Glottal Source Cepstrum Coefficients Applied To NIST SRE 2010

L. M. Mazaira, A. Álvarez, P. Gómez, R. Martínez, C. Muñoz

Grupo De Informática Aplicada al Procesado de Señal e Imagen (GIAPSI)
Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n,
28660 Boadilla del Monte, Madrid – SPAIN
E-MAIL: luismiguel.mazaira@upm.es

Abstract. Through the present paper, a novel feature set for speaker recognition based on glottal estimate information is presented. An iterative algorithm is used to derive the vocal tract and glottal source estimations from speech signal. In order to test the importance of glottal source information in speaker characterization, the novel feature set has been tested in the 2010 NIST Speaker Recognition Evaluation (NIST SRE10). The proposed system uses glottal estimate parameter templates and classical cepstral information to build a model for each speaker involved in the recognition process. ALIZE [1] open-source software has been used to create the GMM models for both background and target speakers. Compared to using mel-frequency cepstrum coefficients (MFCC), the misclassification rate for the NIST SRE 2010 reduced from 29.43% to 27.15% when glottal source features are used.

Keywords: Glotal Source, Speaker Characterization, Speaker Recognition, GMM, Speech production, NIST SRE 2010.

1 Introduction

Speaker Recognition has been an active research area over the last decades. In this time, different classification methods have been proposed (GMM-UBM, SVM, etc.) and different characterization parameters have been tested (short-term spectral features, high-level features, etc.) [2]. Since 1996 National Institute of Standards and Technology) NIST has provided a framework (SRE – Speaker Recognition Evaluation [3]) to test the advances in this area. Sign of the importance of this International Evaluation is that in the NIST SRE 2010 [4] the sites participating in the evaluations reach the number of 49, among them universities and companies, achieving recognition equal error rates (EER) closed to 1% under some specific conditions.

Regardless of the classification method applied, participants in NIST SRE have focused their efforts on incorporating high level features [5] [6] in order to make SR systems more robust. However, most systems still use classical parameterization techniques that take into account the power spectral density of speech as a whole. The present work defends the idea that a parameterization technique which considers the voice signal as a composition of acoustic articulation and phonation gesture will provide better results than classical approaches. For this reason, the system presented to the NIST SRE 2010 uses a parameterization technique taking into account spectral characteristics of vocal tract (acoustic-phonetic) and glottal estimate (phonation-gesture) of voicing speech.

The speech production model, shown in Fig. 1, assumes that the speech signal, *s(n)*, is obtained by filtering a glottal excitation signal *e(n)*, with the transfer function of the vocal tract *h(n)* and a radiation model [7]. The idea that the glottal signal can be represented by a transfer function given by 1/*f*, thus not providing essential information to the characterization process of a particular speaker, has been taken traditionally for granted. However, recent research has shown that the glottal source presents some distinguishable properties, both in time and frequency domain, which can be applied to speaker characterization; especially in gender and age detection, pathology detection and speaker identification, among others [8]-[11]. Moreover, in preliminary work, [11] [12], it has been shown that the glottal source bears essential biometric information, which can be applied in speaker recognition tasks.

The aim of the present work is to show the discriminative power of glottal source estimate in speaker identification tasks. Section 2 will explain the glottal source estimation process. A description of the system submitted to the NIST SRE 2010 which includes the novel parameterization features is presented in section 3. Section 4 describes the experimental results achieved in the evaluation. Finally, conclusions and future work will be exposed in section 6.

2 Estimation of the glottal source

In the physiological speech production model (**Fig. 1**), the voiced speech is generated by a glottal excitation signal, *e(n)*, which is spectrally conditioned by the vocal tract with transfer function given by $F_{vt}(z)$ to produce the speech signal before radiation $s_l(n)$ and after radiation *s(n)*.

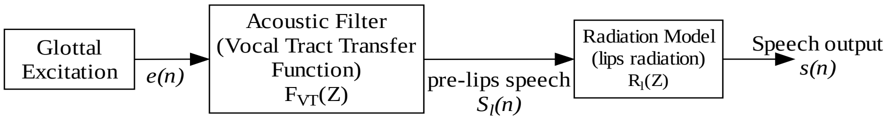
Glottal Excitation
e(n)
Acoustic Filter (Vocal Tract Transfer Function) F_VT(Z)
pre-lips speech S_l(n)
Radiation Model (lips radiation) R_l(Z)
Speech output s(n)

Fig. 1. Block diagram of voiced speech production

Different methods have been proposed to estimate the glottal source from the speech signal [13] [14]. However, in order to grant the orthogonality, in terms of correlation, between the vocal tract and the glottal source estimates an iterative algorithm, described in [15], has been selected. Fig. 2 depicts the block diagram of the applied method.

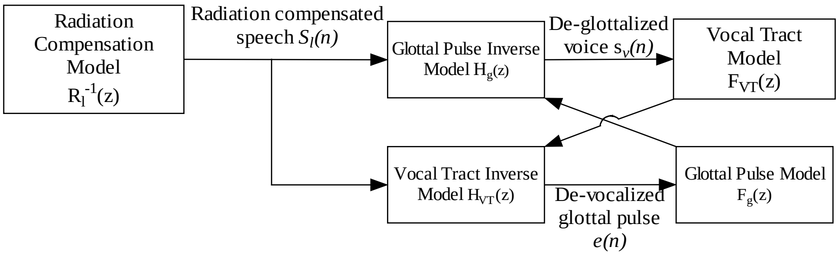
Radiation Compensation Model
$R_l^{-1}(z)$
Radiation compensated speech $S_l(n)$
Glottal Pulse Inverse Model $H_g(z)$
De-glottalized voice $s_v(n)$
Vocal Tract Model $F_{VT}(z)$
Vocal Tract Inverse Model $H_{VT}(z)$
De-vocalized glottal pulse $e(n)$
Glottal Pulse Model $F_g(z)$

Fig. 2. Framework for glottal source separation from voice by adaptive joint estimation.

The iterative algorithm comprises the following steps:

1. Remove the radiation effects from voice s(n), by filtering with $R_l^{-1}(z)$.
2. Remove the glottal pulse generating model $H_g(z)$ from the radiation compensated voice $S_l(n)$. In the first iteration $H_g(z)$ need not be a very precise estimation, as it will be refined by successive iterations.
3. The vocal tract model $F_{VT}(z)$ is estimated from the de-glottalized voice $s_v(n)$, using an adaptive lattice (typically of order 20-30)
4. Remove the vocal tract model from input voice, by filtering with the inverse function $H_{VT}(z)$ to obtain a better estimation of the glottal source e(n).
5. Produce a more precise model of the glottal source $F_g(z)$, which could be used to refine $H_g(z)$.
6. Repeat steps 2-5 to a desired end (typically two iterations are enough).

In order to make the reconstructed glottal source useful to the speaker recognition system, MFCC derived from the magnitude spectrum of the glottal source estimation are extracted and combined with MFCC from speech. (Fig. 3).

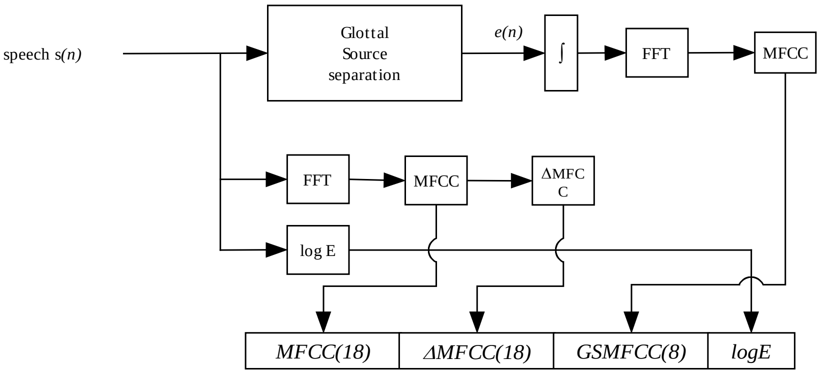

Fig. 3. Parameterization scheme used in the GIAPSI System.

3 GIAPSI System Description for NIST SRE 2010

To test the performance of the new set of parameters when compared with classical parameterization features, a complete system was design to participate in the NIST SRE 2010. Fig. 4 provides a block diagram representation of the GIAPSI system presented in the NIST SRE2010.

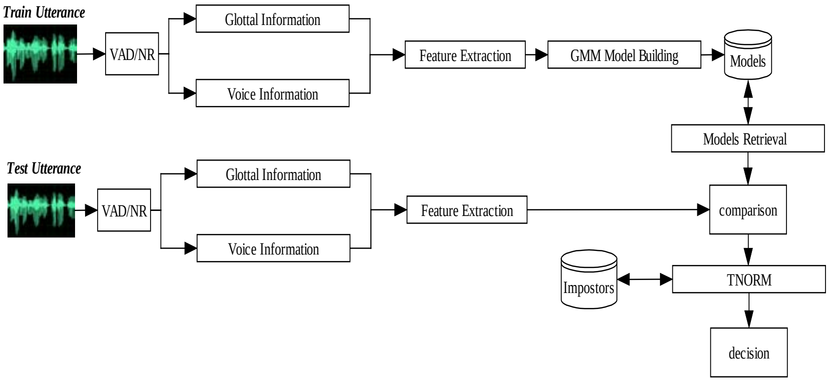

Fig. 4. Block Diagram of NIST SRE GIAPSI System.

System description can be divided into 3 different subsystems:

- Preprocessing and feature extraction.-
  The evaluation data provided by NIST consists of training and test stereo files stored as 8-bit µ-law SPHERE format recorded at 8 KHz sampling frequency. For each file, the SPHERE header contains supplementary information, such as whether or not the data was recorded over a telephone line, whether or not the data is from an interview session, etc. Although evaluation rules allow the use of this auxiliary information for recognition purposes, these have not been used by the system. Word transcripts, produced using automatic speech recognition (ASR) system and provided by NIST have not been also used.

  A preprocessing step, over the recordings contained in each file, is performed including a Voice Activity Detection stage (VAD) and a simple cross-channel speaker cancelation. An adaptive VAD algorithm based on energy detection has been implemented and computed over a 64ms-long Blackman window with 13ms overlap. In the case of speaker cancelation, the algorithm applied consists on removing segments on the channel of interest that coincide with segments classified as including voice in the complementary channel. As evaluation data involves interviews or telephone conversations, detecting voice in both channels at the same time could mean that the specific segment contains information from non-target speaker, or that the target speaker is breathing, laughing or just uttering fillers like “uh”, “um” ; making such segment unsuitable for automatic training or testing.

Additionally, as some of the files include telephone conversations, it was 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 [16].

The last step consists in applying the glottal source extraction process and feature extraction algorithm described in section 2 for each of the segments of interest. For each 32 ms voiced frame (with 8ms overlapping), a 45 feature vector has been extracted (Fig. 3), which contains the following information:

  - 18 MFCC + 18 ΔMFCC (from voice as a whole)
  - 8 MFCC derived from the power spectral density of the glottal source signal
  - Logarithm of the frame energy

- Speaker modeling.-
  Each speaker is represented by a Gaussian Mixture Model (GMM), λ_s, that has been adapted from a gender-dependent Universal Background Model (UBM) using the MAP algorithm [17] to adapt only the distribution means. The UBM is also represented as a GMM, λ_{UBM}, which has been formed from a set of speakers from the evaluation data of the NIST SRE 2006 using the EM-algorithm. Although different number of mixtures has been tested during system development, finally the number of mixtures in the GMM was set to 1024.

  The selection of this modeling technique is twofold. First of all, Gaussian Mixture Model (GMM) is a probabilistic model which has become the *de facto* reference method in text-independent speaker recognition. Second, the availability of an open-source software for model generation, known as ALIZE toolkit [1], thus reducing the development time of the system.

- Scoring.-
  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 represented by a model $\lambda_{S.}$

$$LLR=\log P(X|\lambda_S)-\log P(X|\ \lambda_{UBM})\ . \quad (1)$$

  In order to improve the performance of the scoring process, only the top 16 Gaussians have been used to produce the LLR score for each test file. This is implemented by first calculating log likelihood values for the UBM and finding the top 16 distributions. Then only these top distributions are evaluated for the target model.

  T-NORM normalization has been applied based on a cohort of impostors extracted from the NIST SRE2006 evaluation data. For this reason, the speaker model set also contained 30 female and 30 male impostor models for use in T-NORM score normalization.

4 Experimental Results

The NIST SRE 2010 involves 9 different tasks [4]. However our system only submitted results for two of this task: core-core and 10sec-10sec conditions. In the core-core task condition, the recording data comes both from telephone conversations and microphone recorded interviews. The conversational telephone recordings are all approximately five minutes in duration, while the interview excerpts are of varying duration between three and fifteen minutes. However, the total amount of time in which the speaker of interest is speaking is substantially smaller and not constant for different speaker. The number of speaker models to be created during train phase is 3026 female models and 2434 male models, while the number of trials for the core-core condition task is 610748.

In the case of the 10sec-10sec task, both the training and test files provided are two-channel excerpts from a telephone conversation that are estimated to contain approximately 10 seconds of actual speech in the channel of interest. The number of speaker models to be created during train phase is 1298 female models and 959 male models, while the number of trials for the 10sec-10sec condition task is 55391.

In order to evaluate the influence of the glottal estimate parameters (GEP), an additional system (baseline) was developed in which the cepstral parameters from the glottal estimate have been removed from the feature set. The results obtained by the GIAPSI system in the NIST SRE 2010 are depicted in Fig.5 – Fig. 8 and Table 1.

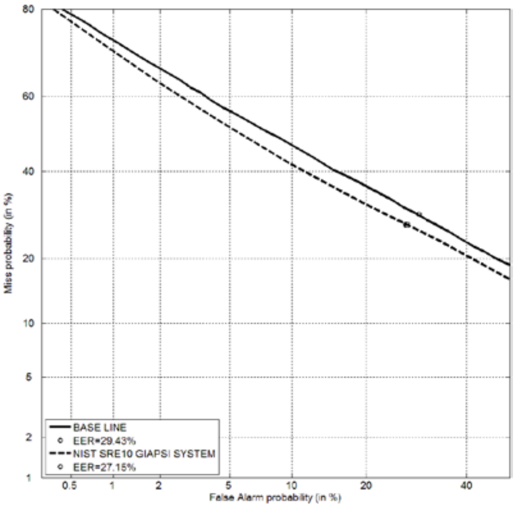

Fig. 5. Core-Core condition DET Curves – baseline vs. GIAPSI System.

The inclusion of the glottal information in the feature set clearly provides an improvement on the recognition rates, with a reduction on Equal Error Rates (ERR) terms of 2.28%.

Although not included in NIST SRE 2010, as no cross-gender trials are defined, gender dependant evaluation has been also performed. Fig. 6 depicts DET curves for core-core task condition, but evaluating independently female and male trials. In this case, the improvement in the recognition rates is clearly larger in male trials, with an EER reduction of approximately 3%, when GEP are used.

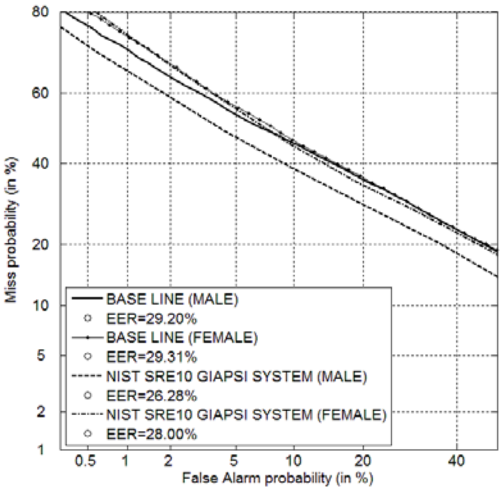

Fig. 6. Gender dependant DET curves.

Fig. 7 presents the recognition results achieved under two different conditions of the core-core task. *Condition 1* (cond-1) refers to the specific case in which all trials involve interview speech from the same microphone in training and testing. Results show again that the inclusion of GEP in feature vectors improves recognition rates achieving 12.5% EER under controlled conditions, making these parameters suitable for security access applications. *Condition 5* refers to trials involving normal vocal effort conversational telephone speech both in training and testing. Under this new situation, the baseline system outperformed the presented system.

Table 1. EER evaluation for baseline and GIAPSI Systems under different conditions.

Condition	***EER% (baseline)***	***EER% (GIAPSI System)***	***DIFF***
Condition 1	14.39	12.52	**-1.87**
Condition 2	26.97	24.36	**-2.61**
Condition 3	31.61	34.90	**3.29**
Condition 4	25.21	21.87	**-3.34**
Condition 5	23.69	26.78	**3.09**
Condition 6	29.04	31.52	**2.48**
Condition 7	26.94	27.47	**0.53**
Condition 8	17.37	22.01	**4.64**
Condition 9	24.84	23.00	**-1.84**

Table 1 provides the results achieved, in terms of EER, for different conditions described in the NIST evaluation plan [4]. A negative difference denotes that the set of parameters proposed increase the recognition rates.

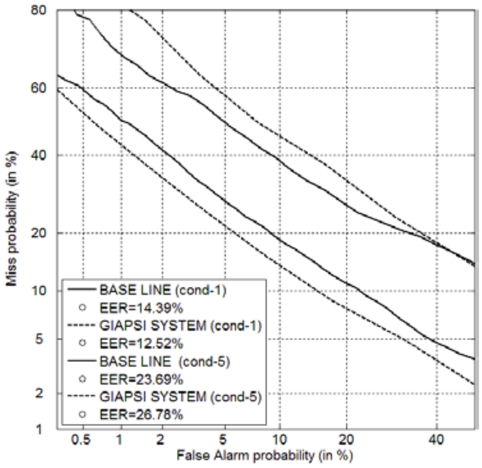

Fig. 7. DET curves for different conditions of the core-core task.

Finally, Fig. 8 depicts DET curves for the specific condition in which normal vocal effort telephone conversations have been used both for training and testing. Dotted line refers to short time enrollment and test utterances (10sec-10sec condition), while solid line refers to long time recordings (core-core condition). Although results are far from being optimal, the DET curves on Figure 9 show that the system proposed is not especially influenced by short enrollment/testing information.

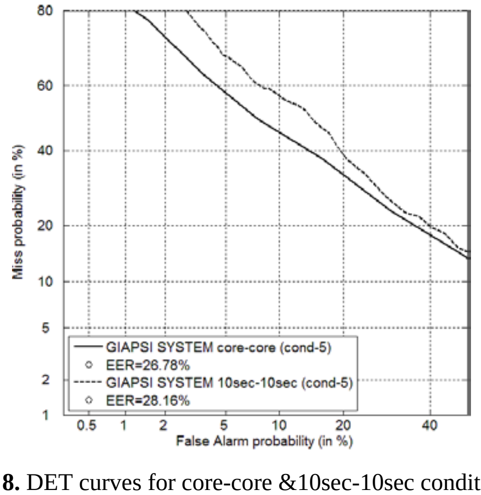

Fig. 8. DET curves for core-core &10sec-10sec conditions.

5 Discussion and Conclusion

In order to test the importance of the glottal information to characterize a given speaker a complete system have been implemented to participate in an international evaluation contest, NIST SRE 2010. Although the results obtained in the evaluation are still far away from the best results achieved by other participants, it has been shown that the incorporation of GEP improves the speaker recognition rate, thus confirming the conclusions in previous works [8][11][18]. On the other hand results on Table 1 show that further work is needed when dealing with telephone recordings.

An additional problem in this case (telephone recordings) is the fact that under some subset of trials, recordings for train and test differ in vocal effort of the speaker. Up to now, we have considered the modality of the phonation as being associated with the speaker's emotional state, thus leaving this situation for further study. However in NIST SRE 2010 the modal phonation has been forced artificially adding another source of variation to the problem of accurately extraction of glottal information.

Several reasons explain the results achieved by the system in the NIST SRE2010. First of all, no channel normalization techniques in order to remove linear channel convolutional effects have been applied. Additionally no side information from recordings has been used, neither during training nor during testing. Finally, more preprocessing work is needed in the reconstruction of the GEP, as the systems must deal with some of these factors: empty recordings both in test and train phase, extremely noisy recordings, gender misclassification, etc.

Acknowledgements

This work is being funded by grants TEC2009-14123-C04-03 from Plan Nacional de I+D+I, Ministry of Science and Innovation, and by project HESPERIA (http.//www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

References

1. Bonastre, J.-F. , Wils, F., and Meignier, S.:"ALIZE, a free toolkit for speaker recognition," in *Proc. of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing.* ICASSP. Philadelphia (USA), vol. 1, pp. 737-740, (2005)
2. Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. Speech Communication, Vol.52, Issue 1, pp.12-40. (2010)
3. National Institute of Standards and Technology (NIST), Information Access Division (IAD), http://www.itl.nist.gov/iad/mig/tests/sre/
4. NIST SRE 2010 evaluation plan, http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

5. Kajarekar, S. S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L., and Bocklet: THE SRI NIST 2008 speaker recognition evaluation system. In: *Proc. of the 2009 IEEE international Conference on Acoustics, Speech and Signal Processing*. ICASSP. Washington, DC (USA), 4205-4208
6. Mason, M., Vogt, R., Baker, B., Sridharan, S.: The QUT NIST 2004 speaker verification system: a fused acoustic and high-level approach. In: *Australian International Conference on Speech Science and Technology*. pp. 398-403. (2004)
7. Nickel, R. M.: Automatic Speech Character Identification. *IEEE Circuits and Systems Magazine*, Vol. 6, No. 4, pp. 8-29. (2006)
8. Plumpe, M.D., Quatieri, T.F., and Reynolds, D.A.: Modeling of the glottal flow derivative waveform with application to speaker identification. IN *IEEE Trans. Speech Audio Process.* , vol. 7(5), pp. 569-85, (1999)
9. Sulter, A. R., and Wit, H. P.: Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age. IN *J. of the Acoustical Society of America*, Vol. 100, pp. 3360-3373, (1996)
10. Gudnason, J. & Brookes, M.: Voice source cepstrum coefficients for speaker identification. IN *Proc. of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. ICASSP. Las Vegas (USA), pp 4821 – 4824, (2008)
11. Gómez, P., Díaz, F., Martínez, R., Godino, J. I., Álvarez, A., Rodríguez, F., Rodellar, V.: A hybrid parameterization technique for speaker identification. IN *16th European Signal Processing Conference* (EUSIPCO 2008), Lausanne, Switzerland, (2008)
12. Gómez, P., Rodellar, V., Álvarez, A., Lázaro, C. A. Murphy, K., Díaz, F., Fernández, R.: Biometrical Speaker Description from Vocal Cord Parameterization. IN *Proc. of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing*. ICASSP, Toulouse, France, pp. 1036-1039. (2006)
13. Akande, O. O. & Murphy, P. J.: Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Communication*, Vol. 46, No. 1, pp. 1-13. (2005)
14. Alku, P.: Parameterisation Methods of the Glottal Flow Estimated by Inverse Filtering. IN *Proc. ISCA Workshop on Voice Quality: Functions, analysis and synthesis (VOQUAL03)*. Geneva, pp. 81-87. (2003)
15. Gómez, P., Martínez, R., Díaz, F., Lázaro, C., Álvarez, A., Rodellar, V., Nieto, V.: Estimation of vocal cord biomechanical parameters by non-linear inverse filtering of voice. IN *Proc. 3rd Int. Conf. on Non-Linear Speech Processing NOLISP'05*. Barcelona, Spain, pp. 174–183, (2005)
16. Ephraim, Y. & Malah, D.: Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Trans. Acoustics, Speech Signal Proc.*, vol.32, pp. 1109-1121, (1984)
17. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, (2000)
18. Gómez, P., Álvarez, A., Mazaira, L. M., Fernández, R., Nieto, V., Martínez, R., Muñoz, C., Rodellar, V.: A hybrid Parameterization Technique for Speaker Identification. *In Proc. 1er WTAC-ASAF Workshop en Tecnologías de Audio Cognitivo para Aplicaciones en Seguridad y Acústica Forense*, Univ. Las Palmas de Gran Canaria, pp. 19-22. (2007)