*, Scientific American Magazine, vol.232, issue 6, pp.76-88, 1975, Scientific American. [ISSN:0036-8733]

**[Jolliffe,2002]** Jolliffe, I.T. *"Principal component analysis"*, 2nd Edition, pp.489, 2002, Springer. [ISBN-10:0387954422 | ISBN-13:978-0387954424]

**[Jones,2006]** Jones, P. *"Banking on vein at the ATM"*, Biometric Technology Today, vol.14, issue 5, pp.8-9, 2006, Elsevier. [ISSN:0969-4765]

**[Juang,1986]** Juang, B. and Rabiner, L. *"Mixture autoregressive hidden Markov models for speaker independent isolated word recognition"*, IEEE International Conference on

Acoustics, Speech and Signal Processing, ICASSP, vol.11, pp.41-44, 1986, Tokio, Japan.

**[Juang,1992]** Juang, B.H. and Katagiri, S. *"Discriminative learning for minimum error classification [pattern recognition]"*, IEEE Transactions on Signal Processing, vol.40, issue 12, pp.3043-3054, 1992, IEEE Computer Society. [ISSN:1053-587X]

**[Kabal,1986]** Kabal, P. and Ramachandran, R.P. *"The computation of line spectral frequencies using Chebyshev polynomials"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.34, issue 6, pp.1419-1426, 1986, IEEE Computer Society. [ISSN:0096-3518]

**[Kailath,1974]** Kailath, T. *"A view of three decades of linear filtering theory"*, IEEE Transactions on Information Theory, vol.20, issue 2, pp.146-181, 1974, IEEE Computer Society. [ISSN:0018-9448]

**[Kajarekar,2003]** Kajarekar, S.; Ferrer, L. et al. *"Speaker recognition using prosodic and lexical features"*, IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU, pp.19-24, 2003, Las Vegas, NV, USA. [ISBN:0-7803-7980-2]

**[Kajarekar,2007]** Kajarekar, S.S. and Stolche, A. *"NAP and WCCN: Comparison of Approaches using MLLR-SVM Speaker Verification System"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.4, pp.249-252, 2007, Honolulu, HI, USA. [ISSN:1520-6149 | ISBN:1-4244-0727-3]

**[Kale,2004-A]** Kale, A.; Sundaresan, A. et al. *"Identification of humans using gait"*, IEEE Transactions on Image Processing, vol.13, issue 9, pp.1163-1173, 2004, IEEE Computer Society. [ISSN:1057-7149]

**[Kale,2004-B]** Kale, A.; Sundaresan, A. et al. *"Gait-based human identification from a monocular video sequence"*, Handbook on Pattern Recognition and Computer Vision (3rd Edition), pp.411-429, 2004, World Scientific Publishing Company Pvt. Ltd., In. [ISBN:978-981-4481-31-1]

**[Karam,2007]** Karam, Z.N. and Campbell, W.M. *"A new Kernel for SVM MLLR Based Speaker Recognition"*, Proceedings of the 8th Annual Conference of the International Speech Communication Association, pp.290-293, 2007, Antwerp, Belgium.

**[Karayiannis,1995]** Karayiannis, N.B. and Pai, P. *"Fuzzy vector quantization algorithms and their application in image compression"*, IEEE Transactions on Image Processing, vol.4, issue 9, pp.1193-1201, 1995, IEEE Computer Society. [ISSN:1057-7149]

**[Kasprzak,2003]** Kasprzak, J. *"Polish Methods of Earprint Identification"*, The Information Bulletin for Shoeprint/Toolmark Examiners (IBSTE), vol.9, issue 3, pp.20-22, 2003, National Bureau of Investigation, Finland. [ISSN 1455-4194]

**[Kass,1995]** Kass, R.E. and Raftery, A.E. *"Bayes Factors"*, Journal of the American Statistical Association, vol.90, issue 430, pp.773-795, 1995, American Statistical Association. [ISSN:0162-1459]

**[Keller,1999]** Keller, P.E. *"Overview of electronic nose algorithms"*, International Joint Conference on Neural Networks, IJCNN, vol.1, pp.309-312, 1999, Washington, DC, USA. [ISSN:1098-7576 | ISBN:0-7803-5529-6]

**[Kenny,2005-A]** Kenny, P. *"Joint factor analysis of speaker and session variability: Theory and algorithms"*, Technical Report, CRIM-06/08-14, pp.1-17, 2005, CRIM, Montreal, Quebec, Canada.

**[Kenny,2005-B]** Kenny, P.; Boulianne, G. and Dumouchel, P. *"Eigenvoice modeling with sparse training data"*, IEEE Transactions on Speech and Audio Processing, vol.13, issue 3, pp.345-354, 2005, IEEE Computer Society. [ISSN:1063-6676]

**[Kenny,2005-C]** Kenny, P.; Boulianne, G. et al. *"Factor Analysis Simplified"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.637-640, 2005, Philadelphia, PA, USA. [ISSN:1520-6149 | ISBN:0-7803-8874-7]

**[Kenny,2008]** Kenny, P.; Ouellet, P. et al. *"A Study of Interspeaker Variability in Speaker Verification"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.16, issue 5, pp.980-988, 2008, IEEE Computer Society. [ISSN:1558-7916]

**[Kersta,1962]** Kersta, L.G. *"Voiceprint Identification"*, Nature, vol.196, issue 4861, pp.1253-1257, 1962.

**[Kholmatov,2005]** Kholmatov, A. and Yanikoglu, B. *"Identity authentication using improved online signature verification method"*, Pattern Recognition Letters, vol.26, issue 15, pp.2400-2408, 2005, Elsevier. [ISSN:0167-8655]

**[Kholmatov,2009]** Kholmatov, A. and Yanikoglu, B. *"SUSIG: an on-line signature database, associated protocols and benchmark results"*, Pattern Analysis & Applications, vol.12, issue 3, pp.227-236, 2009, Springer-Verlag. [ISSN:1433-7541]

**[Khoury,2013]** Khoury, E.; Vesnicer, B. et al. *"The 2013 Speaker Recognition Evaluation in Mobile Environment"*, International Conference on Biometrics (ICB), pp.1-8, 2013, Madrid, Spain. [ISBN:978-1-4799-0310-8]

**[Kim,1998]** Kim, N.S. and Un, C. *"Deleted strategy for MMI-based HMM training"*, IEEE Transactions on Speech and Audio Processing, vol.6, issue 3, pp.299-303, 1998, IEEE Computer Society. [ISSN:1063-6676]

**[Kim,2004-A]** Kim, J.; Ko, D.Y. and Na, S.N. *"Implementation and enhancement of GMM face recognition systems using flatness measure"*, 13th IEEE International Workshop on Robot and Human Interactive Communication, ROMAN, pp.247-251, 2004, Kurashiki, Okayama Japan. [ISBN:0-7803-8570-5]

**[Kim,2004-B]** Kim, S.; Eriksson, T. et al. *"A pitch synchronous feature extraction method for speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.405-408, 2004, Montreal, Canada. [ISSN:1520-6149 | ISBN:0-7803-8484-9]

**[Kim,2007]** Kim, S.; Kim, M. and Yu, H. *"Speaker Recognition Using Temporal Decomposition of LSF for Mobile Environment"*, Embedded Software and Systems (LNCS), vol.4523, pp.338-346, 2007, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-72684-5]

**[Kim,2008]** Kim, D.J. and Hong, K.S. *"Multimodal biometric authentication using teeth image and voice in mobile environment"*, IEEE Transactions on Consumer Electronics, vol.54, issue 4, pp.1790-1797, 2008, IEEE Computer Society. [ISSN:0098-3063]

**[Kinnunen,2009]** Kinnunen, T. and Alku, P. *"On separating glottal source and vocal tract information in telephony speaker verification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.4545-4548, 2009, Taipei, Taiwan. [ISSN:1520-6149 | ISBN:978-1-4244-2353-8]

**[Kinnunen,2010]** Kinnunen, T. and Li, H. *"An overview of text-independent speaker recognition: From features to supervectors"*, Speech Communication, vol.52, issue 1, pp.12-40, 2010, Elsevier. [ISSN:0167-6393]

**[Kiran,2001]** Kiran, G.V.; Kunte, R.S.R. and Samuel, S. *"On-line signature verification system using probabilistic feature modelling"*, 6th International Symposium on Signal Processing and its Applications, vol.1, pp.355-358, 2001, Kuala Lupmur,Malaysia. [ISBN:0-7803-6703-0]

**[Ko,2006]** Ko, J.G.; Gil, Y.H. and Yoo, J.H. *"Iris Recognition using Cumulative SUM based Change Analysis"*, International Symposium on Intelligent Signal Processing and Communications, ISPACS, pp.275-278, 2006, Yonago, Japan. [ISBN:0-7803-9732-0]

**[Kohler,2001]** Kohler, M.A.; Andrews, W.D. et al. *"Phonetic speaker recognition"*, Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, vol.2, pp.1557-1561, 2001, Pacific Grove, CA, USA. [ISSN:1058-6393 | ISBN:0-7803-7147-X]

**[Kohonen,1996]** Kohonen, T.; Hynninen, J. et al. *"LVQ_PAK: The Learning Vector Quantization Program Package"*, Technical Report A30, pp.1-28, 1996, Helsinki University of Technology, Laboratory of Computer and Information Science, Finland. [ISSN:0783-7445 | ISBN:951-22-2948-X]

**[Kong,2001]** Kong, W.K. and Zhang, D. *"Accurate iris segmentation based on novel reflection and eyelash detection model"*, Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, ISIMP, pp.263-266, 2001, Hong Kong, China. [ISBN:962-85766-2-3]

**[Kong,2007]** Kong, S.G.; Heo, J. et al. *"Multiscale Fusion of Visible and Thermal IR Images for Illumination-Invariant Face Recognition"*, International Journal of Computer Vision, vol.71, issue 2, pp.215-233, 2007, Kluwer Academic Publishers. [ISSN:0920-5691]

**[Konig,1998]** Konig, Y.; Heck, L. et al. *"Nonlinear discriminant feature extraction for robust text-independent speaker recognition"*, In Proc. RLA2CESCA Speaker Recognition and its Commercial and Forensic Applications, pp.72-75, 1998, Avignon, France.

**[Korotkaya,2003]** Korotkaya, Z. *"Biometrics Person Authentication: Odor"*, Technical Report, pp.1-16, 2003, Department of Information Technology, Laboratory of Applied Mathematics, Lappeenranta University of Technology. [on-line http://www2.it.lut.fi/kurssit/03-04/010970000/seminars/Korotkaya.pdf 2014]

**[Kressel,1999]** Kressel, U. *"Pairwise classification and support vector machines"*, Advances in Kernel Methods - Support Vector Learning, issue 15, pp.255-268, 1999, The MIT Press, Cambridge, MA, USA. [ISBN-10:0-262-19416-3 | ISBN-13:978-0262194167]

**[Krishna,2008]** Krishna, B.G. and Sreenivas, T.V. *"A comparative study of speaker adaptation methods"*, IEEE Region 10 Conference TENCON, pp.1-4, 2008, Hyderabad, India. [ISBN:978-1-4244-2408-5]

**[Kumar,2003]** Kumar, A.; Wong, D.C.M. et al. *"Personal Verification Using Palmprint and Hand Geometry Biometric"*, Audio- and Video-Based Biometric Person Authentication (LNCS), vol.2688, pp.668-678, 2003, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-40302-9]

**[Lanitis,2010]** Lanitis, A. *"A survey of the effects of aging on biometric identity verification"*, International Journal of Biometrics (IJBM), vol.2, issue 1, pp.34-52, 2010, Inderscience Publishers, Geneva, SWITZERLAND. [ISSN:1755-8301]

**[Lee,2002]** Lee, L. and Grimson, W.E.L. *"Gait analysis for recognition and classification"*, Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition, pp.148-155, 2002, Washington, DC, USA. [ISBN:0-7695-1602-5]

**[Lee,2006]** Lee, K.H.; Min,S.Y. et al. *"An Improvement of the Processing Delay for the G.723.1 Vocoder"*, Advances in Multimedia Modeling (LNCS), vol.4352, pp.568-575, 2006, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-69428-1]

**[Leggetter,1995-A]** Leggetter, C.J. and Woodland, P.C. *"Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models"*, Computer Speech and Language, vol.9, issue 2, pp.171-185, 1995, Elsevier. [ISSN:0885-2308]

**[Leggetter,1995-B]** Leggetter, C.J. and Woodland, P.C. *"Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression"*, Proceedings of the ARPA Spoken Language Technology Workshop, pp.110-115, 1995, Austin, TX, USA.

**[Lei,2006]** Lei, J. and Lu, C. *"Face Recognition by Spatiotemporal ICA Using Facial Database Collected by AcSys FRS Discover System"*, 7th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, SNPD, pp.218-224, 2006, Las Vegas, NV, USA. [ISBN:0-7695-2611-X]

**[Levinson,1986]** Levinson, S.E. *"Continuously variable duration hidden Markov models for speech analysis"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.11, pp.1241-1244, 1986, Tokio, Japan.

**[Li,1988]** Li, K.P. and Porter, J. E. *"Normalizations and selection of speech segments for speaker recognition scoring"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.595-598, 1988, New York, NY, USA. [ISSN:1520-6149]

**[Li,2006]** Li, B.; Zhang, D. and Wang, K. *"Online signature verification based on null component analysis and principal component analysis"*, Pattern Analysis & Applications, vol.8, issue 4, pp.345-356, 2006, Springer-Verlag. [ISSN:1433-7541]

**[Li,2009]** Li, H.; Ma, B. et al. *"The I4U system in NIST 2008 speaker recognition evaluation"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.4201-4204, 2009, Taipei, Taiwan. [ISSN:1520-6149 | ISBN:978-1-4244-2353-8]

**[Lin,2004]** Lin, C.L. and Fan,K.C. *"Biometric verification using thermal images of palm-dorsa vein patterns"*, IEEE Transactions on Circuits and Systems for Video Technology, vol.14, issue 2, pp.199-213, 2004, IEEE Computer Society. [ISSN:1051-8215]

**[Lin,2006]** Lin, T.C.; Chen, S.H. et al. *"Speaker Classification Using Support Vector Machine and Wavelets"*, IEEE Region 10 Conference TENCON, pp.1-4, 2006, Hyderabad, India. [ISBN:1-4244-0548-3]

**[Linde,1980]** Linde, Y.; Buzo, A. and Gray, R. *"An Algorithm for Vector Quantizer Design"*, IEEE Transactions on Communications, vol.28, issue 1, pp.84-95, 1980, IEEE Computer Society. [ISSN:0090-6778]

**[Liporace,1982]** Liporace, L. *"Maximum likelihood estimation for multivariate observations of Markov sources"*, IEEE Transactions on Information Theory, vol.28, issue 5, pp.729-734, 1982, IEEE Computer Society. [ISSN:0018-9448]

**[Liu,1990]** Liu, C.S; Wang, W.J. et al. *"Study of line spectrum pair frequencies for speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.277-280, 1990, Alburquerque, NM, USA. [ISSN:1520-6149]

**[Liu,1995]** Liu, C.S.; Lee, C.H. et al. *"A study on minimum error discriminative training for speaker recognition"*, The Journal of the Acoustical Society of America, vol.97, issue 1, pp.637-648, 1995, Acoustic Society of America. [ISSN:0001-4966]

**[Liu,2003]** Liu, C. and Wechsler, H. *"Independent component analysis of Gabor features for face recognition"*, IEEE Transactions on Neural Networks, vol.14, issue 4, pp.919-928, 2003, IEEE Computer Society. [ISSN:1045-9227]

**[Liu,2004]** Liu, Q.; Cheng, J. et al. *"Modeling face appearance with nonlinear independent component analysis"*, Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, pp.761-766, 2004, Washington, DC, USA. [ISBN:0-7695-2122-3]

**[Liu,2005]** Liu, X.; Bowyer, K.W. and Flynn, P.J. *"Experimental Evaluation of Iris Recognition"*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, pp.158-158, 2005, San Diego, CA, USA. [ISSN:1063-6919 | ISBN:0-7695-2372-2]

**[Liu,2006-A]** Liu, M.; Xie, Y. et al. *"A New Hybrid GMM/SVM for Speaker Verification"*, Proceedings of the 18th International Conference on Pattern Recognition, vol.4, pp.314-317, 2006, Hong Kong, China. [ISSN:1051-4651 | ISBN:0-7695-2521-0]

**[Liu,2006-B]** Liu, Z. and Sarkar, S. *"Improved gait recognition by gait dynamics normalization"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.28, issue 6, pp.863-876, 2006, IEEE Computer Society. [ISSN:0162-8828]

**[Liu,2008]** Liu, N.N and Wang, Y.H. *"Fusion of global and local information for an on-line Signature Verification system"*, International Conference on Machine Learning and Cybernetics, vol.1, pp.57-61, 2008, Kunming, China. [ISBN:978-1-4244-2095-7]

**[Lloyd,1982]** Lloyd, S. *"Least squares quantization in PCM"*, IEEE Transactions on Information Theory, vol.28, issue 2, pp.129-137, 1982, IEEE Computer Society. [ISSN:0018-9448]

**[Longworth,2009]** Longworth, C. and Gales, M.J.F. *"Combining Derivative and Parametric Kernels for Speaker Verification"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.17, issue 4, pp.748-757, 2009, IEEE Computer Society. [ISSN:1558-7916]

**[Loy,2007]** Loy, C.C.; Lai, W.K. and Lim, C.P. *"Keystroke Patterns Classification Using the ARTMAP-FD Neural Network"*, Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP, vol.1, pp.61-64, 2007, Kaohsiung, China. [ISBN:978-0-7695-2994-1]

**[Lv,2006]** Lv, H.R. and Wang, W.Y. *"Biologic verification based on pressure sensor keyboards and classifier fusion techniques"*, IEEE Transactions on Consumer Electronics, vol.52, issue 3, pp.1057-1063, 2006, IEEE Computer Society. [ISSN:0098-3063]

**[Lv,2008]** Lv, H.R.; Lin, Z.L. et al. *"Emotion recognition based on pressure sensor keyboards"*, IEEE International Conference on Multimedia and Expo, pp.1089-1092, 2008, Hannover, Germany. [ISBN:978-1-4244-2570-9]

**[Ma,2003]** Ma, L.; Tan, T. et al. *"Personal identification based on iris texture analysis"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.25, issue 12, pp.1519-1533, 2003, IEEE Computer Society. [ISSN:0162-8828]

**[Mak,2006]** Mak, M.; Hsiao, R. and Mak, B. *"A Comparison of Various Adaptation Methods for Speaker Verification With Limited Enrollment Data"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.929-932, 2006, Toulouse, France. [ISSN:1520-6149 | ISBN:1-4244-0469-X]

**[Makhoul,1975]** Makhoul, J. *"Linear prediction: A tutorial review"*, Proceedings of the IEEE, vol.63, issue 4, pp.561-580, 1975, IEEE Computer Society. [ISSN:0018-9219]

**[Maltoni,2009]** Maltoni, D.; Maio, D. et al. *"Handbook of Fingerprint Recognition"*, 2nd Edition, pp.496, 2009, Springer. [ISBN-10:1848822537 | ISBN-13:978-1848822535]

**[Mariethoz,2001]** Mariethoz, J. and Bengio, S. *"A Comparative Study of Adaptation Methods for Speaker Verification"*, Proceedings of the 7th International Conference on Spoken Language Processing, pp.581-584, 2001, Denver, CO, USA.

**[Mariño,2006]** Mariño, C.; Penedo, M.G. et al. *"Personal authentication using digital retinal images"*, Pattern Analysis & Applications, vol.9, issue 1, pp.21-33, 2006, Springer-Verlag. [ISSN:1433-7541]

**[Matrouf,2007]** Matrouf, D.; Scheffer, N. et al. *"A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification"*, Proceedings of the 8th Annual Conference of the International Speech Communication Association, pp.1242-1245, 2007, Antwerp, Belgium.

**[Matsui,1992]** Matsui, T. and Furui, S. *"Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.157-160, 1992, San Francisco, CA, USA. [ISSN:1520-6149 | ISBN:0-7803-0532-9]

**[Matsumoto,2002]** Matsumoto, T.; Matsumoto, H. et al. *"Impact of artificial 'gummy' fingers on fingerprint systems"*, Proceedings of the SPIE, Optical Security and Counterfeit Deterrence Techniques IV, vol.4677, issue 1, pp.275-289, 2002, SPIE.

**[McCool,2012]** McCool, C.; Marcel, S. et al. *"Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data"*, IEEE International Conference on Multimedia and Expo Workshops, ICMEW, pp.635-640, 2012, Melbourne, VIC, Australia. [ISBN:978-1-4673-2027-6]

**[Melin,2005]** Melin, P. and Castillo, O. *"Human Recognition using Face, Fingerprint and Voice"*, Hybrid Intelligent Systems for Pattern Recognition Using Soft Computing - Studies in Fuzziness and Soft Computing, vol.172, pp.241-256, 2005, Springer Berlin Heidelberg. [ISSN:1434-9922 | ISBN:978-3-540-24121-8]

**[Mercer,1909]** Mercer, J. *"Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations"*, Philosophical Transactions of the Royal Society of London.Series A, Containing Papers of a Mathematical or Physical Character, vol.209, pp.415-446, 1909, The Royal Society. [ISSN:0264-3952]

**[Messer,1999]** Messer, K.; Matas, J. et al. *"XM2VTSDB: The Extended M2VTS Database"*, Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA), pp.72-77, 1999, Washington, DC, USA.

**[Middleton,2005]** Middleton, L.; Buss, A.A. et al. *"A floor sensor system for gait recognition"*, IEEE Workshop on Automatic Identification Advanced Technologies, pp.171-176, 2005, Buffalo, NY, USA. [ISBN:0-7695-2475-3]

**[Mittal,2008]** Mittal, N. and Walia, E. *"Face Recognition Using Improved Fast PCA Algorithm"*, Congress on Image and Signal Processing, CISP, vol.1, pp.554-558, 2008, Sanya, Hainan, China. [ISBN:978-0-7695-3119-9]

**[Miura,2004]** Miura, N.; Nagasaka, A. and Miyatake, T. *"Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification"*, Machine Vision and Applications, vol.15, issue 4, pp.194-203, 2004, Springer-Verlag. [ISSN:0932-8092]

**[Miyazawa,2005]** Miyazawa,K.; Ito,K. et al. *"A Phase-Based Iris Recognition Algorithm"*, Advances in Biometrics (LNCS), vol.3832, pp.356-365, 2005, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-31111-9]

**[Miyazawa,2006]** Miyazawa, K.; Ito, K. et al. *"An Iris Recognition System Using Phase-Based Image Matching"*, IEEE International Conference on Image Processing, pp.325-328, 2006, Atlanta, GA, USA. [ISSN:1522-4880 | ISBN:1-4244-0480-0]

**[Monro,2008]** Monro,D. *"University of Bath Iris Image Database"*, University of Bath, Bath, UK. [on-line acc. 2008 http://www.bath.ac.uk/elec-eng/research/sipg/irisweb/]

**[Monrose,1997]** Monrose, F. and Rubin, A. *"Authentication via keystroke dynamics"*, Proceedings of the 4th ACM conference on Computer and communications security, CCS, pp.48-56, 1997, Zurich, Switzerland. [ISBN:0-89791-912-2]

**[Monrose,2000]** Monrose, F. and Rubin, A.D. *"Keystroke dynamics as a biometric for authentication"*, Future Generation Computer Systems, vol.16, issue 4, pp.351-359, 2000, Elsevier. [ISSN:0167-739X]

**[Monwar,2008]** Monwar, M. and Gavrilova, M. *"FES: A System for Combining Face, Ear and Signature Biometrics Using Rank Level Fusion"*, 5th International Conference on Information Technology: New Generations, ITNG, pp.922-927, 2008, Las Vegas, NV, USA. [ISBN:0-7695-3099-0]

**[Moreno,1999]** Moreno, B.; Sanchez, A. and Velez, J.F. *"On the use of outer ear images for personal identification in security applications"*, Proceedings of the IEEE 33rd International Carnahan Conference on Security Technology, pp.469-476, 1999, Madrid, Spain. [ISBN:0-7803-5247-5]

**[Mu,2005]** Mu, Z.; Yuan,L. et al. *"Shape and Structural Feature Based Ear Recognition"*, Advances in Biometric Person Authentication (LNCS), vol.3338, pp.663-670, 2005, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-24029-7]

**[Muller,2001]** Muller, K.; Mika, S. et al. *"An introduction to kernel-based learning algorithms"*, IEEE Transactions on Neural Networks, vol.12, issue 2, pp.181-201, 2001, IEEE Computer Society. [ISSN:1045-9227]

**[Mulyono,2008]** Mulyono, D. and Jinn, H.S. *"A study of finger vein biometric for personal identification"*, International Symposium on Biometrics and Security Technologies, ISBAST, pp.1-8, 2008, Islamabad, Pakistan. [ISBN:978-1-4244-2427-6]

**[Muñoz,2010]** Muñoz, C.; Martínez, R. et al. *"Discriminación de género basada en nuevos parametros MFCC"*, 1er Workshop de Tecnologías Multibiométricas para la identificación de personas (WTM-IP), vol.1, pp.22-24, 2010, Las Palmas de Gran Canaria, Spain. [ISBN:978-84-693-3389-1]

**[Muroi,2008]** Muroi, T.; Takiguchi, T. and Ariki, Y. *"Speaker Independent Phoneme Recognition Based on Fisher Weight Map"*, International Conference on Multimedia and Ubiquitous Engineering, MUE, pp.253-257, 2008, Busan,South Korea. [ISBN:978-0-7695-3134-2]

**[Murty,2006]** Murty, K.S.R. and Yegnanarayana, B. *"Combining evidence from residual phase and MFCC features for speaker recognition"*, IEEE Signal Processing Letters, vol.13, issue 1, pp.52-55, 2006, IEEE Computer Society. [ISSN:1070-9908]

**[Nalwa,1997]** Nalwa, V.S. *"Automatic on-line signature verification"*, Proceedings of the IEEE, vol.85, issue 2, pp.215-239, 1997, IEEE Computer Society. [ISSN:0018-9219]

**[Nandini,2008]** Nandini, C. and RaviKumar, C.N. *"An approach to gait recognition"*, International Symposium on Biometrics and Security Technologies, ISBAST, pp.1-3, 2008, Islamabad, Pakistan. [ISBN:978-1-4244-2427-6]

**[Naylor,2007]** Naylor, P.A.; Kounoudes, A. et al. *"Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 1, pp.34-43, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Nazeer,2007]** Nazeer, S.A.; Omar, N. and Khalid, M. *"Face Recognition System using Artificial Neural Networks Approach"*, International Conference on Signal Processing, Communications and Networking, ICSCN, pp.420-425, 2007, Chennai, India. [ISBN:1-4244-0997-7]

**[Nickel,2006]** Nickel, R.M. *"Feature - Automatic speech character identification"*, IEEE Circuits and Systems Magazine, vol.6, issue 4, pp.10-31, 2006, IEEE Computer Society. [ISSN:1531-636X]

**[NIST,2010]** NIST *"The NIST Year 2010 Speaker Recognition Evaluation Plan"*, pp.1-11, 2010. [on-line acc. 2014 http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf]

**[Nixon,2006]** Nixon, M.S. and Carter, J.N. *"Automatic Recognition by Gait"*, Proceedings of the IEEE, vol.94, issue 11, pp.2013-2024, 2006, IEEE Computer Society. [ISSN:0018-9219]

**[Obaidat,1997]** Obaidat, M.S. and Sadoun, B. *"Verification of computer users using keystroke dynamics"*, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol.27, issue 2, pp.261-269, 1997, IEEE Computer Society. [ISSN:1083-4419]

**[Oliver,1996]** Oliver, J.; Baxter, R. and Wallace, C. *"Unsupervised Learning Using MML"*, Proceedings of the Thirteenth International Conference In Machine Learning, ICML, pp.364-372, 1996, Bari, Italy. [ISBN:1-55860-419-7]

**[Onshaunjit,2008]** Onshaunjit, J. and Srinonchat,J. *"LSP Trajectory Analysis for Speech Recognition"*, Fifth International Conference on Computer Graphics, Imaging and Visualisation, CGIV, pp.276-279, 2008, Penang, Malaysia . [ISBN:978-0-7695-3359-9]

**[Ormoneit,1995]** Ormoneit, D. and Tresp,V. *"Improved Gaussian Mixture Density Estimates Using Bayesian Penalty Terms and Network Averaging"*, Advances in Neural Information Processing Systems (NIPS), vol.8, pp.542-548, 1995, Denver, CO, USA. [ISBN:0-262-20107-0]

**[Ortega,2006]** Ortega, M.; Marino, C. et al. *"Biometric authentication using digital retinal images"*, Proceedings of the 5th WSEAS International Conference on Applied Computer Science, pp.422-427, 2006, Hangzhou, China. [ISBN:960-8457-43-2]

**[Ortega-Garcia,2000]** Ortega-Garcia, J.; Gonzalez-Rodriguez, J. and Marrero-Aguiar, V. *"AHUMADA: A large speech corpus in Spanish for speaker characterization and identification"*, Speech Communication, vol.31, issue 2-3, pp.255-264, 2000, Elsevier. [ISSN:0167-6393]

**[Ortega-Garcia,2003]** Ortega-Garcia, J.; Fierrez-Aguilar, J. et al. *"MCYT baseline corpus: a bimodal biometric database"*, IEEE Proceedings Vision, Image and Signal Processing, vol.150, issue 6, pp.395-401, 2003, IEEE Computer Society. [ISSN:1350-245X]

**[Ostendorf,1996]** Ostendorf, M.; Digalakis, V.V. and Kimball, O.A. *"From HMM's to segment models: a unified view of stochastic modeling for speech recognition"*, IEEE Transactions on Speech and Audio Processing, vol.4, issue 5, pp.360-378, 1996, IEEE Computer Society. [ISSN:1063-6676]

**[Osuna,1997]** Osuna, E.; Freund, R. and Girosi, F. *"An improved training algorithm for support vector machines"*, Proceedings of the 1997 IEEE Workshop Neural Networks for Signal Processing, pp.276-285, 1997, Amelia Island, FL, USA. [ISSN:1089-3555 | ISBN:0-7803-4256-9]

**[Özgündüz,2005]** Özgündüz, E.; Sentürk, T. and Karsligil, M.E. *"Off-line Signature Verification and Recognition by SVM"*, The 13th European Signal Processing Conference (EUSIPCO), pp.113-116, 2005, Antalya, Turkey. [ISSN 2219-5491]

**[Padma Polash,2007]** Padma Polash,P. and Maruf Monwar,M. *"Human iris recognition for biometric identification"*, 10th International Conference on Computer and Information Iechnology, iccit, pp.1-5, 2007, Dhaka,Bangladesh. [ISBN:978-1-4244-1550-2]

**[Palka,2007]** Palka, S. and Hamilton, B.A. *"Fingerprint Readers: Vulnerabilities to Front- and Back- end Attacks"*, 1st IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS, pp.1-5, 2007, Crystal City, VA,USA. [ISBN:978-1-4244-1597-7]

**[Pandey,2010]** Pandey, B.; Ranjan, A. et al. *"Multilingual speaker recognition using ANFIS"*, 2nd International Conference on Signal Processing Systems, ICSPS, vol.3, pp.714-718, 2010, Dalian, China. [ISBN:978-1-4244-6892-8]

**[Park,2006]** Park, W.; Oh, J.C. et al. *"An open-set speaker identification system using genetic learning classifier system"*, Proceedings of the 8th annual conference on Genetic and evolutionary computation, GECCO, pp.1597-1598, 2006, New York, NY, USA. [ISBN:1-59593-186-4]

**[Parsa,2000]** Parsa, V. and Jamieson, D.G. *"Identification of Pathological Voices Using Glottal Noise Measures"*, Journal of Speech, Language, and Hearing Research, vol.43, issue 2, pp.469-485, 2000, American Speech-Language-Hearing Association. [ISSN:1092-4388]

**[Paul,1991]** Paul, D.B. *"The Lincoln tied-mixture HMM continuous speech recognizer"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.329-332, 1991, Toronto, Ontario, Canada. [ISSN:1520-6149 | ISBN:0-7803-0003-3]

**[Pelecanos,2000]** Pelecanos, J.; Myers, S. et al. *"Vector quantization based Gaussian modeling for speaker verification"*, Proceedings of the 15th International Conference on Pattern Recognition, vol.3, pp.294-297, 2000, Barcelona, Spain. [ISSN:1051-4651 | ISBN:0-7695-0750-6]

**[Pelecanos,2001]** Pelecanos, J.W. and Sridharan, S. *"Feature Warping for Robust Speech Verification"*, 2001: A Speaker Odyssey - The Speaker Recognition Workshop, pp.213-218, 2001, Crete, Greece.

**[Petrovska-Delacrétaz,2007]** Petrovska-Delacrétaz, D.; El Hannani, A. and Chollet, G. *"Text-Independent Speaker Verification: State of the Art and Challenges"*, Progress in Nonlinear Speech Processing (LNCS), vol.4391, pp.135-169, 2007, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-71503-0]

**[Phillips,2006]** Phillips, P.J.; Flynn, P. et al. *"Preliminary Face Recognition Grand Challenge Results"*, Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, FGR, pp.15-24, 2006, Southampton, UK. [ISBN:0-7695-2503-2]

**[Phillips,2007]** Phillips,J. *"ICE 2005 Dataset"*, NIST, USA. [on-line acc. 2007 http://iris.nist.gov/ice/ICE_Home.htm]

**[Phillips,2009]** Phillips, P.J.; Scruggs,W.T. et al. *"FRVT 2006 and ICE 2006 Large-Scale Experimental Results"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, issue 5, pp.831-846, 2009, IEEE Computer Society. [ISSN:0162-8828]

**[Picone,1993]** Picone, J.W. *"Signal modeling techniques in speech recognition"*, Proceedings of the IEEE, vol.81, issue 9, pp.1215-1247, 1993, IEEE Computer Society. [ISSN:0018-9219]

**[Pinto,2004]** Pinto, E.; Charlet, D. et al. “Development of new telephone speech databases for French : the NEOLOGOS Project”, 4th International Conference on Language Resources and Evaluation, LREC, pp.603-606, 2004, Lisbon, Portugal. [ISBN:2-9517408-1-6]

**[Platt,1998]** Platt, J.C. *"Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines"*, Technical Report, MSR-TR-98-14, pp.1-21, 1998, Microsoft Research.

**[Platt,1999-A]** Platt, J.C. *"Fast training of support vector machines using sequential minimal optimization"*, Advances in Kernel Methods - Support Vector Learning, issue

12, pp.185-208, 1999, The MIT Press, Cambridge, MA, USA. [ISBN-10:0-262-19416-3 | ISBN-13:978-0262194167]

**[Platt,1999-B]** Platt, J.C.; Cristianini, N. and Shawe-Taylor, J. *"Large margin DAGs for multiclass classification"*, Advances in Neural Information Processing Systems (NIPS), vol.12, pp.547-553, 1999, Denver, CO, USA. [ISBN:0-262-19450-3]

**[Plumpe,1999]** Plumpe, M.D.; Quatieri, T.F. and Reynolds, D.A. *"Modeling of the glottal flow derivative waveform with application to speaker identification"*, IEEE Transactions on Speech and Audio Processing, vol.7, issue 5, pp.569-586, 1999, IEEE Computer Society. [ISSN:1063-6676]

**[Polat,2008]** Polat, Ö. and Yildirim, T. *"Hand geometry identification without feature extraction by general regression neural network"*, Expert Systems with Applications: An International Journal, vol.34, issue 2, pp.845-849, 2008, Tarrytown, NY, USA. [ISSN:0957-4174]

**[Poritz,1982]** Poritz,A. *"Linear predictive hidden Markov models and the speech signal"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.7, pp.1291-1294, 1982, Paris, France.

**[Porwik,2007]** Porwik, P. and Para, T. *"Some Handwritten Signature Parameters in Biometric Recognition Process"*, 29th International Conference on Information Technology Interfaces, ITI, pp.185-190, 2007, Cavtat,Croatia. [ISSN:1330-1012 | ISBN:953-7138-10-0]

**[Price,1989]** Price, P.J. *"Male and female voice source characteristics: Inverse filtering results"*, Speech Communication, vol.8, issue 3, pp.261-277, 1989, Elsevier. [ISSN:0167-6393]

**[Pun,2004]** Pun, K.H. and Moon, Y.S. *"Recent advances in ear biometrics"*, Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, pp.164-169, 2004, Washington, DC, USA. [ISBN:0-7695-2122-3]

**[Quatieri,2001]** Quatieri, T.F. *"Discrete-Time Speech Signal Processing: Principles and Practice"*, 1st Edition, pp.816, 2001, Prentice Hall PTR. [ISBN-10:8177587463 | ISBN-13:978-8177587463]

**[Rabiner,1989]** Rabiner, L. *"A tutorial on hidden Markov models and selected applications in speech recognition"*, Proceedings of the IEEE, vol.77, issue 2, pp.257-286, 1989, IEEE Computer Society. [ISSN:0018-9219]

**[Rabiner,1993]** Rabiner, L. and Juang, B. *"Fundamentals of speech recognition"*, 1st Edition, pp.496, 1993, Prentice Hall. [ISBN-10:0130151572 | ISBN-13:978-0130151575]

**[Rahal,2006]** Rahal, S.M.; Aboalsamah, H.A. and Muteb, K.N. *"Multimodal Biometric Authentication System - MBAS"*, Information and Communication Technologies, 2006. ICTTA '06. 2nd, vol.1, pp.1026-1030, 2006, Damascus, Siria. [ISBN:0-7803-9521-2]

**[Rahman,2008]** Rahman, N.A.; Mohamed, A.S. and Rasmy, M.E. *"Retinal Identification"*, Proceedings of the 2008 Cairo International Biomedical Engineering Conference (CIBEC), pp.1-4, 2008, Cairo, Egypt. [ISBN:978-1-4244-2694-2]

**[Ramos,2008]** Ramos, D.; Gonzalez-Rodriguez, J. et al. *"Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish"*, Proceedings of the 9th Annual Conference of the International Speech Communication Association, pp.1493-1496, 2008, Brisbane, Australia.

**[Raphael,2006]** Raphael, L.J.; Borden, G.J. and Harris, K. S. *"Speech Science Primer: Physiology, Acoustics, and Perception of Speech"*, 5th Edition, pp.336, 2006, Lippincott Williams & Wilkins. [ISBN-10:078177117X | ISBN-13:978-0781771177]

**[Rattani,2007]** Rattani, A.; Kisku, D.R. et al. *"Feature Level Fusion of Face and Fingerprint Biometrics"*, 1st IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS, pp.1-6, 2007, Crystal City, VA,USA. [ISBN:978-1-4244-1597-7]

**[Revett,2007]** Revett, K. *"A Bioinformatics Based Approach to Behavioural Biometrics"*, Frontiers in the Convergence of Bioscience and Information Technologies, FBIT, pp.665-670, 2007, Jeju, South Korea. [ISBN:978-0-7695-2999-8]

**[Reynolds,1992]** Reynolds, D.A. and Rose, R.C. *"An integrated speech-background model for robust speaker identification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.185-188, 1992, San Francisco, CA, USA. [ISSN:1520-6149 | ISBN:0-7803-0532-9]

**[Reynolds,1995-A]** Reynolds, D.A. *"Speaker identification and verification using Gaussian mixture speaker models"*, Speech Communication, vol.17, issue 1-2, pp.91-108, 1995, Elsevier. [ISSN:0167-6393]

**[Reynolds,1995-B]** Reynolds, D.A. and Rose, R.C. *"Robust text-independent speaker identification using Gaussian mixture speaker models"*, IEEE Transactions on Speech and Audio Processing, vol.3, issue 1, pp.72-83, 1995, IEEE Computer Society. [ISSN:1063-6676]

**[Reynolds,1996]** Reynolds, D.A. *"The effects of handset variability on speaker recognition performance: experiments on the Switchboard corpus"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.113-116, 1996, Atlanta, GA, USA. [ISSN:1520-6149 | ISBN:0-7803-3192-3]

**[Reynolds,2000]** Reynolds, D.A.; Quatieri, T.F. and Dunn, R.B. *"Speaker verification using adapted gaussian mixture models"*, Digital Signal Processing, vol.10, issue 1-3, pp.19-41, 2000, Elsevier. [ISSN:1051-2004]

**[Reynolds,2003]** Reynolds, D.; Andrews, W. et al. *"The SuperSID project: exploiting high-level information for high-accuracy speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.4, pp.784-787, 2003, Hong Kong, China. [ISSN:1520-6149 | ISBN:0-7803-7663-3]

**[Reynolds,2005]** Reynolds, D.A.; Campbell, W. et al. *"The 2004 MIT Lincoln Laboratory Speaker Recognition System"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.177-180, 2005, Philadelphia, PA, USA. [ISSN:1520-6149 | ISBN:0-7803-8874-7]

**[Richiardi,2005]** Richiardi, J.; Ketabdar, H. and Drygajlo, A. *"Local and global feature selection for on-line signature verification"*, Proceedings of the 8th International Conference on Document Analysis and Recognition, vol.2, pp.625-629, 2005, Seoul, South Korea. [ISSN:1520-5263 | ISBN:0-7695-2420-6]

**[Rissanen,1978]** Rissanen, J. *"Modeling by shortest data description"*, Automatica, vol.14, issue 5, pp.465-471, 1978, Elsevier. [ISSN:0005-1098]

**[Rodriguez,2008]** Rodriguez, L.P.; Crespo, A.G. et al. *"Study of Different Fusion Techniques for Multimodal Biometric Authentication"*, IEEE International Conference

on Wireless and Mobile Computing Networking and Communications, WIMOB '08, pp.666-671, 2008, Avignon, France. [ISBN:978-0-7695-3393-3]

**[Rosenberg,1998]** Rosenberg, A.E.; Siohan, O. and Parathasarathy, S. *"Speaker verification using minimum verification error training"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.105-108, 1998, Seattle, WA, USA. [ISSN:1520-6149 | ISBN:0-7803-4428-6]

**[Ross,2007]** Ross, A. *"An introduction to multibiometrics"*, The 15th European Signal Processing Conference (EUSIPCO), pp.20-24, 2007, Poznan, Poland. [ISSN 2219-5491]

**[Roy,2007]** Roy, K.; Hudgin, D. et al. *"Iris Recognition: A Java based implementation"*, 10th international conference on Computer and information technology, iccit, pp.1-6, 2007, Dhaka,Bangladesh. [ISBN:978-1-4244-1550-2]

**[Rudin,2001]** Rudin, N. and Inman, K. "An Introduction to Forensic DNA Analysis", 2nd Edition, pp.312, 2001, CRC Press. [ISBN-10:0849302331 | ISBN-13:978-0849302336]

**[Russell,1985]** Russell, M. and Moore, R. *"Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.10, pp.5-8, 1985, Tampa, FL, USA.

**[Russell,1993]** Russell, M. *"A segmental HMM for speech pattern modelling"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.499-502, 1993, Minneapolis, MN, USA. [ISSN:1520-6149 | ISBN:0-7803-7402-9]

**[Rybnik,2008]** Rybnik, M.; Tabedzki, M. and Saeed, K. *"A Keystroke Dynamics Based System for User Identification"*, 7th Computer Information Systems and Industrial Management Applications, CISIM, pp.225-230, 2008, Ostrava, Czech Republic. [ISBN:978-0-7695-3184-7]

**[Saevanee,2009]** Saevanee, H. and Bhattarakosol, P. *"Authenticating User Using Keystroke Dynamics and Finger Pressure"*, 6th IEEE Consumer Communications and Networking Conference, CCNC, pp.1-2, 2009, Las Vegas, NV, USA. [ISBN:978-1-4244-2308-8]

**[Sahidullah,2010]** Sahidullah, M. and Saha, G. *"On the use of perceptual Line Spectral Pairs Frequencies for speaker identification"*, National Conference on Communications, NCC, pp.1-5, 2010, Chennai, India. [ISBN:978-1-4244-6383-1]

**[Samad,2007]** Samad, S.A.; Ramli, D.A. and Hussain, A. *"A multi-sample single-source model using spectrographic features for biometric authentication"*, 6th International Conference on Information, Communications & Signal Processing, pp.1-5, 2007, Singapore. [ISBN:978-1-4244-0983-9]

**[San Segundo,2013]** San Segundo, E. and Gómez Vilda, P. *"Matching Twin and non-Twin Siblings from Phonation Characteristics"*, VII Jornadas de Reconocimiento Biométrico de Personas, pp.10-17, 2013, Zamora, Spain. [ISBN-10:84-616-5690-3 | ISBN-13:978-84-616-5690-5]

**[Sana,2007]** Sana,A.; Gupta,P. and Purkait,R. *"Ear biometrics : a new approach"*, Proceedings of the Sixth International Conference on Advances in Pattern Recognition, ICAPR, vol.1, issue 6, pp.46-50, 2007, Kolkata, India. [ISBN:978-981-270-553-2]

**[Sanchez-Reillo,2000-A]** Sanchez-Reillo, R. *"Hand geometry pattern recognition through Gaussian mixture modelling"*, Proceedings of the 15th International Conference on Pattern Recognition, vol.2, pp.937-940, 2000, Barcelona, Spain. [ISSN:1051-4651 | ISBN:0-7695-0750-6]

**[Sanchez-Reillo,2000-B]** Sanchez-Reillo, R.; Sanchez-Avila, C. and Gonzalez-Marcos, A. *"Biometric identification through hand geometry measurements"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, issue 10, pp.1168-1171, 2000, IEEE Computer Society. [ISSN:0162-8828]

**[Sarkar,2005]** Sarkar, S.; Phillips, P.J. et al. *"The humanID gait challenge problem: data sets, performance, and analysis"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, issue 2, pp.162-177, 2005, IEEE Computer Society. [ISSN:0162-8828]

**[Scheffer,2005]** Scheffer, N. and Bonastre, J.F. *"Speaker Detection Using Acoustic Event Sequences"*, Proceedings of the 9th European Conference on Speech Communication and Technology, pp.3065-3068, 2005, Lisbon, Portugal.

**[Schölkopf,1997]** Schölkopf, B.; Smola, A. and Müller, K.R. *"Kernel principal component analysis"*, Artificial Neural Networks - ICANN'97 (LNCS), vol.1327, pp.583-588, 1997, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-63631-1]

**[Schölkopf,2000]** Schölkopf, B.; Smola, A. et al. *"New Support Vector Algorithms"*, Neural Computation, vol.12, issue 5, pp.1207-1245, 2000, Cambridge, MA, USA. [ISSN:0899-7667]

**[Schölkopf,2001]** Schölkopf, B. and Smola, A.J. *"Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond"*, 1st Edition, pp.648, 2001, The MIT Press. [ISBN-10:0262194759 | ISBN-13:978-0262194754]

**[Schwarz,1978]** Schwarz, G. *"Estimating the Dimension of a Model"*, The Annals of Statistics, vol.6, issue 2, pp.461-464, 1978, The Institute of Mathematical Statistics. [ISSN:0090-5364]

**[Seddik,2004]** Seddik, H.; Rahmouni, A. and Sayadi, M. *"Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier"*, First International Symposium on Control, Communications and Signal Processing, pp.631-634, 2004, Hammamet, Tunisia. [ISBN:0-7803-8379-6]

**[Senoussaoui,2010]** Senoussaoui, M.; Kenny, P. et al. *"An i-vector extractor suitable for speaker recognition with both microphone and telephone speech"*, Odyssey 2010: The Speaker and Language Recognition Workshop, paper 6, 2010, Brno, Czech Republic.

**[Shahin,2008-A]** Shahin, M.; Badawi, A. and Kamel, M. *"Biometric Authentication Using Fast Correlation of Near Infrared Hand Vein Patterns"*, World Academy of Science, Engineering and Technology, International Science Index, vol.13, issue 2 (1), pp.770-775, 2008, World Academy of Science, Engineering and Technology.

**[Shahin,2008-B]** Shahin, M.K.; Badawi, A.M. and Rasmy, M.E. *"A Multimodal Hand Vein, Hand Geometry, and Fingerprint Prototype Design for High Security Biometrics"*, Proceedings of the 2008 Cairo International Biomedical Engineering Conference (CIBEC), pp.1-6, 2008, Cairo, Egypt. [ISBN:978-1-4244-2694-2]

**[Shang,2006]** Shang, L.; Huang, D.S. et al. *"Palmprint Recognition Using ICA Based on Winner-Take-All Network and Radial Basis Probabilistic Neural Network"*, Advances in Neural Networks (LNCS), vol.3972, pp.216-221, 2006, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-34437-7]

**[Sharkas,2008]** Sharkas, M. and Elenien, M.A. *"Eigenfaces vs. fisherfaces vs. ICA for face recognition; a comparative study"*, 9th International Conference on Signal Processing, ICSP, pp.914-919, 2008, Beijing, China. [ISBN:978-1-4244-2178-7]

**[Shi,2007]** Shi, R. and Sun, D. *"A New Security Scheme based on Palmprint Biometrics for Signature"*, 1st IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS, pp.1-6, 2007, Crystal City, VA,USA. [ISBN:978-1-4244-1597-7]

**[Shiv Subramaniam,2007]** Shiv Subramaniam, K.N.; Raj Bharath, S. and Ravinder, S. *"Improved Authentication Mechanism Using Keystroke Analysis"*, International Conference on Information and Communication Technology, ICICT, pp.258-261, 2007, Dhaka,Bangladesh. [ISBN:984-32-3394-8]

**[Shum,2010]** Shum,S.; Dehak,N. et al. *"Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification"*, Odyssey 2010: The Speaker and Language Recognition Workshop, paper 16, 2010, Brno, Czech Republic.

**[Shutler,2004]** Shutler, J.D.; Grant, M.G. et al. *"On a Large Sequence-Based Human Gait Database"*, Applications and Science in Soft Computing - Advances in Soft Computing, vol.24, pp.339-346, 2004, Springer Berlin Heidelberg. [ISSN:1867-5662 | ISBN:978-3-540-40856-7]

**[Siohan,1998]** Siohan, O.; Rosenberg, A.E. and Parthasarathy, S. *"Speaker identification using minimum classification error training"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.109-112, 1998, Seattle, WA, USA. [ISSN:1520-6149 | ISBN:0-7803-4428-6]

**[Smola,2004]** Smola, A.J. and Schölkopf, B. *"A tutorial on support vector regression"*, Statistics and Computing, vol.14, issue 3, pp.199-222, 2004, Kluwer Academic Publishers. [ISSN:0960-3174]

**[Snelick,2005]** Snelick, R.; Uludag, U. et al. *"Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, issue 3, pp.450-455, 2005, IEEE Computer Society. [ISSN:0162-8828]

**[SOCIALab,2014]** SOCIALab *"UBIRIS Database"*, Department of Computer Science, University of Beira Interior, Covilhã Portugal. [on-line acc. 2014 http://iris.di.ubi.pt/]

**[Solewicz,2007]** Solewicz, Y.A. and Koppel, M. *"Using Post-Classifiers to Enhance Fusion of Low- and High-Level Speaker Recognition"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.2063-2071, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Solomonoff,2004]** Solomonoff, A.; Quillen, C. and Campbell, W.M. *"Channel Compensation for SVM Speaker Recognition"*, Odyssey 2004: The speaker and Language Recognition Workshop, pp.57-62, 2004, Toledo, Spain.

**[Solomonoff,2005]** Solomonoff, A.; Campbell, W.M. and Boardman, I. *"Advances In Channel Compensation For SVM Speaker Recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.629-632, 2005, Philadelphia, PA, USA. [ISSN:1520-6149 | ISBN:0-7803-8874-7]

**[Sönmez,1998]** Sönmez, K.; Shriberg, E. et al. *"Modeling dynamic prosodic variation for speaker verification"*, Proceedings of the 5th International Conference on Spoken Language Processing, pp.3189-3192, 1998, Sydney, Australia.

**[Soong,1985]** Soong, F.; Rosenberg, A. et al. *"A vector quantization approach to speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.10, pp.387-390, 1985, Tampa, FL, USA.

**[Soong,1988]** Soong, F.K. and Rosenberg, A.E. *"On the use of instantaneous and transitional spectral information"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.36, issue 6, pp.871-879, 1988, IEEE Computer Society. [ISSN:0096-3518]

**[Stolcke,2005]** Stolcke, A.; Ferrer, L. et al. *"MLLR Transforms as Features in Speaker Recognition"*, Proceedings of the 9th European Conference on Speech Communication and Technology, pp.2425-2428, 2005, Lisbon, Portugal.

**[Stolcke,2007-A]** Stolcke, A.; Kajarekar, S.S. et al. *"Speaker Recognition With Session Variability Normalization Based on MLLR Adaptation Transforms"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.1987-1998, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Stolcke,2007-B]** Stolcke, A.; Shriberg, E. et al. *"Speech Recognition as Feature Extraction for Speaker Recognition"*, IEEE Workshop on Signal Processing Applications for Public Security and Forensics, SAFE, pp.1-5, 2007, Washington, DC, USA. [ISBN:1-4244-1226-9]

**[Stolcke,2008]** Stolcke, A.; Kajarekar, S. and Ferrer, L. *"Nonparametric feature normalization for SVM-based speaker verification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.1577-1580, 2008, Las Vegas, NV, USA. [ISSN:1520-6149 | ISBN:978-1-4244-1483-3]

**[Story,1995]** Story, B.H. and Titze, I.R. *"Voice simulation with a body-cover model of the vocal folds"*, The Journal of the Acoustical Society of America, vol.97, issue 2, pp.1249-1260, 1995, Acoustic Society of America. [ISSN:0001-4966]

**[Story,2002]** Story, B.H. *"An overview of the physiology, physics and modeling of the sound source for vowels"*, Acoustical Science and Technology, vol.23, issue 4, pp.195-206, 2002, The Acoustical Society of Japan. [ISSN:1346-3969]

**[Sturim,2002]** Sturim, D.E.; Reynolds, D.A. et al. *"Speaker verification using text-constrained Gaussian mixture models"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.677-680, 2002, Orlando, FL, USA. [ISSN:1520-6149 | ISBN:0-7803-7402-9]

**[Sturim,2009]** Sturim, D.E.; Campbell, W.M. et al. *"The MIT LINCOLN Laboratory 2008 Speaker Recognition System"*, Proceedings of the 10th Annual Conference of the International Speech Communication Association, pp.2359-2362, 2009, Brighton, UK.

**[Svec,2000]** Svec, J.G.; Horacek, J. et al. *"Resonance properties of the vocal folds: In vivo laryngoscopic investigation of the externally excited laryngeal vibrations"*, The Journal of the Acoustical Society of America, vol.108, issue 4, pp.1397-1407, 2000, Acoustic Society of America. [ISSN:0001-4966]

**[Szewczyk,2007]** Szewczyk, R. *"New Features Extraction Method for People Recognition on the Basis of the Iris Pattern"*, 14th International Conference on Mixed

Design of Integrated Circuits and Systems, MIXDES, pp.645-650, 2007, Ciechocinek, Poland. [ISBN:83-922632-9-4]

**[Tabatabaee,2006]** Tabatabaee, H.; Fard, A.M. and Jafariani, H. *"A novel Human Identifier System using Retina Image and fuzzy Clustering"*, Information and Communication Technologies, 2006. ICTTA '06. 2nd, vol.1, pp.1031-1036, 2006, Damascus, Siria. [ISBN:0-7803-9521-2]

**[Tadj,1998]** Tadj, C.; Dumouchel, P. and Ouellet, P. *"GMM based speaker identification using training-time-dependent number of mixtures"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.761-764, 1998, Seattle, WA, USA. [ISSN:1520-6149 | ISBN:0-7803-4428-6]

**[Tanabian,2005]** Tanabian, M.M.; Tierney, P. and Azami, B.Z. *"Automatic speaker recognition with formant trajectory tracking using CART and neural networks"*, Canadian Conference on Electrical and Computer Engineering, pp.1225-1228, 2005, Saskatoon, Sask, Canada. [ISSN:0840-7789 | ISBN:0-7803-8885-2]

**[Theodoridis,2006]** Theodoridis, S. and Koutroumbas, K. *"Pattern Recognition"*, 4th Edition, pp.984, 2006, Academic Press. [ISBN-10:1597492728 | ISBN-13:978-1597492720]

**[Thornton,2007]** Thornton, J.; Savvides, M. and Kumar, B.V.K.V. *"An Evaluation of Iris Pattern Representations"*, 1st IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS, pp.1-6, 2007, Crystal City, VA,USA. [ISBN:978-1-4244-1597-7]

**[Timms,1992]** Timms, S.R. and King, R.A. *"Speaker verification utilising artificial neural networks and biometric functions derived from time encoded speech (TES) data"*, Communications on the Move, ICCS/ISITA, vol.2, pp.447-450, 1992, Singapore. [ISBN:0-7803-0803-4]

**[Tisby,1991]** Tisby, N.Z. *"On the application of mixture AR hidden Markov models to text independent speaker recognition"*, IEEE Transactions on Signal Processing, vol.39, issue 3, pp.563-570, 1991, IEEE Computer Society. [ISSN:1053-587X]

**[Titze,1973]** Titze, I.R. *"The Human Vocal Cords: A Mathematical Model"*, Phonetica (International Journal of Phonetic Science), vol.28, issue 3-4, pp.129-170, 1973, Karger. [ISSN:0031-8388]

**[Titze,1974]** Titze, I.R. *"The Human Vocal Cords: A Mathematical Model. Part II"*, Phonetica (International Journal of Phonetic Science), vol.29, issue 1-2, pp.1-21, 1974, Karger. [ISSN:0031-8388]

**[Titze,1988]** Titze, I.R. *"The physics of small-amplitude oscillation of the vocal folds"*, The Journal of the Acoustical Society of America, vol.83, issue 4, pp.1536-1552, 1988, Acoustic Society of America. [ISSN:0001-4966]

**[Titze,1994-A]** Titze, I.R. *"Principles of Voice production"*, pp.354, 1994, Prentice Hall. [ISBN-10:013717893X | ISBN-13:978-0137178933]

**[Titze,1994-B]** Titze, I.R. *"Summary Statement"*, Workshop on Acoustic Voice Analysis (WAVA), pp.1-36, 1994, Denver, CO, USA.

**[Tong,2006]** Tong, R.; Ma, B. et al. *"Fusion of accoustic and tokenization features for speaker recognition"*, Chinese Spoken Language Processing (LNCS), vol.4274, pp.566-577, 2006, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-49665-6]

**[Turk,1991-A]** Turk, M. and Pentland, A. *"Eigenfaces for recognition"*, Journal of Cognitive Neuroscience, vol.3, issue 1, pp.71-86, 1991, Cambridge, MA, USA. [ISSN:0898-929X]

**[Turk,1991-B]** Turk, M.A. and Pentland, A.P. *"Face recognition using eigenfaces"*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, pp.586-591, 1991, Maui, HI, USA. [ISSN:1063-6919 | ISBN:0-8186-2148-6]

**[Tydlitat,2007]** Tydlitat, B.; Navratil, J. et al. *"Text-Independent Speaker Verification in Embedded Environments"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.4, pp.293-296, 2007, Honolulu, HI, USA. [ISSN:1520-6149 | ISBN:1-4244-0727-3]

**[Uludag,2004]** Uludag, U. and Jain, A.K. *"Attacks on Biometric Systems: A Case Study in Fingerprints"*, Proc. SPIE-EI 2004, Security, Seganography and Watermarking of Multimedia Contents VI (SSWMC), vol.5306, pp.622-633, 2004 .

**[van Leeuwen,2007]** van Leeuwen, D.A. and Brummer, N. *"An Introduction to Application-Independent Evaluation of Speaker Recognition Systems"*, Speaker Classification I (LNCS), vol.4343, pp.330-353, 2007, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-74186-2]

**[Van Lierde,2005]** Van Lierde, K.M.; Vinck, B. et al. *"Genetics of Vocal Quality Characteristics in Monozygotic Twins: A Multiparameter Approach"*, Journal of Voice, vol.19, issue 4, pp.511-518, 2005, Elsevier. [ISSN:0892-1997]

**[Vapnik,1996]** Vapnik, V.N.; Golowich, S. E. and Smola, A.J. *"Support Vector method for function approximation, regression estimation, and signal processing"*, Advances in Neural Information Processing Systems (NIPS), vol.9, pp.281-287, 1996, Denver, CO, USA.

**[Vapnik,1998]** Vapnik, V.N. *"Statistical Learning Theory"*, 1st Edition, pp.768, 1998, Wiley-Interscience. [ISBN-10:0471030031 | ISBN-13:978-0471030034]

**[Vapnik,2000]** Vapnik, V.N. *"The nature of statistical learning theory"*, 2nd Edition, pp.314, 2000, Springer. [ISBN-10:0387987800 | ISBN-13:978-0387987804]

**[Vapnik,2006]** Vapnik, V.N. and Kotz, S. *"Estimation of Dependence Based on Empirical Data"*, 2nd Edition, pp.505, 2006, Springer. [ISBN-10:0387308652 | ISBN-13:978-0387308654]

**[Victor,2002]** Victor, B.; Bowyer,K. W. and Sarkar,S. *"An evaluation of face and ear biometrics"*, Proceedings of the 16th International Conference on Pattern Recognition, vol.1, pp.429-432, 2002, Quebec, Canada. [ISSN:1051-4651 | ISBN:0-7695-1695-X]

**[Viterbi,1967]** Viterbi, A.J. *"Error bounds for convolutional codes and an asymptotically optimum decoding algorithm"*, IEEE Transactions on Information Theory, vol.13, issue 2, pp.260-269, 1967, IEEE Computer Society. [ISSN:0018-9448]

**[Vogt,2006]** Vogt, R. and Sridharan, S. *"Experiments in session variability modelling for speaker verification"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.897-900, 2006, Toulouse, France. [ISSN:1520-6149 | ISBN:1-4244-0469-X]

**[Vogt,2008]** Vogt, R.; Kajarekar, S. and Sridharan, S. *"Discriminant NAP for SVM Speaker Recognition"*, Odyssey 2008: The Speaker and Language Recognition Workshop, paper 10, 2008, Stellenbosch, South Africa.

**[Wallace,1968]** Wallace, C.S. and Boulton, D.M. *"An Information Measure for Classification"*, The Computer Journal, vol.11, issue 2, pp.185-194, 1968, Oxford University Press. [ISSN:0010-4620]

**[Wallace,2005]** Wallace, C.S. *"Statistical and Inductive Inference by Minimum Message Length"*, 1st Edition, pp.432, 2005, Springer. [ISBN-10:038723795X | ISBN-13:978-0387237954]

**[Wan,2003]** Wan, V. and Renals,S. *"SVMSVM: support vector machine speaker verification methodology"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.221-224, 2003, Hong Kong, China. [ISSN:1520-6149 | ISBN:0-7803-7663-3]

**[Wan,2005]** Wan, V. and Renals, S. *"Speaker verification using sequence discriminant support vector machines"*, IEEE Transactions on Speech and Audio Processing, vol.13, issue 2, pp.203-210, 2005, IEEE Computer Society. [ISSN:1063-6676]

**[Wang,2006]** Wang, G. and Ou, Z. *"Face Recognition Based on Image Enhancement and Gabor Features"*, The Sixth World Congress on Intelligent Control and Automation, WCICA, vol.2, pp.9761-9764, 2006, Dalian, China. [ISBN:1-4244-0332-4]

**[Wang,2008]** Wang, L.; Leedham, G. and Cho,D.S.Y. *"Minutiae feature analysis for infrared hand vein pattern biometrics"*, Pattern Recognition, vol.41, issue 3, pp.920-929, 2008, Elsevier. [ISSN:0031-3203]

**[Weaver,2006]** Weaver, A.C. *"Biometric authentication"*, Computer, vol.39, issue 2, pp.96-97, 2006, IEEE Computer Society. [ISSN:0018-9162]

**[Wellekens,1987]** Wellekens,C. *"Explicit time correlation in hidden Markov models for speech recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.12, pp.384-386, 1987, Dallas, TX, USA.

**[Welling,2002]** Welling, L.; Ney, H. and Kanthak, S. *"Speaker adaptive modeling by vocal tract normalization"*, IEEE Transactions on Speech and Audio Processing, vol.10, issue 6, pp.415-426, 2002, IEEE Computer Society. [ISSN:1063-6676]

**[Whiteside,2001]** Whiteside, S.P. *"Sex-specific fundamental and formant frequency patterns in a cross-sectional study"*, The Journal of the Acoustical Society of America, vol.110, issue 1, pp.464-478, 2001, Acoustic Society of America. [ISSN:0001-4966]

**[Wiener,1949]** Wiener, N. *"Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications"*, 1st Edition, pp.163, 1949, The MIT Press. [ISBN-10:026223002X | ISBN-13:978-0262230025]

**[Wildes,1997]** Wildes, R.P. *"Iris recognition: an emerging biometric technology"*, Proceedings of the IEEE, vol.85, issue 9, pp.1348-1363, 1997, IEEE Computer Society. [ISSN:0018-9219]

**[Wiskott,1997]** Wiskott, L.; Fellous, J.M. et al. *"Face recognition by elastic bunch graph matching"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, issue 7, pp.775-779, 1997, IEEE Computer Society. [ISSN:0162-8828]

**[Woodland,1994]** Woodland, P.C.; Odell, J.J. et al. *"Large vocabulary continuous speech recognition using HTK"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.125-128, 1994, Adelaide, South Australia, Australia. [ISSN:1520-6149 | ISBN:0-7803-1775-0]

**[Xiang,2003]** Xiang, B. and Berger,T. *"Efficient text-independent speaker verification with structural Gaussian mixture models and neural network"*, IEEE Transactions on Speech and Audio Processing, vol.11, issue 5, pp.447-456, 2003, IEEE Computer Society. [ISSN:1063-6676]

**[Xiao,2008]** Xiao, X. and Li, L. *"Face Recognition Based on the Probability Support Vector Machines"*, International Conference on Computer Science and Software Engineering, vol.1, pp.907-910, 2008, Wuhan, Hubei, China. [ISBN:978-0-7695-3336-0]

**[Xu,2005]** Xu, Z.W.; Guo,X.X. et al. *"The blood vessel recognition of ocular fundus"*, Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol.7, pp.4493-4498, 2005, Guangzhou, China. [ISBN:0-7803-9091-1]

**[Yan,2006]** Yan, P. and Bowyer, K.W. *"An Automatic 3D Ear Recognition System"*, Third International Symposium on 3D Data Processing, Visualization, and Transmission, pp.326-333, 2006, Chapel Hill, NC, USA. [ISBN:0-7695-2825-2]

**[Yan,2007]** Yan, P. and Bowyer, K.W. *"Biometric Recognition Using 3D Ear Shape"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.29, issue 8, pp.1297-1308, 2007, IEEE Computer Society. [ISSN:0162-8828]

**[Yang,2005]** Yang, P.; Yang, Y. and Wu, Z. "Exploiting Glottal Information in Speaker Recognition Using Parallel GMMs", Audio- and Video-Based Biometric Person Authentication (LNCS), vol.3546, pp.804-812, 2005, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-27887-0]

**[Yang,2007]** Yang, H.; Dong, Y. et al. *"Cluster adaptive training weights as features in SVM-based speaker verification"*, Proceedings of the 8th Annual Conference of the International Speech Communication Association, pp.2013-2016, 2007, Antwerp, Belgium.

**[Ye,2007]** Ye, B. and Wen,Y.M. *"Gait recognition based on DWT and SVM"*, International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR, vol.3, pp.1382-1387, 2007, Beijing, China. [ISBN:78-1-4244-1065-1]

**[Yeung,2004]** Yeung, D.; Chang, H. et al. *"SVC2004: First International Signature Verification Competition"*, Biometric authentication (LNCS), vol.3072, pp.16-22, 2004, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-22146-3]

**[Yin,2007]** Yin, S.C.; Rose, R. and Kenny, P. *"A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification"*, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 7, pp.1999-2010, 2007, IEEE Computer Society. [ISSN:1558-7916]

**[Yöruk,2006]** Yöruk, E.; Konukoglu, E. et al. *"Shape-based hand recognition"*, IEEE Transactions on Image Processing, vol.15, issue 7, pp.1803-1815, 2006, IEEE Computer Society. [ISSN:1057-7149]

**[Yu,1995]** Yu, K.; Mason, J. and Oglesby, J. *"Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation"*, IEE Proceedings Vision, Image and Signal Processing, vol.142, issue 5, pp.313-318, 1995, IEEE Computer Society. [ISSN:1350-245X]

**[Yuan,2005]** Yuan,L.; Mu,Z. and Xu,Z. *"Using Ear Biometrics for Personal Recognition"*, Advances in Biometric Person Authentication (LNCS), vol.3781, pp.221-228, 2005, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-29431-3]

**[Yun,2000]** Yun, Y.S. and Oh, Y.H. *"A segmental-feature HMM for speech pattern modeling"*, IEEE Signal Processing Letters, vol.7, issue 6, pp.135-137, 2000, IEEE Computer Society. [ISSN:1070-9908]

**[Zamalloa,2006]** Zamalloa, M.; Bordel, G. et al. *"Feature Selection Based on Genetic Algorithms for Speaker Recognition"*, IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, pp.1-8, 2006, San Juan, Puerto Rico. [ISBN:1-424400471-1]

**[Zhang,2003]** Zhang, W.; Yang, Y. and Wu, Z. *"Exploiting PCA classifiers to speaker recognition"*, Proceedings of the International Joint Conference on Neural Networks, vol.1, pp.820-823, 2003, Portland, OR, USA. [ISSN:1098-7576 | ISBN:0-7803-7898-9]

**[Zhang,2004]** Zhang, P.F; Li, D.S. and Wang, Q. *"A novel iris recognition method based on feature fusion"*, Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol.6, pp.3661-3665, 2004, Shanghai, China. [ISBN:0-7803-8403-2]

**[Zhao,1998]** Zhao, W.; Chellappa, R. and Krishnaswamy, A. *"Discriminant analysis of principal components for face recognition"*, Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp.336-341, 1998, Nara, Japan. [ISBN:0-8186-8344-9]

**[Zhao,2003]** Zhao, W.; Chellappa, R. et al. *"Face recognition: A literature survey"*, ACM Computing Surveys (CSUR), vol.35, issue 4, pp.399-458, 2003, New York, NY, USA. [ISSN:0360-0300]

**[Zhao,2009]** Zhao, Z.D.; Zhang, J. et al. *"An effective identification method for speaker recognition based on PCA and double VQ"*, International Conference on Machine Learning and Cybernetics, vol.3, pp.1686-1689, 2009, Baoding, China. [ISBN:978-1-4244-3702-3]

**[Zheng,2004]** Zheng, N. and Ching, P.C. *"Using Haar transformed vocal source information for automatic speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.1, pp.77-80, 2004, Montreal, Canada. [ISSN:1520-6149 | ISBN:0-7803-8484-9]

**[Zheng,2006]** Zheng, N.; Wang, N. et al. *"Speaker Recognition Using Complementary Information from Vocal Source and Vocal Tract"*, Chinese Spoken Language Processing (LNCS), vol.4274, pp.518-528, 2006, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-49665-6]

**[Zheng,2007]** Zheng, N.; Lee, T. and Ching, P.C. *"Integration of Complementary Acoustic Features for Speaker Recognition"*, IEEE Signal Processing Letters, vol.14, issue 3, pp.181-184, 2007, IEEE Computer Society. [ISSN:1070-9908]

**[Zhong,2008]** Zhong, C.; Sun, Z. et al. *"Robust 3D face recognition in uncontrolled environments"*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, pp.1-8, 2008, Anchorage, AK, USA. [ISSN:1063-6919 | ISBN:978-1-4244-2242-5]

**[Zhou,1996]** Zhou, R.W. and Quek, C. *"An automatic fuzzy neural network driven signature verification system"*, IEEE International Conference on Neural Networks, vol.2, pp.1034-1039, 1996, Washington, DC, USA. [ISBN:0-7803-3210-5]

**[Zhou,2006]** Zhou, X.; Peng,Y. and Yang, M. *"Palmprint Recognition Using Wavelet and Support Vector Machines"*, PRICAI 2006: Trends in Artificial Intelligence (LNCS),

vol.4099, pp.385-393, 2006, Springer Berlin Heidelberg. [ISSN:0302-9743 | ISBN:978-3-540-36667-6]

**[Zilea,2003]** Zilea, R.D.; Navratil, J. and Ramaswamy, G.N. *"Depitch and the role of fundamental frequency in speaker recognition"*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol.2, pp.81-84, 2003, Hong Kong, China. [ISSN:1520-6149 | ISBN:0-7803-7663-3]

# II. APPENDICES

## II.1. Software Tools

As previously stated some of the algorithms that have been presented and used in the course of this work have open source implementations, including those concerning feature extraction, which constitutes the core work of the present thesis. However, the available implementations or software tools are not always efficient or easily connectable with other modules in a complete speaker recognition system. Furthermore, the gender-dependent biometric parameterisation module constitutes a contribution of this thesis, which obviously requires an ad hoc implementation. For this reason, we have implemented, in C code, each of the algorithms that have been uses in this work, namely:

- Feature extraction module: The major contribution of this thesis is concentrated in this module, it provides the user with a power tool to generate not only classical features extracted from the voice signal, i.e. MFCC, but the biometric coefficients extracted from the glottal source and vocal tract estimates. Through a configuration file the user can select not only the type of parameters that are going to be generated but also the following settings:
    - Format of input recordings: *Wav* files (format used in HESPERIA database), *sphere* file (format used in NIST databases), *ses* file (format used in ALBAYZIN database).
    - Format of output files: HTK files (in order to make the module compatible with HTK toolkit), *TAB* file (binary tabular format, used by the other modules implemented in this work)
    - Features: The user is allowed to select not only the number of MFCC to be extracted, but also the configuration used to extract these parameters, i.e. filters in the filter bank or bandwidth limits. Additionally, $\Delta$ and $\Delta\Delta$ coefficients can be included in the output feature vector. The extraction of the MFCC parameters extracted from the glottal source and vocal tract estimates follows the same procedure, so the same type of settings applied.
    - Source-Tract separation algorithm: Depending on the quality of the recordings, and the gender of the speaker, it will be useful to modify the order of the filter and the *forgetting factor* parameter.
    - Extra parameters: The user is allowed to select whether to include or not additional parameters in form of energy, $\Delta$E, pitch estimation or third formant estimation, extracted from the voice signal.
    - Feature Normalisation algorithms: Given the characteristic of some of the recordings in the different databases used in the set of experiments carried out, the use of certain features normalisation algorithms has proven its value. Therefore, the user can select whether to apply or not each of the following normalisations:
        - Cepstral mean subtraction [Furui,1981].
        - Feature warping [Pelecanos,2001].
        - RASTA filtering [Hermansky,1994].

Despite being focused on MFCC parameters, it is a powerful and yet flexible module that allows the user to generate a wide variety of parameters while controlling the process at low level.

- Class modelling/class decision: Regarding the modelling, classification and scoring stages, we have implemented the most common techniques that we can find in the state of the art to address the speaker recognition problem. Specifically, we have implemented the GMM/UBM paradigm, a GMM-supervector classifier and the *i*-vector approach. Depending on the approach used to face a specific problem, additional modules are needed. For instance, in the case of using the GMM/UBM paradigm, MAP algorithm has been implemented and LLR scoring step is also provided, while ZNorm and TNorm normalisation scores are available if the user decides to use them. For the case of applying the GMM-supervector approach, the PCA dimensionality reduction technique is available and cosine distance is provided as scoring method (also available on the *i*-vectors approach). No matter whether the *i*-vector or the GMM-supervector approach is used, standard compensation techniques are available for used, such as LDA or WCCN. SNorm has been implemented so it can be used when cosine distance is applied as scoring method.

# II.2. Glossary

## II.2.1. Acronyms

**CMS:** Cepstral Mean Subtraction. Robust feature enhancement which removes channel induced effects as well as any other stationary speech component, by removing the mean cepstral coefficient from each feature over the duration of an utterance (See section 5.1.1).

**DET:** Detection Error Trade-off (see DET curves)

**EM:** Expectation-Maximisation. Iterative algorithm used to produce maximum likelihood estimates of the parameters in statistical models, given some observed data. (See section 3.3.1).

**EER:** Equal Error Rate. Statistic used to compare biometric system performance, it can be define as the location on a DET curve where the FAR and FRR are equal. In general, the lower the EER value, the higher the accuracy of the system.

**FAR:** False Acceptance Rate. In the verification task, it can be defined as the percentage of times a system produces a false match of false acceptance error (i.e. an impostor is accepted as a genuine user) (See section 1.4).

**FRR:** False Rejection Rate. In the verification task, it can be defined as the percentage of times a system produces a false non-match of false rejection error (i.e. a genuine user is not correctly recognised) (See section 1.4).

**FW:** Feature Warping. Robust feature enhancement which consists in transforming the cepstral coefficients so that they follow a specific distribution over a window of speech frames. (See section 5.1.1).

**GMM:** Gaussian Mixture Model is a probability density function represented as a weighted sum of Gaussian component densities. It has become the *de facto* reference method in text-independent speaker recognition, but it is also used in other biometric systems. (See section 3.3).

**JFA:** Joint Factor Analysis. In speaker recognition it aims at modelling speaker and session variability in generative-based models through the use of traditional statistical methods. (See section 3.5.1).

**HMM:** Hidden Markov Model. Stochastic model that can efficiently model temporal sequences of events. (See section 3.2).

**HTK:** Hidden Markov Model Toolkit (http://htk.eng.cam.ac.uk/) is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research although it has been used for numerous other applications including research into speech synthesis, character recognition and DNA sequencing. (See section 3.2.3).

**LDA:** Linear Discriminant Analysis. Widely used technique in the field of pattern recognition both for better discrimination between different classes and for dimensionality reduction. The main goal of LDA is to find a transformation matrix W such that when applied to the original feature space, the between-class variance is maximised while the intra-class variance is minimised. (See Section 3.5.2)

**LLR:** Log-likelihood ratio. Applied to speaker recognition, it represents the logarithmic form of the likelihood ratio which express whether the

observed data is more likely to be produce by the target speaker or by the alternative model.

**MAP:** Maximum A Posteriori. In speaker recognition applications it is used to derive a speaker model from a universal background model. (See section 3.3.2).

**MFCC:** Mel-Frequency Cepstral Coefficients.

**NAP**: Nuisance Attribute Projection. Mathematical method used to find a transformation from the original expanded space into a subspace in which the unwanted variability is removed. (See section 3.3.2).

**NIST:** National Institute of Standards and Technology. A non-regulatory federal agency within the U.S. Department of Commerce that works with industry to develop and apply technology, measurements, and standards. (www.nist.gov).

**PCA:** Principal Component Analysis. Mathematical method aiming to reduce dimensionality of data while preserving the information on the data set, but also used to identify patterns in the data. (See section 3.5.1)

**RASTA:** RelAtive SpecTrA. RASTA filtering consist in the application of a IIR filter in order to remove spectral components that changes at different rate than the one present in speech, i.e. tries to remove convolutional and additive noise (See section 5.1.1).

**SRE:** Speaker Recognition Evaluation. An ongoing series of evaluations of speaker recognition systems.

**SVM:** Support Vector Machine. Discriminative two-class classifier whose purpose is to model the boundaries between two classes as a separating hyperplane. (See section 3.4)

**UBM:** Universal Background Model. In speaker recognition applications it is used to refer to the entire space of possible alternatives to the hypothesised speaker. (See section 3.3.2)

**WCCN:** Within-Class Covariance Normalisation. Compensation technique used to identify and weight orthonormal directions in feature space that maximise task-relevant information while minimizing a particular upper bound on error rate. It is used to remove channel effects on speaker recognition task. (See Section 3.5.2)

## II.2.2. Terms

**Authentication:** Term used in biometrics as a generic synonym of verification (See Verification).

**Behavioural Biometric Characteristic:** Any biometric characteristic that is learned or acquired by an individual over the time. (See Section 1.4.1)

**Biological Biometric Characteristic:** Any biometric characteristic that is primarily based on anatomical or physiological characteristics. (See Section 1.4.1)

**Biometrics:** General term that can be used indistinctly for (www.biometrics.gov):

- A measurable biological or behavioural characteristic of an individual that can be used for recognition purposes.
- Automated process for recognizing an individual based on physiological or behavioural characteristic.

**Biometric System:** Automated system aimed to determine the identity of an individual based on one or more biological or behavioural characteristics. Specifically the system must be capable of (www.biometrics.gov):

- Capturing a biometric sample from an end user.
- Extracting and processing the biometric data from that sample.
- Storing the extracted information in a database.
- Comparing the biometric data with data contained in one or more references.
- Deciding how well they match and indicating whether or not an identification or verification of identity has been achieved.

**Database:** A comprehensive collection of data organised for ease and rapid access, generally in a computer.

**DET Curve:** The Detection Error Trade-off curve represents FRR and FAR on a non-linear scale (see Section 1.4).

**Enrolment:** Process in which biometric samples from an individual are captured in order to get a biometric template that can be stored in a database fort later access.

**Feature:** Distinctive mathematical characteristic derived from a biometric sample.

***i*-vector:** Applied to speaker recognition, set of total factors that simultaneously represent speaker and channel variability (See section 3.5.2)

**Identification:** A one-to-all comparison performed by a biometric system to identify a person by searching within the templates of all enrolled users. We can distinguish between closed-set identification and open-set identification. In the former case, the target individual is known to be present in the database so an exact identification is possible, whereas in the later case there is no warranty for the target speaker to be in the database, so "*unknown individual*" can be a possible output of the system.

**Impostor:** Person who attempt to claim a false identity against a biometric system no matter whether in an intentional way or not.

**Matching:** In biometrics, it can be defined as the process of comparing a biometric sample with a previously acquire template to provide a measure of similarity.

**Model:** In biometrics, it can be defined as mathematical representation of an individual's distinguishing features extracted from a biometric sample.

**Recognition:** Generic term used to describe a biometric system that refers to its main function, regardless whether it is set to operate in verification, closed-set identification or open-set identification mode.

**Threshold:** User defined operating point for verification or open-set identification tasks in biometric systems. In other words, the acceptance or rejection of biometric data is dependent on the match score falling above or below the threshold. Thus there is always a trade-off between false match rates and false non-match rates in every security system. (www.biometrics.gov)

**Score:** Value provided by a biometric system reflecting the similarity between an input biometric sample and a stored template.

**Speaker Recognition:** Biometric process aimed at determining the identity of an individual through the voice.

**Speech Recognition:** Automatic process to identify spoken words.

**Supervector:** Term used in speaker recognition to refer to a vector that provides a robust representation of utterances through the use of just a single vector in a high-dimensional space, allowing sequences of different durations to be compared and classified directly using traditional machine learning approaches, such as SVMs. (See Section 3.5.1)

**Verification:** A one-to-one comparison performed by a biometric system to determine whether the captured data provided by an individual corresponds to the stored biometric template of the person's claimed identity**.**