Emotion Recognition from Speech with Acoustic, Non-Linear and Wavelet-based Features Extracted in Different Acoustic Conditions

Juan Camilo Vásquez Correa

Department of Electronic and Telecommunication Engineering
Research group on Embedded Systems and Computational Intelligence
SISTEMIC

Master in Telecommunications Engineering

University of Antioquia UdeA February 2016

Emotion Recognition from Speech with Acoustic, Non-Linear and Wavelet-based Features Extracted in Different Acoustic Conditions

Juan Camilo Vásquez Correa

Document in partial fulfillment of the requirements for the degree of:

Master of Science

Director:

Ph.D. Jesús Francisco Vargas Bonilla

University of Antioquia

Department of Electronic and Telecommunication Engineering

Research group on Embedded Systems and Computational Intelligence

SISTEMIC

Medellín, Colombia

February 2016

Programming is the sweet spot, the high leverage point in a digital society. If we don't learn to program, we risk being programmed ourselves.
Douglas Rushkoff.

Computers are incredibly fast, accurate and stupid. Human beings are incredibly slow, inaccurate and brilliant. Together they are powerful beyond imagination.
Yan Ayrton.
composer & scientist.

Acknowledgements

There are many people to I would like to thank, who made possible the development of this work.

First of all I would like to thank to Prof. Jesús Francisco Vargas Bonilla, director of this study, for his invaluable support, and dedication throughout the development of this work. His guidance allowed me to grow up in a personal and professional way during the development of this study. Also to Prof. Juan Rafael Orozco Arroyave for allowing me discover the world of digital signal processing and pattern recognition, and for the support in the development of this study. I would like to thank also to the Prof. Dr. Ing. Elmar Nöth for the opportunity to share and learn as a member of team in the Pattern Recognition Lab. of the Friedrich Alexander Universität Erlangen-Nürnberg.

To the team of digital signal processing and pattern recognition from the University of Antioquia: Tatiana Villa, Elkyn Belalcazar, and Tomas Arias, for the academic and personal support. Also, to the undergraduate students who provide support for the development of this study.

To my parents, my sister and the rest of my family for their invaluable support during all these years, to all my friends for their unconditional personal support, and for allowing me to share with them both in the good and bad times.

To the founders of this study: The International Speech and Communication Association (ISCA) for award me with one of the ISCA student grants to attend the INTERSPEECH 2015 in Dresden. The Colombian Administrative Department of Science, Technology and Innovation (COLCIENCIAS) for support this study through the young researchers and innovators program, in 2014. And the COLCIENCIAS project # 111556933858, titled "Analysis of the discriminant capacity of phonation, articulation and prosody features from patients with Parkinson's disease on preclinic and advanced stages for the development of computer aided tools for supporting the diagnosis and monitoring of the patients".

Abstract

In the last years, there has a great progress in automatic speech recognition. The challenge now it is not only recognize the semantic content in the speech but also the called "paralinguistic" aspects of the speech, including the emotions, and the personality of the speaker.

This research work aims in the development of a methodology for the automatic emotion recognition from speech signals in non-controlled noise conditions. For that purpose, different sets of acoustic, non-linear, and wavelet based features are used to characterize emotions in different databases created for such purpose. The acoustic analysis considers a standard feature set developed for emotion recognition from speech called OpenEAR, and a set of spectral and noise derived measures. The non-linear analysis is based on non-linear dynamic measures and include the correlation dimension, the largest Lyapunov exponent, the Hurst exponent, and the Lempel Ziv complexity. Also it is proposed a set of measures derived from parametric non-stationary analysis using time dependent ARMA models. The wavelet based measures consider features derived from the wavelet packet transform, and different wavelet time-frequency representations such as the bionic wavelet transform, and the synchro-squeezed wavelet transform.

Different non-controlled noise conditions are tested considering four different scenarios: (1) the original recordings, (2) the signals degraded by two additive noisy environments: street and a cafeteria babble, (3) the re-captured signals in two natural noisy environments as street and office, and (4) the recordings compressed by seven different codecs used for the transmission through mobile, VoIP, and web based telephone channels. Also two different speech enhancement algorithms are tested to evaluate if they are suitable to improve the results in the classification of emotions in noisy speech signals. A classification scheme based on the combination of Gaussian mixture models and Support vector machines is used for the analysis.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Speech is the most natural method of communication between humans, which has motivated researches to consider speech as one of the main methods for interaction between humans and machines. In the last years, there has a great progress in automatic speech recognition, which refers to the transcription of the human speech in a sequence of words. The challenge now is not only recognize the semantic content from speech but also the called “paralinguistic" aspects of the speech in order to reach a natural interaction between humans and machines. The paralinguistic aspects are related to “how" is transmitted the message, and include the emotions, the personality, and others cognitive aspects of the speakers.

This study addresses the development of methodologies based on digital signal processing and machine learning to the development of computational tools that can automatically recognize the emotions of a person according to the information provided only by speech. In the following sections are presented the motivation of this study, a survey of the state of the art, followed by the actual issues related to emotion recognition from speech, and the main contributions of this study.

1.1 Motivation

The interest in recognition of emotions and affect has been increased in the field of speech and language processing over the last decade. Emotion recognition can improve the quality of services and even the quality of life. While automatic speech recognition is a part of most intelligent systems such as virtual assistants or mobile phones, such systems do not have the human ability of observe and react according to the affective response of humans. The automatic emotion recognition is essential to render speech-based systems more human-like in order to reach out a more natural interaction. For instance, in the seventh framework programme for information and communication technology of the European commission,

efforts are devoted to increasing accessibility and efficiency of spoken dialogue systems by integrating emotional and other paralinguistic cues [2].

There are a great deal of potential application that may use technologies related to automatic emotion recognition in real scenarios. Such applications are described in the following subsection.

1.1.1 Applications

Most of applications are motivated from a user-centric perspective, where the main objective is to increase the quality of service or even the quality of life [2]. Technologies related to emotion recognition from speech can be used to support for customer and emergency call-centers, to improve tutoring systems for education, in public surveillance systems, to detect stress of drivers, for supporting psychological treatment, and in the entertainment sector.

Call-centers

The service provided by call-centers can be managed and optimized by evaluating the emotional state of the callers and/or the agents. The main aim is to detect changes such as an increase in the number of angry callers or an increase in the average stress of the agents [3]. Angry callers could also be detected in order to specially trained agents handle the situation and calm the customer by special dialog strategies [2]. As example, the company Nemesysco developed a technology for voice analysis for application in emergency and customer call centers to detect and measure anger, stress and other emotions that may arise in call center conversations [4, 5].

Tutoring systems, and virtual agents

In this case might be useful the knowledge of certainly emotions of the user including stress or deception in order to adapt the teaching pace. As example, The European SEMAINE project, a sensitive artificial listener (SAL) able to sustain a conversation with a human using social interaction skills [6].

Surveillance

Several situations related to security such as crisis management, piloting, stress level detection, or vandalism surveillance, which may be aided by the analysis of “fear-type" emotions

such as fear, anger, disgust, and desperation. In this case can be analyzed the aggressiveness of potential aggressors, or fear in the potential victims [7].

Psychological treatment

Automatic emotion recognition might be useful in the assessment of patients under clinical depression, which is marked by emotional disturbances consist of prolonged periods of excessive sadness, reduced emotional expression and physical drive [8]. Recent studies show that features related to prosody and glottal waves are useful to detect and manage major depression disorders [9, 10].

Entertainment

Several applications could be developed in the entertainment industry based on technologies for emotion recognition from speech. For example in on-line role playing games in order to improve the credibility of characters and the immersion of the user [2].

1.2 State of the Art

1.2.1 Databases

One of the main problems in the automatic analysis of emotions from speech is to collect the data necessary to train statistical classifiers. A catalog of datasets for emotion recognition was created by the project HUMAINE [11]. The web site lists three different categories: multi-modal datasets (audio-visual, audio and physiological data), speech only datasets, and data formed only with facial expressions. The datasets for emotion recognition from speech can be classified into three groups according to the type of emotion elicitation: acted, evoked, and natural, each one with different characteristics in the recording process and the kinds of emotion.

Acted datasets

Many of corpus available for emotion recognition consist of emotional speech produced by professional or lay actors, based on certain emotion labels. In this case the emotional content is more intense, as was found by Williams and Stevens in [12]. They concluded that acted emotions tend to be more exaggerated than the real ones. The acted recordings usually are captured with high audio quality avoiding problems in signal processing with reverberant or noisy speech. Another advantage of the acted datasets is that a balanced distribution of all

emotions can be guaranteed improving the performance of the classification methods. The main disadvantage of these kind of datasets is related that the emotions under evaluation are limited to only a few basic emotions. The actors are also influenced by stereotypes of vocal expression in the emotions, producing that the natural expression of emotions may be missed in some cases [13]. The most common acted datasets are described as follows.

Berlin emotional dataset [14]: It contains 534 voice recordings produced by 10 German native speakers who acted 7 different emotions including anger, disgust, fear, happiness, sadness, boredom, and neutral. The recordings were sampled at 16 kHz.

IEMOCAP [15]: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multi-modal, and multispeaker database. It contains approximately 12 hours of audiovisual data, including video, speech recordings, motion capture of the face, and text transcriptions. The audio files consist of 10039 recordings sampled at 16 kHz uttered by 10 English native speakers who acted 10 different emotions. The recording process considered several dyadic sessions where actors performed improvisations or scripted scenarios, specifically selected to elicit emotional expressions. The recordings were labeled by multiple annotators into categorical labels such as anger, happiness, sadness, and neutrality, as well as dimensional labels such as valence and arousal.

GVEESS [16]: The Geneva Vocal Emotion Stimulus Set (GVEESS) contains emotional speech samples of 14 different emotions uttered by 12 actors. The 14 emotions are anxiety, disgust, happiness, hot anger, interest, cold anger, boredom, panic fear, shame, pride, sadness, elation, contempt and desperation. A selection of 224 emotional speech samples (16 sentences $\times$ 14 emotions) recorded at 44.1 kHz was obtained.

SAVEE [17]: The Surrey Audio-Visual Expressed Emotion (SAVEE) database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion. The data were recorded in a visual media lab with high quality audio-visual equipment with a sampling frequency of 44.1 kHz.

GEMEP [18]: The Geneva Multimodal Emotional Portrayal (GEMEP) corpus is a multimodal database recorded by 10 actors (five male, five female), which record 1260 speech samples with a sampling frequency of 44.1 kHz. The database contains recordings of 18 different emotions including among others sadness, joy, fear, and hot anger.

SAFE [19]: The Situation Analysis in a Fictional and Emotional (SAFE) corpus consist of 400 audio-visual sequences from 8 seconds to 5 minutes extracted from 30 fiction movies. The database contains recordings of both normal and abnormal situations. The corpus contains four types of emotions: (1) fear-type emotions such as stress and anxiety, (2) other negative emotions such as anger, sadness, and disgust, (3) neutral state, and (4) positive emotions.

Evoked datasets

The induction or evocation techniques try to effectively change the emotional state of the subject. There are several procedures to the induction of the emotions. The more commons are: (1) the free mental generation, which may include hypnosis. (2) The guided mental generation which include the imagination or remembering of a situation which involves the emotion or the listening of pieces of music. (3) The presentation of emotional material such as movies or stories where the stimulus is presented to the subject. (4) The generation of physiological states by the administration of drugs like anti-depresants or adrenaline to produce arousal in the subjects.

There are induction techniques designed to produce emotions from speech. For example a difficult spelling task to elicit negative emotions [20], a mental arithmetic task to evoke stress [21], and the Wizard of Oz scenario, where a malfunctioning system is simulated to evoke anger [22]. There are some evoked databases for emotion recognition from speech, which are described as follows.

enterface05 [23]: It contains 1317 audio-visual recordings with 6 emotions produced by 44 speakers, including anger, disgust, fear, happiness, sadness, and surprise. In this database each subject was instructed to listen six successive short stories. After each story the subject had to react to the situation by reading predefined sentences closely related to each story. The recordings were sampled at 44.1 kHz

Semaine [24]: This database was created by the European Semaine project in order to build a Sensitive Artificial Listener. The database includes audiovisual recordings of natural human computer interactions. Instead of assigning an emotional label for each sentence, the subjective evaluations correspond to continuous assessment of the emotional content in real time (50 values per second). The dataset contains 69 recordings in English language sampled at 48 kHz.

Natural datasets

There are several works that have analyzed emotions from speech using natural emotion databases. In [3] the authors use real call-center data to recognize happiness, anger, and neutral emotions with the purpose to help call-center supervisors to monitor the calls, and to identify agents who are not able to satisfy the customer. Another studies that consider real world call-center speech recordings are the presented in [25, 26]. Such databases are private, and no comparisons can be made. However, there are some natural databases available to the study of emotions from speech.

FAU-Aibo [27]: This corpus contains 18216 recordings sampled at 16 kHz uttered by 51 kids from two German schools (26 and 25 subjects, respectively) interacting with an Aibo pet robot. The database considers 5 different emotions (anger, emphatic, neutral, positive, and rest). For the construction of this database, the children were led to believe that Aibo was responding to their commands by producing a series of fixed and predetermined behaviors. As Aibo sometimes disobeyed the commands, it induced different emotional reactions in the children.

SUSAS [28]: The Speech under Simulated and Actual Stress (SUSAS) database is partitioned into four domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male), with ages ranging from 22 to 76 were employed to generate in excess of 16000 utterances recorded at 8 kHz. The database also contains several longer speech files from four Apache helicopter pilots. A common highly confusable vocabulary set of 35 aircraft communication words make up the database.

RECOLA [29]: The RECOLA database includes 7 hours of multimodal recordings of spontaneous collaborative and affective interactions in French. The database contains audio, video, Electrocardiography and Electro-dermal activity signals, which were continuously and synchronously recorded from 34 participants. The signals are labeled according to the degree of arousal, and valence.

Summary

Table 1.1 summarizes the description of each one of the databases, the Table contains the name of database, the number of recordings, the sampling frequency, the emotions included, and the type of dataset (acted, evoked, natural). The experiments in this study are performed over five of the databases described: Berlin, SAVEE, enterface05, FAU-Aibo, and IEMOCAP.

Table 1.1 Summary of databases for emotion recognition from speech

Dataset	# Recordings	# Speakers	Fs (k Hz)	Type	Emotions
Berlin [14]	534	12	16	Acted	Hot anger, Boredorm Disgust, Fear, Neutral Happiness, Sadness
IEMOCAP [15]	10309	10	16	Acted	Hot anger, Happiness Disgust, Fear, Neutral Sadness, Surprise Excitation, Frustration Others
GVEESS [16]	224	12	44.1	Acted	Anxiety, Disgust Happiness, Hot anger Interest, Cold anger Boredom, Panic, Pride Sadness, Shame, Elation Contempt, Desperation
SAVEE [17]	480	4	44.1	Acted	Hot anger, Happiness Disgust, Fear, Neutral Sadness, Surprise
GEMEP [18]	1260	10	44.1	Acted	Admiration, Amusement Anxiety, Cold anger Contempt, Despair, Shame Disgust, Elation, Pride Hot anger, Interest, Panic Sadness, Pleasure, Relief Surprise, Tenderness
SAFE [19]	400	30 movies	-	Acted	Fear, Other negatives Neutral, Positives
enterface05 [23]	1317	44	44.1	Evoked	Hot anger, Happiness Disgust, Fear Sadness, Surprise
Semaine [24]	69 conversation	-	48	Evoked	Hot anger, Happiness Disgust, Fear, Sadness Amusement, Contempt
FAU-Aibo [27]	18216	51	16	Natural	Hot anger, Emphatic Neutral, Joy Rest
SUSAS [28]	16000	7	8	Natural	Low stress, Middle stress High stress, Neutral
RECOLA [29]	7 hours	35	-	Natural	Arousal degree (1-5) Valence degree (1-5)

1.2.2 Feature Extraction

One of the main aims in recognition of emotions from speech is to find suitable features to represent the emotional content of the speaker. Currently, the feature extraction process has been focused on large sets of acoustic features including measures derived from prosody (pitch, energy and the speaking rate), spectral and cepstral such as the Mel frequency cepstral

coefficients (MFCC), and voice quality such as noise measures, jitter, and shimmer [30]. There are studies that also consider the use of non-linear measures, and features derived from the wavelet transform to recognize emotions from speech.

In this subsection are described several works for recognition of emotions from speech using acoustic, non-linear, and wavelet-based features.

Acoustic analysis

The most common measures for the analysis of emotions from speech include features derived from the acoustic analysis. Features related to the the fundamental frequency, the energy content, spectral features, the MFCC, perturbation measures such as jitter and shimmer, and voice quality measures such as the harmonic to noise ratio (HNR) are commonly used for emotion recognition from speech.

One of the main features used is the fundamental frequency (F_0). In [31], 39 different statistics derived from the F_0 are analyzed to discriminate between emotional and neutral speech. The authors use the Berlin database and implement a classifier based on the linear discriminant analysis. The authors report an accuracy of 80.9%. The authors conclude that gross F_0 contour statistics such as mean, maximum, minimum, and range are more emotionally prominent than features describing only the shape of the F_0. In [32], 36 different acoustic measures are used. The authors focuses on the features related to the contour of the F_0 combined with MFCC, the duration of unvoiced frames, and the speech rate to classify the seven emotions in the Berlin database. A support vector machine (SVM) with a Gaussian kernel is used for recognition following a 10 folds cross-validation strategy. The authors do not indicate if those folds are speaker dependent or independent. The authors report accuracies of up to 84.9%, and conclude that boredom presents a maximum F_0 lower than the other emotions. In [33], different features related to the contour of the F_0, the first three formant frequencies, the gains of the vocal tract, and glottal parameters were proposed. 54 features for each voiced segment are calculated to recognize the five emotions in the FAU-Aibo database. The authors consider a hidden Markov Model (HMM) for classification and use the validation strategy followed in the “INTERSPEECH 2009 emotional challenge" [34]. The best result reported corresponds to an unweighted average recall (UAR) of 40.3% considering only nine selected features formed only the based on F_0, and the formants. In [35], 16 features related to the F_0, noise measures, jitter, and shimmer are used to classify happiness, sadness, anger, and neutral state in the Berlin database. The authors perform a gender-dependent modeling considering different GMMs for male and female speakers. The reported accuracies are around 96% and 95% for male and female speakers, respectively.

The authors conclude that the separation of gender before the classification of emotions is useful to obtain better results.

Other common measures are derived from spectral, and prosody analysis. In [36], a standard toolbox called OpenEAR was presented. Such toolkit compute 5967 measures formed by 117 descriptors $\times$ 51 statistical functionals. The descriptors include spectral, cepstral, and prosodic features. The authors recognize the seven emotions in Berlin, and the six emotions in enterface05 using a SVM for classification with a 10-fold stratified cross-validation. The authors report accuracies of up to 89.5% in Berlin, and 75.2% in enterface05 databases, but the cross-validation strategy used does not guarantee speaker independence in the results. The presented toolkit has been used in many related works. In [37], six different large scale acoustic feature sets are applied, the sets are based on the OpenEAR toolkit. One of the sets corresponds to the 384 features used as baseline in the “INTERSPEECH 2009 emotion challenge" to evaluate different emotions in the speech recordings of several databases including Berlin and enterface05. The authors use a multi-class SVM, and follow a leave one speaker out (LOSO) cross-validation strategy to optimize the parameters of the classifier. The reported accuracies are around 96% in Berlin, and 76% in enterface05 classifying the two levels of the arousal dimension, i.e., high and low; 80% in Berlin and 65% in enterface05 for the classification of the valence dimension, i.e., positive and negative emotions; and 80% in Berlin, and 68% in enterface05 for the multi-class, which means classifying a total of seven classes in Berlin and six classes in enterface05. In [38], the author use the same subset of 384 features to model anger, sadness, happiness, and neutrality from the IEMOCAP database, and the five emotions included in the FAU-Aibo database. The authors propose a new classification scheme based on a hierarchical binary decision tree which decisions on each node are taken using a SVM. The reported UAR for the two databases, considering a speaker independent validation strategy is 58.4% in IEMOCAP and 39.9% in FAU-Aibo. In [39], the complete feature set computed using OpenEAR, formed by 6552 features was used in Berlin, and enterface05 databases. The authors propose a classifier based on the generalized discriminant analysis based on a deep neural network, and follow a leave one group speaker out strategy to guarantee speaker independence in the results. The reported accuracies for Berlin database correspond to 97.4% for arousal dimension, 87.5% for valence dimension, and 81.9% for the classification of the seven emotions. For the enterface05 database, the results are 80.8% for arousal, 79.7% for valence, and 61.1% for the classification of the six emotions. In [40], the subset of 1582 features used as baseline in the “2010 INTERSPEECH computational paralinguistic challenge" was used by the authors to classify the six emotions in enterface05 database, and four classes in the FAU-Aibo database, including emphatic, neutral, motherese, and negatives emotions. The authors propose a

method based on least square regression (LSR) for recognition, and report an UAR of 69.3% in enterface05 database, and 60.5% in FAU-Aibo database. In [41], the same subset of 1582 features is used to train a deep denoising autoencoder (DAE). One different DAE is created per gender. The authors model four emotions from the IEMOCAP database and use a SVM with a radial basis kernel function for classification. An UAR of 63.1% in the recognition of the four emotions is reported. The same authors, in [42], use the subset of 384 features described previously and used in [37, 38] to train a denoising autoencoder with the aim of perform a cross corpus experiment, where the SUSAS, and the Airplane behavior corpus are used as training sets, and the FAU-Aibo database is used for test. The authors discriminate between negative and idle, according to the valence dimension. A SVM with linear kernel is used as classifier, and the reported UAR is of up to 64.2%. In [43], the authors calculate 513 features related to spectral and prosodic measures using openEAR. The authors use factor analysis to quantify the dependence of the acoustic features with traits such as the speakers, the lexical content, and the emotions, finding that 76% of the variability of the features is associated with the lexical content instead of the emotional content. Based on that fact, the authors propose a lexical and speaker normalization to compensate the effects introduced by these external factors. The IEMOCAP database is used to recognize happiness, anger, sadness, and neutral, and a SVM is used for classification. The reported accuracies are 55.32% without applying the proposed normalization, and 56.75%, after the lexical normalization.

There are works focused on another acoustic features such as modulation spectral, perceptual evaluation, and the combination of audio, and video information. In [44], features derived from modulation spectral were proposed to capture both acoustic frequency, and the temporal modulation frequency components. 82 features related to modulation spectral, including the energy, and the centroids of sub-bands were calculated. The proposed features are combined with 75 prosodic features to classify the seven emotions in Berlin database. The authors use a SVM and follow a LOSO cross-validation strategy, and report an accuracy of 80.9%. In [45], 106 speech features based on F_0, the energy content, duration, and MFCC are merged with features derived from the facial expression. The seven emotions of the SAVEE database are used for classification. The set of features are reduced using principal component analysis (PCA) and LDA. A Gaussian classifier is used. The authors report accuracies of up to 68.5% using only speech-based features, and 97.9% using the combination of speech, and visual features. In [46], the information of speech signals is combined with the obtained with the facial expression to recognize the six emotions in the enterface05 database. The speech based features include the formant frequencies and MFCC. The authors use a neural network for classification, and perform feature selection using ANOVA. The reported accuracy only

for speech based features is 55%, while using the combination of speech, and video features is 70.3%. In [47], 145 speech based features are combined with 504 video based features. The acoustic features include both prosodic and spectral features such as F_0, the energy content, and the energy of Mel frequency filter bank outputs. The authors classify happiness, sadness, anger, and neutral in the IEMOCAP database, and propose a classification based on a deep neural network. The authors report accuracies of up to 66.1% considering the combination of audio, and video features. In [48], a new set of features based on perceptual quality measures using the perceptual evaluation of audio quality (PEAQ) standard was proposed. The feature set was formed by 9 measures related to spectral envelope, perceptual bandwidth, and the harmonic content. The authors use the Berlin database to perform three different classification tasks: the classification of the seven emotions, the classification according to the arousal dimension, and according to the valence dimension. The authors use a SVM for classification, and report accuracies of up to 85.9% for the recognition of the seven emotions, 95.1% to discriminate between high vs low arousal, and 95.6% for the classification of positive vs negative valence emotions. In [49], 14 MFCC, and their first and second derivatives are computed to characterize emotions in speech. The authors consider all of the seven emotions included in the Berlin database and the six emotions included in the enterface05 database. The classification is performed using a hybrid system based on Deep Neural Networks and Hidden Markov Models (DNN-HMM). The authors randomly select 60% of the data for train, and the remaining 40% for test. Both databases contain several voice recordings per speaker, thus as the train and test sets are formed randomly, the speaker independence is not guaranteed. The reported accuracies are 77.9%, and 53.9% for Berlin and enterface05, respectively. In [50], 12 MFCC and their first, and second derivatives are used to classify the five emotions in the FAU-Aibo database. The authors propose a new classification strategy based on a measure of distance between the emotional classes. Such measure is calculated according to the euclidean distance between the log-likelihood obtained from a GMM generated for the test data, and the GMM that represents each class. The reported UAR is 44.2%. Recently, in [51], a new set of features based on a computational model of the human auditory system is proposed. The model simulates the process from the *pinna* through the auditory nerve. The features are extracted from the output of the auditory model. The feature set consist of the mean and standard deviation of 283 modulation filtered signals obtained as outputs of the auditory nerve model. The authors consider the seven emotions in the Berlin database, and six emotions in the SAVEE database, and propose a classifier based on a hierarchical binary decision tree using a SVM. The best accuracies reported are 72.3%, and 73.8% both for Berlin, and SAVEE databases, respectively. In both

cases, LOSO cross-validation strategy is performed to guarantee speaker independence in the results.

Non-linear analysis

The speech production model involves some non-linear process such as the non-linear pressure flow in glottis, and the non-linearity that occurs in the vocal fold collision [52]. These processes can not be characterized using classical measures. In order to resolve this problem, the non-linear dynamics (NLD) analysis has been established as a mathematical alternative for the analysis of this kind of process. The NLD analysis describes the temporal evolution of a system through a multiple dimension space on which the speech signal is reconstructed. The use of NLD or complexity measures from speech processing tasks has been increased in the last years and have proved to be useful in the analysis of voice quality, and the voice pathology detection [52–54], also have been used from speech emotion recognition [55–57].

In [58] a new feature called smoothed non-linear energy operator (SNEO)-based amplitude modulation cepstral coefficient (AMCC) is proposed to recognize emotions from speech. The feature set is formed by 256 features to recognize the five classes in the FAU-Aibo database. The authors consider a GMM to model the emotions, and report an UAR of 44.5%, following a cross-validation strategy of nine folds speaker independence. In [55], features related to non-linear dynamics are proposed to model the non-linear effects in emotional speech. A 24-dimensional feature vector is formed by 6 descriptors $\times$ 4 statistical functions. The descriptors include the first minimum of the mutual information, the Shannon entropy, the correlation dimension, the correlation entropy, the Lempel-Ziv complexity, and the Hurst exponent. The authors classify anger, fear, and neutral state in Berlin database, using a neural network. The reported accuracy is of up to 75.4%. The authors conclude that recordings associated with fear, and anger show more complexity than the associated with neutral state. The reason could be refereed to that in such emotions, people tend to use more fricative sounds, than in neutral state, these fricative sounds are more noisy and complex than the voiced sounds.

Wavelet analysis

There exist several works related to the use of the wavelet packet transform (WPT) for emotion recognition in speech. The wavelet analysis allows a multi-resolution analysis of time, frequency, and energy.

In [59], the log-energy and features derived from the Teager Energy Operator (TEO) are calculated on 14 bands of the sixth level of the WPT considering the Daubachies10 wavelet function to characterize seven different emotions in a local indian database that contains recordings of speech in five different languages. The authors consider a GMM for classification and report accuracies of up to 94% considering the log-energy based features, and 99% using the TEO-based features. The results are compared to the obtained using MFCC, which are around to 88%. In [60] the authors consider the energy content on decomposed bands from WPT to discriminate between anger, neutral, and happiness. The authors decompose the signal into four bands using two levels of the WPT and use a Haar mother function. The study considers recordings of a database in Marathi language. The classification is performed with a threshold that is set according to the amplitude of the coefficients. The authors report accuracies of 85%, 65%, and 80% for the recognition of anger, happiness, and neutral speech, respectively. In [61], a set of measures based on wavelet decompositions calculated on specific frequency bands is proposed. The wavelet perceptual packets are considered to model the wavelet transform according to the Bark scale. A total of 17 critical bands are obtained from the 3rd, 4th, and 5th levels of the WPT. The auto correlation envelope area associated with each wavelet coefficient is calculated. Six of the seven emotions of the Berlin database are classified. The classification system consists of two stages, the first one is a HMM and the second one is a neural network that is trained with the posterior probabilities obtained from the HMM. The authors report accuracies of 68.8% for the first stage and 91.8% for the second stage. In [62], a new type of features based on the energy entropy calculated on selected bands from WPT obtained from speech and glottal signals is proposed. The authors recognize the seven emotions in the Berlin database, using a GMM for classification and report accuracies of up to 54%. Recently, in [56] a new set of features which combines the NLD analysis, and the wavelet decomposition is proposed to recognize different emotions and to detect stress in speech. The feature set is formed with the Hurst exponent (HE) obtained from the detail coefficients of the discrete wavelet transform (DWT) in the first five levels. The authors model the seven classes of the Berlin database and use a GMM-based classifier. The reported accuracy is 68.1%. In [63], a new set of features based on WPT is proposed. The authors calculate minimum, maximum, mean, median, and standard deviation from all coefficients of the fifth level of the WPT, using the Daubachies2 wavelet function. The seven emotions of Berlin database are classified, using a multi-class SVM. An accuracy of 60% is obtained using only the first 8 wavelet coefficients instead of the 32 of the fifth level. The authors conclude that the low frequency coefficients are the most important to recognize emotions from speech signals. In [64], a new feature extraction approach based on WPT is proposed, adapting the concept of the conventional MFCC to the

wavelet domain. The authors called Coiflet wavelet packet cepstral coefficients (WPCC). The authors consider only six emotions of the Berlin database for the experiments, excluding disgust, and use a SVM as classifier. The authors also compare different wavelet mother functions, and obtain the best results considering the Coiflet3 wavelet, reporting an accuracy of up to 81.1%.

Summary

Tables 1.2, 1.3, 1.4, 1.5, and 1.6 contain the more relevant works, and the methodologies evaluated in the described datasets for emotion recognition from speech: the Berlin, enterface05, FAU-Aibo, IEMOCAP, and SAVEE. The tables include the reference of the work, the number and the description of the features used, the main result obtained, and the classification task evaluated.

Table 1.2 Results reported in Berlin database

Source	# Feat.	Description	Result	task
[31]	39	F_0	80.9%	neutral vs emotional
[32]	36	F_0, MFCC, formants	84.9%	7 classes
[35]	16	F_0, jitter, shimmer HNR	95%	anger, sadness happiness, neutral
[36]	5967	OpenEAR,	89.5%	7 classes
[37]	384	OpenEAR subset 2009 INTERSPEECH challenge	80.0% 96.0% 80.0%	7 classes arousal dimension valence dimension
[39]	6552	OpenEAR,	81.9% 97.4% 87.5%	7 classes arousal dimension valence dimension
[44]	157	Modulation spectral and Prosody	80.9%	7 classes
[48]	9	Perceptual evaluation of audio quality	85.9% 95.1% 94.3%	7 classes arousal dimension valence dimension
[49]	42	MFCC derivatives	77.9%	7 classes
[51]	566	Auditory model	72.3%	7 classes
[55]	24	non-linear dynamics	75.4%	neutral, fear anger
[61]	17	WPP	91.8%	6 classes
[62]	120	WPT energy-entropy	54%	7-classes
[56]	12	Hurts exponent in DWT	68.1%	7-classes
[63]	120	statistics from WPT, and Δ	60.0%	7 classes
[64]	36	WPCC	81.1%	6 classes

Table 1.3 Results reported in enterface05

Source	# Feat.	Description	Result	task
[36]	5967	OpenEAR,	75.2%	6 classes
[37]	384	OpenEAR subset 2009 INTERSPEECH challenge	68.0% 76.0% 65.0%	6 classes arousal dimension valence dimension
[39]	6552	OpenEAR	61.1% 80.8% 79.7%	6 classes arousal dimension valence dimension
[40]	1582	OpenEAR subset 2010 INTERSPEECH challenge	69.3%	6 classes
[46]	80	MFCC, formants	55.0%	6 classes
[49]	42	MFCC derivatives	53.9%	6 classes

Table 1.4 Results reported in FAU-Aibo database

Source	# Feat.	Description	Result	task
[33]	9	F_0, Formants	40.3%	5 classes
[38]	384	OpenEAR subset 2009 INTERSPEECH challenge	39.9%	5 classes
[40]	1582	OpenEAR subset 2010 INTERSPEECH challenge	60.5%	4 classes: Emphatic, Neutral, Motherese, Negatives
[42]	384	OpenEAR subset 2009 INTERSPEECH challenge	64.2%	valence dimension
[50]	36	MFCC derivatives	44.2%	5 classes
[58]	256	SNEO based AMCC	44.5%	5 classes

For the Berlin dataset, the reported results for the classification of the seven emotions range from 60% to 89%, using different kinds of features and classification strategies, note that the best results are those which contains more features such as [36]. For the case of the detection of high vs low arousal, the results reported are around to 96%, and for the discrimination between positive and negative valence emotions, the results range from 80% to 94%.

For the case of enterface05, the highest result for the classification of the six emotions is of up to 75%, the results range from 53% to 75%. As in Berlin, the highest results consider a large set of acoustic features related to F_0, the energy content, duration, MFCC, and voice

Table 1.5 Results reported in IEMOCAP database

Source	# Feat.	Description	Result	task
[38]	384	OpenEAR subset INTERSPEECH 2009 challenge	56.3%	Angry, Happy Sad, Neutral
[43]	513	OpenEAR	56.7%	Angry, Happy Sad, Neutral
[41]	1582	OpenEAR subset, INTERSPEECH 2010 challenge	63.1%	Angry, Happy Sad, Neutral
[47]	685	audio and video	66.1%	Angry, Happy Sad, Neutral

Table 1.6 Results reported in SAVEE

Source	# Feat.	Description	Result	task
[45]	106	Acoustics	68.5%	7 classes
[51]	566	Auditory model	73.8%	7 classes

quality measures. For the 2-class recognition experiments, the results obtained for arousal dimension range from 76% to 80%, and for valence dimension range from 75% to 79%.

For the case of FAU-Aibo, the results reported range from 39.4% to 44.5% for the classification of the five emotions. For the discrimination between idle and negatives emotions the results are around to 64.0%. The work proposed in [40] report the highest result for the multi-class problem, but only consider four of the five emotions.

Besides the IEMOCAP database contains ten different emotions, the main works which use this database only consider four of them: anger, happiness, sadness, and neutral state because most of the recordings are labeled with these emotions. For this case, the highest results are reported in [47], but it consider both speech, and video-based features. The highest result that consider only speech-based features is the proposed in [41], which report an accuracy of up to 63.1%.

For the SAVEE database, there are few works that consider such database, the highest result reported is 73.8%, for the classification of the seven emotions.

1.2.3 Evaluation in non-controlled noise conditions

There are relatively few studies that consider the influence of noise and telephone channels in the recognition of emotions from speech, but currently this aspect has gained the attention of the research community. There are two different approaches for the evaluation of the non-controlled noise conditions. The first one consider addition of different kinds of noise to the conventional databases. The second one consider the evaluation of the speech signals recorded in non-controlled noise conditions scenarios such as customer service, and emergency call centers.

According to the first approach, in [65] the authors consider the Berlin database, and white noise addition in a range from -10 to 20 dB to evaluate the effect of emotion recognition under noise conditions. The authors consider a large set of features formed by 4000 measures based on statistical functions calculated from the contours of intonation, intensity, formants, HNR, and MFCC. The classification process is performed using a SVM with a polynomial kernel, and a cross-validation strategy based on 10 folds stratified. The reported accuracy is 86.7% for the noise-free recordings, while for the noisy signals, the accuracy ranges from 67.2% to 83.4%. Other kind of noise is considered in [66], where the authors evaluate the effect of the acoustic in-car noise conditions for recognition of emotions from speech. Recordings of Berlin database are used for the addition of the noise produced by different vehicles considering both convertibles, and non-convertibles. The authors consider a set of features formed by 1400 measures related to energy, duration, and perturbation measures. A classifier based on a SVM with linear kernel, using the LOSO cross validation strategy is proposed. The reported accuracy for the noise-free speech signals is 74.9%, while for the noisy signals the results range from 65.6% to 74.5%, depending on the kind of vehicle, the speed, and the contact surface. In [67], a database recorded in a car environment is considered to discriminate between positives, negatives, and neutral state. Also the Berlin database is considered and degraded with different kinds of noise such as Gaussian, parking lot, highway, and city street with different SNR levels from 5 to 15 dB. A speech enhancement algorithm based on an adaptive threshold from the wavelet transform is considered. The authors calculate a 1054-dimensional feature vector using several measures including MFCC, F_0, and intensity. The classification is performed using a SVM w following a 10-folds cross-validation strategy. The obtained accuracy considering the original acoustic conditions is 84% for Berlin database, and 88.1% for the in-car environment database. For the case of the noisy signals, the accuracy ranges from 16.8% to 37% depending on the kind of the added noise. For the evaluation of speech enhancement, the accuracy ranges from 37.5% to 63%. In [68] the authors use the Berlin database to detect anger from speech. A model of telephone channel based on the adaptive multi-rate (AMR) codec, and additive noise is

used to simulate non-controlled noise conditions. The authors use features derived from MFCC, RASTA filters, and linear prediction coefficients (LPC), and use a GMM-based classifier. The authors report an accuracy of 90.1% for the noise-free conditions, while in the non-controlled noise conditions, the accuracy ranges from 77.4% to 88.2% depending on the kind of noise. Recently, in [69] the authors use the Berlin database to classify the arousal dimension, and the valence dimension. The authors consider signals filtered by telephone channels, and contaminated by different kinds of noise such as car, factory, and babble with SNR=0 dB, and use features related to MFCC, and the log-energy filtered by an auto-regressive (AR) model. The authors perform the classification using GMMs. For the classification of arousal dimension, the authors report accuracies of 93.3% in the noise-free recordings, while in non-controlled noise conditions the accuracy ranges from 79.8% to 91.8%. For the classification of valence dimension, the authors obtain accuracies of 80.0% in the noise-free conditions, while in non-controlled noise conditions the accuracy ranges from 60.0% to 74.8%.

There are also some studies that consider real world scenarios such as call centers. In [70] the authors consider a corpus recorded in a medical emergency call center to discriminate between positive and negative emotions. The feature extraction is performed with measures derived from the F_0, the energy content, duration, spectral analysis, disfluency and non-linguistic event features. The authors report accuracies of up to 83.5%, using a classifier based on a logistic model tree. In [3] the authors use real call-center data to recognize happiness, anger, and neutral emotions in a real call center database using acoustic features related to the energy content and F_0. The authors report accuracies of 57.1% for happiness, 60.0% for anger, and 50% for neutral emotion, using a GMM-based classifier. In [26] the authors discriminate between positive, negative and neutral emotional speech in a real-world corpus collected from a complaint call center. A 374-dimensional feature vector derived from acoustic measures such as MFCC, the F_0, formants, and the energy content is considered. The classification is performed using a SVM with a radial basis kernel. The authors report a F1 Score of 0.54, which is a measure of accuracy that takes into account the precision and the recall.

Table 1.7 summarizes the results of related works for automatic emotion recognition in non-controlled noise conditions. It consider both the studies that use the conventional databases with additive noise and phone channels, and the studies focused on the analysis in real-world scenarios.

Table 1.7 Results of the evaluation of non-controlled noise conditions

Source	Features	Conditions	Result	Task
[65]	acoustics	Clean Berlin DB	86.7%	7-classes
		white noise SNR from -10 to 20 dB	67.2% to 83.4%	
[66]	acoustics	Clean Berlin DB	74.9%	7-classes
		Car noise at different speed, and surfaces	65.6% to 74.5%	
[67]	acoustics	Clean Berlin DB	84.01%	7-classes
		Different street noise	16.8% to 37%	
		Speech enhancement	37.5% to 63%	
[68]	MFCC, energy	Clean Berlin DB	90.1%	anger vs others
		environment noise+telephone channel	77.4% to 88.2%	
[69]	MFCC, energy	Clean Berlin DB	93.3%	Arousal dimension
		environment noise+telephone channel	79.8% to 91.8%	
		Clean Berlin DB	80.0%	Valence dimension
		environment noise+telephone channel	60.0% to 74.8%	
[70]	acoustics	emergency call center database	83.49%	Valence dimension
[3]	acoustics	customer service call center database	57.1%	happy, anger, neutral
[26]	acoustics	customer service call center database	54%	pos, neg, neutral

1.3 Problems and issues

According to the comprehensive revision of the literature presented here, there are still several issues related to the emotion recognition from speech that need to be addressed. More feature sets might be proposed and evaluated. Also the effect produced by other non-controlled noise conditions needs to be studied bearing in mind the use of speech enhancement techniques to improve the quality of the noisy speech signals.

1.4 Contribution

This study presents a contribution to solve the current issues for emotion recognition from speech.

Different feature sets are proposed and used for the feature extraction of the emotional content from speech. Five different approaches for feature extraction are proposed including features related to acoustic, non-linear, and wavelet-based measures: (1) conventional acoustic features derived from MFCCs, energy, duration, and the F_0. (2) Features derived from the NLD analysis including the correlation dimension, the largest Lyapunov exponent, the Hurst exponent, the Lempel-Ziv complexity, and entropy measures. (3) Features computed from the time dependent auto-regressive moving average (TARMA) models to model the non stationary process related to the emotions from the speech signals. (4) Different features extracted from the WPT in different decomposition levels. (5) Features derived from three

time-frequency representations based on the wavelet transform such as the traditional representation of the continuous wavelet transform (CWT), the bionic wavelet transform (BWT), and the synchro-squeezed wavelet transform (SSWT). All the approaches for feature extraction are compared to the standard feature set used in the “2009 INTERSPEECH emotional challenge" formed by 384 acoustic features.

For the evaluation of non controlled noise conditions, different environments are evaluated considering both the effect of background noise and the compression of telephony codecs. To evaluate the effect of background noise, the databases used in this study are degraded by different kinds of additive noise such as street noise and cafeteria babble. Other experiment performed consist in the reproduction, and posterior re-capture of the databases in noisy conditions in order to obtain a more natural acoustic scenario. In both cases, the effect of two different speech enhancement algorithms are used to improve the quality of the noisy signals, and the performance of the classification.

For the case of the evaluation of the compression of the speech, the databases are coded by the different state-of-art codecs for speech compression such as the adaptive multi-rate (AMR), the global system for mobile communications (GSM), the SILK codec used by Skype, different models of VoIP codecs, and Opus which is used for WebRTC frameworks.

1.5 Structure of this Study

This work is divided into 7 chapters, chapter 2 contains the description about the main concepts and definitions related to speech and emotions, the chapter describes the speech production process, and gives a definition about *What is emotion*, also provides a review about the main psychological models of emotions. Finally, describes the relationship between emotions and the speech signals, according to some physiological aspects. Chapter 3 contains the description about the main feature estimation approaches used in this work. The chapter defines the acoustic analysis, the non-linear dynamics concepts, the parametric non-stationary analysis using TARMA models, and the wavelet analysis. Chapter 4 describes the non-controlled noise conditions evaluated in this study. It contains two scenarios: evaluation of background noise such as street noise, and cafeteria babble; and evaluation of telephony codecs used both in mobile communications, and in VoIP networks. Chapter 5 defines the methodology evaluated in this study. It contains the description about the classification tasks evaluated in each one of the databases, all the feature sets used, and how are evaluated the non-controlled noise conditions. Chapter 6 shows and discuss the results obtained in this study, using all the features sets proposed. The results of the evaluation of the feature extraction approaches in the non-controlled noise conditions are also discussed. Finally

Chapter 7 provides the conclusions derived from this study, the main contribution to the state of art in the speech emotion recognition problem, and the open questions and issues to be addressed in future work.

Chapter 2

Speech and Emotions

In order to develop suitable methodologies for automatic recognition of emotions from speech, it is important to review important issues given by the physiological aspects, and answer important questions such as what the emotions are, and how are related to the speech.

This chapter describes the main concepts and definitions about emotions and speech. First is described the main aspects related to speech, voice production, and emotions. Then are explained the principal theories for modeling emotions, and how could be classified. Finally, the relations between emotions and speech is described.

2.1 Terminology

2.1.1 Speech Signals and Speech Production Process

Most of features for speech processing tasks such as emotion recognition from speech are based on the human speech production process. For this case the speech production system could be divided into four complementary sub-systems: the respiratory system, the source model, the vocal tract, and the radiation model [71].

The lungs in the respiratory system generate an airflow, which is pressed through the glottis. If the vocal chords are tensed, a quasi-periodic excitation signal with a fixed period is produced. Otherwise, a white noise-like signal is generated. Thus, in the respiratory system could be produced both voiced or unvoiced signals that are passed to vocal tract, which act as filter and gives certain properties to the signal to articulate different tones. The vocal tract are formed by the pharynx, the nasal cavity, and the oral cavity. The vocal tract can be modeled as a series of tubes with similar length but different areas [72]. Finally, the signals filtered by the vocal tract are emitted through the radiation model of the mouth, and the nose. Figure 2.1 summarizes the structure of the speech production process.

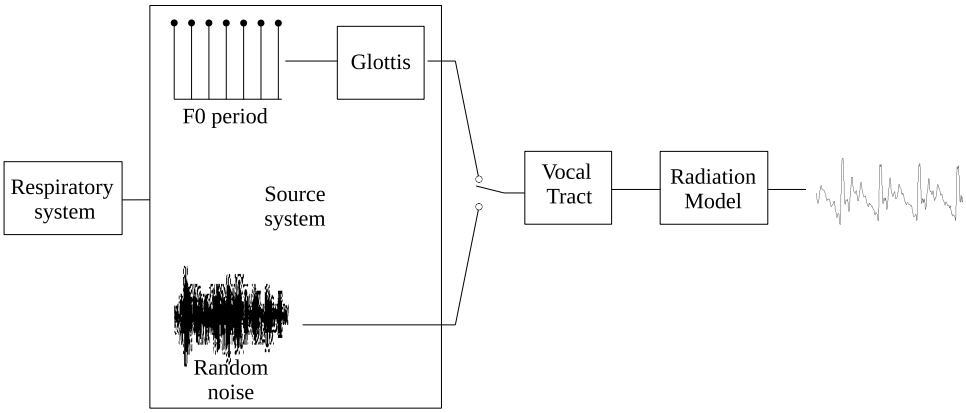

Fig. 2.1 Model of speech production process

2.1.2 Emotions

In psychological research the emotions reflect short term states, usually related and bound to a specific event or action [73]. The emotions reflect the reaction of a human to a specific experience. The theory presented by Scherer in [13] conclude that emotions are the result of the evaluation of events causing specific reactions [71]. The emotions produce organic changes in human body and may affect the facial expression, the speech, and another bio-markers such as the breathing, the heart rate, and the electro-dermal activity.

2.2 Models of Emotion

According to several studies performed in psychology, there have been created many theories about models of emotions in humans. Each one of them capturing and explaining some aspects of the complex phenomenon “emotion" [27]. The most common theories for emotion modeling are described as follows.

2.2.1 Discrete models

These models suggest the existence of primarily discrete emotions such as anger, fear, disgust sadness, surprise and happiness, and the rest of emotions could be considered as a combination of these basic emotions. The primarily emotions are mainly distinguished by their specific stimulus conditions and their corresponding physiological response. In general, the number of basic emotions varies between six and 14 [27]. Particularly, Ekman propose the existence of six basic emotions according to facial expressions, the emotions include happiness, sadness, fear, disgust, anger and surprise. This set of emotions is called the “Big-Six" [74].

2.2.2 Uni-dimensional models

This model consider a single dimension to discriminate the emotions. This dimension could be the “activation" related to the arousal level of the emotion, or the “valence" dimension, related to the subjective feeling of pleasantness or unpleasantness of the subject.

For the case of the activation dimension, the major difference between emotions is the degree of arousal from very low to very high. For example emotions such as sadness and happiness can be discriminated according to this dimension because the low arousal of sadness and the high arousal of happiness. Emotions such as happiness and anger are not discriminated according to this dimension because both of them are high arousal emotions. For the case of the “valence" dimension, the most important difference between the emotions is the degree of pleasantness, which ranges from negative or disagreeable feelings to positive or agreeable. In this way, positive emotions such as happiness and calm can be differentiated from negative emotions such as sadness and anger [75].

2.2.3 Multi-dimensional models

In this model the emotional state can be represented as coordinates in a multidimensional space. In 1954, Schlosberg in [76] shows the relevance of two dimensions called “valence" and “arousal", which create a plane where the emotions are represented. Cowie in [77] call this two-dimensional space “activation-evaluation" plane highlighting that this representation describes the emotional states in an easier and more treatable way, than using several discrete emotions, and it is specially attractive to the research oriented to affective computing [77, 78]. Figure 2.2 shows an example of this plane to represent the different emotions.

In related works has been found that the recognition of emotions in the arousal plane provides better results than the recognition in the valence plane [44]. This fact motivates further research in characterization in order to improve the performance in the recognition of emotions in the valence dimension.

2.2.4 Fear-type Emotions

In the last years the interest of the research community in automatic emotion recognition from speech for security applications have increased and have been focused on the detection of “fear-type" emotions such as anxiety, fear, anger, disgust, desperation and those reflecting that the life or the human integrity are at risk [7, 79, 80].

Fear-type emotions appear in abnormal situations, specially in unplanned events where human life could be threatened. In these situations could be detected the hot anger in a

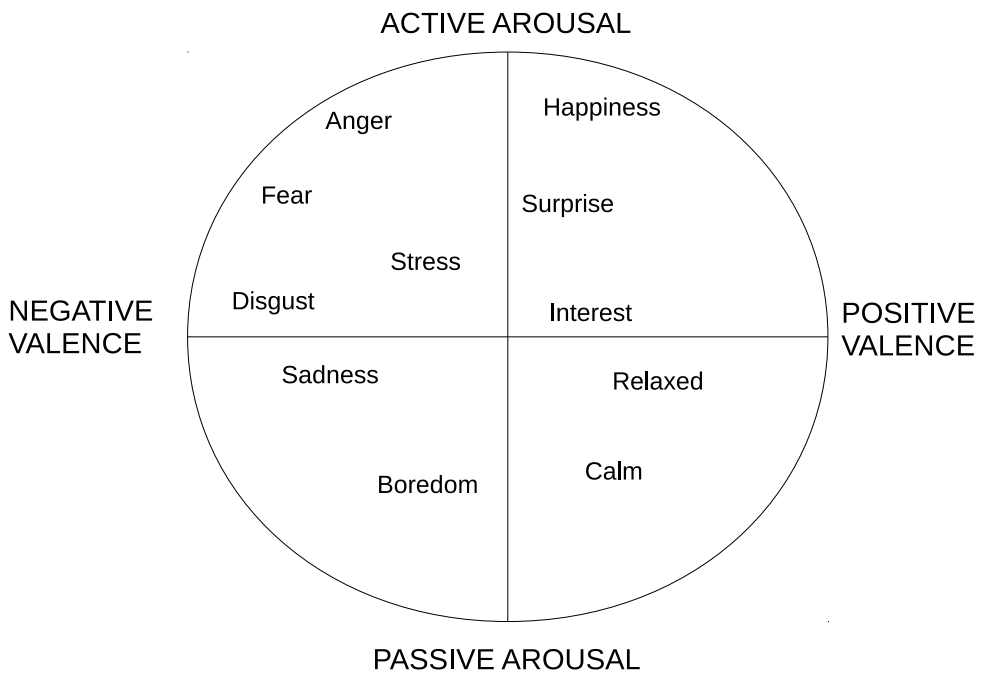

Fig. 2.2 Two dimensional representation of emotions in arousal-valence plane

potential aggressor or the fear in a potential victim. Even more complex emotional states could also be detected ranging from worry to panic [7].

The recognition of fear-type emotions from speech have been considered in several applications: detection of stress in drivers [81], where the speech signals from 15 drivers were considered to classify three levels of stress with an accuracy of up to 88.2%. In real emergency call-centers [70] to detect anxiety, stress, relief, annoyance, and others both for users and agents with and accuracy of up to 83%. In public surveillance systems [7] to discriminate between emotions related with fear and neutral state with an accuracy close to 70%.

2.3 Emotions from Speech

There are various physiological changes associated with emotions that affect different aspects of speech, producing effects on the breathing, phonation, articulation, and prosody. Several works have been focused on finding the vocal signs of emotion [16]. In general, the arousal dimension might affect measures related to effort such as the intensity, the mean voice pitch, and the speech rate. The tremor associated with fear and anger would be expected to produce

oscillations in pitch. Also it has been suggested that unpleasantness is related to stress of the vocal tract walls, altering the spectral balance.

According to physiological studies performed by Williams and Stevens in [82], the sympathetic nervous system is aroused with emotions related to anger, happiness, and fear, which induces an increase in the sub-glottal pressure, a dryness of the mouth, and occasional muscle tremor. These aspects produces louder and faster speech, which is characterized by strong high frequency energy, a higher average pitch, and a wider pitch range. On the other hand the low arousal emotions such as sadness and boredom affect the parasympathetic nervous system, producing speech characterized by slow rate, low pitch, and with little high frequency energy.

Other studies suggest that the emotional content in speech is related to the voice quality [1]. However, there is an ambiguity and subjectivity in the description of voice quality terms such as tense, harsh, and breathy. Several studies debate whether tense voice is associated with anger, joy, and fear; lax voice is related to sadness, and breathy voice is associated with both anger and happiness, while sadness is associated with a resonant voice quality [48].

Table 2.1 summarizes the relationship between emotions and the speech parameters. Note that the more affected speech parameters according to the emotional content are related to the F_0, and the energy content. The relation of those parameters with the emotional content is described as follows.

Table 2.1 Emotions and speech parameters, from [1]

Feature	**Anger**	**Happiness**	**Sadness**	**Fear**	**Disgust**
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
F_0	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
F_0 Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Energy content	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy chest blaring tone	Breathy	Resonant voicing	Irregular chest tone	Grumble
F_0 Changes	Abrupt on stressed	smooth, upward inflections	Downward inflections	Normal	Wide, downward inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

Emotions in the F_0

The contour of the F_0 has been found as a great marker to differentiate emotions from speech. In [83] was shown that the neutral speech produces a narrower F_0 range than the emotional speech. The F_0 for angry speech has also a high median, wide range, and high rate of change [84] than the others emotions. The vowels of angry speech exhibit the highest F_0, and have downward slopes relative to the neutral speech and other emotions [13]. The F_0 for fear has a high median, wide range, and a moderate rate of change [84].

For the case of low arousal emotions such as sadness or disgust, it has been found that the F_0 exhibits a lower mean and a narrower range for sadness; and a low median, wide range, and lower rate of change for disgust [84, 85].

Figure 2.3 shows an example about the difference in the F_0 contour produced by the emotions. Figure contains the F_0 contour of angry, and neutral speech recordings of the Berlin database. The recordings are uttered by the same speaker, who pronounces the same sentence in the two emotional states. Note that the F_0 values are higher for anger than for neutral speech. Also can be observed a difference in the range of the F_0, for anger the F_0 ranges from 140 to 300 Hz, while for neutral speech ranges from 100 to 160 Hz.

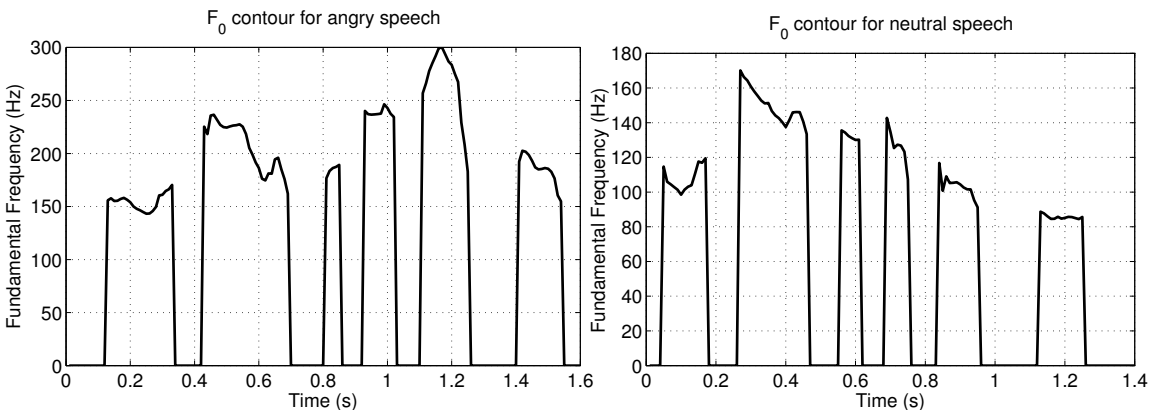

Fig. 2.3 F_0 contour for angry (left) and Neutral speech (right)

Emotions in the energy content

The energy content of the speech signal is the another marker more affected according to the emotion of the speaker. In [85] was found that angry speech exhibits a higher energy envelope than the others emotions. A similar result was reported for happiness; while sadness is associated with a decreased energy [13]. These characteristics agree with the reported in the models of emotions: the high arousal emotions such as anger and happiness generally exhibit higher energy than the observed in low arousal emotions.

Chapter 3

Features for Emotion Recognition from Speech

One of the main objectives in speech processing tasks such as automatic emotion recognition is to find suitable features to represent the emotional state of the speaker. In section 1.2 an overview of the current characterization approaches was presented. In this chapter a description of the proposed feature extraction methods is provided.

The first step to find the adequate features is the pre-processing of the speech signal which includes among others the normalization, segmentation, and noise reduction techniques. Then the feature extraction process is performed. Five different approaches for characterization of speech signals are proposed: (1) the conventional acoustic analysis used to represent both the speech production process or the human auditory system, (2) the nonlinear analysis to represent the complexity and the long term dependence of time series, (3) the non-stationary analysis provided by parametric representations such as time dependent ARMA models, (4) the features derived from wavelet transform which enables to analyze the signals in a time-frequency multi-resolution perspective, and (5) different time-frequency representations derived from the wavelet transform.

3.1 Pre-processing

This stage consider different methods applied on the speech signal before the feature extraction process. The main objective of pre-processing is to conditioning, normalize, remove perturbation caused by the recording conditions, and segment the speech signals. In this stage the speech enhancement (SE) algorithms are used to remove the background noise, and the segmentation into voiced and unvoiced segments is performed to analyze the representation

capability of those segments separately. The SE algorithms are explained with details in appendix A, and the segmentation procedure is explained as follows.

3.1.1 Segmentation

This pre-processing step allows to separate the speech signal into different segments and remove non-relevant information such as silence parts. One of the main segmentation procedures for speech processing tasks is the voiced-unvoiced segmentation. This separation has proven to be useful in the recognition of emotions in speech and other paralinguistic aspects [7, 32].

Voiced-Unvoiced Segmentation

The speech signals is formed by two kind of frames: The voiced segments, which are produced by the vocal folds vibration in a quasi-periodic way due to the glottis closure, and the unvoiced segments, which are produced by turbulent air flow at constriction, or by the release or closure in the vocal tract.

Figure 3.1 shows the temporal view of the voiced and unvoiced segments. Note the oscillatory behavior of the voiced segment, and the similarity between the unvoiced segment with noise.

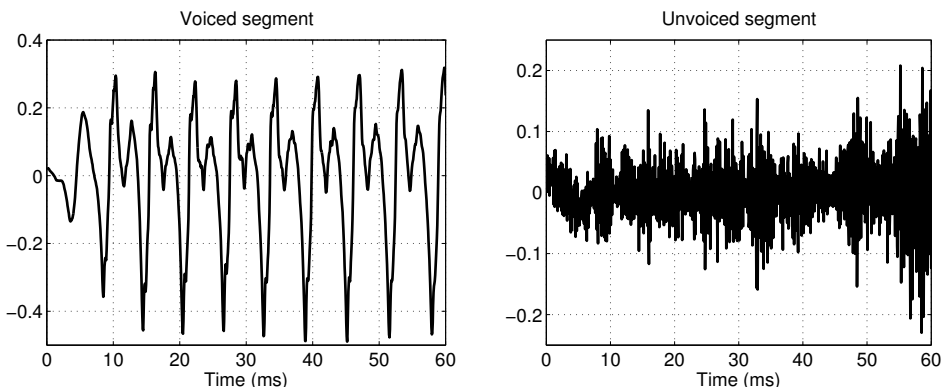

Fig. 3.1 Voiced segment (left). Unvoiced segment (right)

One of the main methods for voiced-unvoiced segmentation is based on the presence or not of the fundamental frequency of speech (F_0) in short time frames. The more used method for the estimation of F_0 is the based on auto-correlation function of speech signal [86]. The

auto-correlation of the speech signal $x[n]$ is expressed according to Equation 3.1. N is the length of analyzed frame.

$$\mathbf{r}_x[n] = \sum_{k=0}^{N} x[k]x[n-k] \tag{3.1}$$

Then, the normalized auto-correlation is calculated using Equation 3.2. Where $r_w[n]$ corresponds to the auto-correlation function of a Hanning, or Gaussian window.

$$\mathbf{r}_{xn}[n] = \frac{r_x[n]}{r_w[n]} \tag{3.2}$$

The fundamental period T_0 corresponds to the distance between two consecutive peaks in the $r_{xn}[n]$ function. Finally F_0 is calculated as $F_0 = 1/T_0$. This method is implemented in Praat software [87], which has become in a standard package to analyze speech signals and is widely used for the research community [31, 88, 89].

3.2 Acoustic Analysis

The acoustic analysis includes the conventional features used in speech processing tasks. The acoustic features can be divided into prosody analysis, which is related to measures derived from the contour of the fundamental frequency, energy, and duration; perturbation measures such as jitter and shimmer, spectral features such as the energy content in different frequency bands, cepstral features such as MFCC, and voice quality features such as noise measures.

3.2.1 Prosody Analysis

Features related to prosody are derived from measures calculated from the contour of the F_0, the energy content, and different duration patterns, and are one of the most used for recognition of emotions from speech [31, 32, 34].

The measures related to the contour of the F_0 include statistics functions such as the mean, maximum, minimum, range, standard deviation, skewness, kurtosis, median, among others. Those measures are calculated in order to evaluate intonation changes in the emotional speech. The same statistics also can be calculated in the derivatives of the F_0 contour to measure the dynamics of the intonation along the time. The measures of the contour of the energy are the same to the calculated for the F_0 contour to evaluate changes in the intensity pattern due to the emotional state of the speaker. Finally, the duration analysis is described

according to the the relation between the duration of voiced, unvoiced, and silence segments, according to the following Equations.

$$dur_1 = \frac{duration\ silence}{duration\ voiced\ +\ duration\ unvoiced} \tag{3.3}$$

$$dur_2 = \frac{duration\ voiced}{duration\ unvoiced} \tag{3.4}$$

$$dur_3 = \frac{duration\ unvoiced}{duration\ voiced\ +\ duration\ unvoiced} \tag{3.5}$$

$$dur_4 = \frac{duration\ voiced}{duration\ voiced\ +\ duration\ unvoiced} \tag{3.6}$$

$$dur_5 = \frac{duration\ voiced}{duration\ silence} \tag{3.7}$$

$$dur_6 = \frac{duration\ unvoiced}{duration\ silence} \tag{3.8}$$

3.2.2 Perturbation Measures

These features provide a measure about the temporal variation of frequency and amplitude in the phonation process. In this work are considered jitter and shimmer. These features have been identified as indicators about stress, and emotions in speech [90].

Jitter

Jitter provides information about the temporal variation of F_0 in the voiced segments. This feature is calculated using Equation 3.9. $F_0[j]$ is the fundamental frequency in the frame j, and $\overline{F_0}$ corresponds to the average value of the fundamental frequency measured in the previous three frames.

$$Jitter[j](\%) = 100 \cdot \frac{|F_0[j] - F_0[j-1]|}{\overline{F_0}} \tag{3.9}$$

Shimmer

Shimmer provides information about the amplitude changes in speech signal along the utterance. This feature is calculated using Equation 3.10. $A[j]$ is the maximum amplitude of

the frame j, and $\overline{A}$ corresponds to the average value of the maximum amplitude measured in the previous three frames.

$$Shimmer[j](\%) = 100 \cdot \frac{|A[j] - A[j-1]|}{\overline{A}} \tag{3.10}$$

3.2.3 Spectral and Cepstral Analysis

Features from this group provide information about the short time frequency content of the speech signal. In this study are considered the Mel frequency cepstral coefficients (MFCC), and the energy distributed in frequency bands according to the Bark scale.

Mel Frequency Cepstral Coefficients

The MFCC are one of the most used features in speech processing tasks. They are based on the perception of human auditory system according to the Mel scale, proposed by Stevens in [91] to reflect the non-linear relationship between the frequency of a tone and the perceived pitch. This relationship is given by Equation 3.11.

$$f_{Mel} = 1127.01048 \cdot log_e \left(1 + \frac{f_{Hz}}{700}\right) \tag{3.11}$$

The procedure to calculate the MFCC is as follows. First the speech signal is enframed using a Hamming window. Then the fast Fourier transform (FFT) is calculated to obtain the power spectrum of the signal. Subsequently, a filter bank in Mel scale is created to obtain a higher resolution at lower frequencies. Figure 3.2 shows the structure of such filterbank. Finally the log-energy of the output signals from each filter is calculated, and the discrete cosine transform (DCT) is applied.

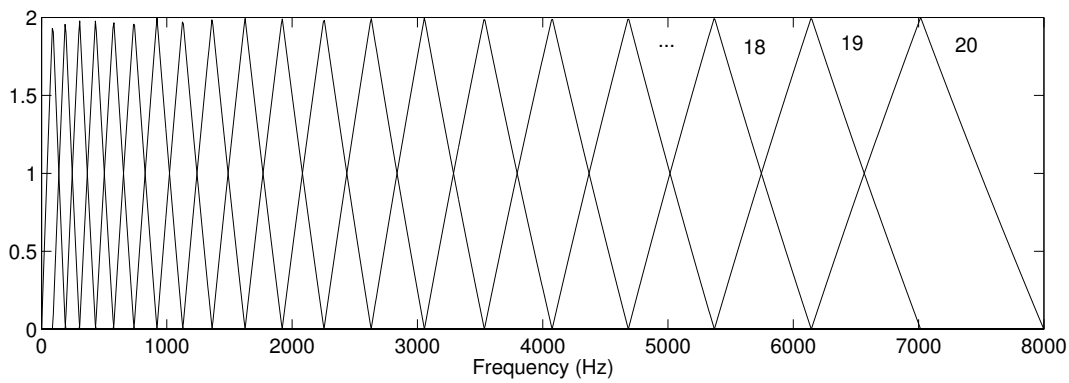

Fig. 3.2 Mel filterbank consisting of 20 triangular filters

In order to compensate the effects of speaker and channel, the speech based cepstral mean subtraction (CMS) is considered [92]. In this case, the cepstrum of the speech signal is normalized using Equation 3.12. y_t is the cepstrum of the original signal, z_t is the cepstrum of the normalized signal, and m_{spe} is calculated according to Equation 3.13. Where w_t corresponds to the probability $p(speech|y_t)$, or the output of a VAD [92].

$$\mathbf{z}_t = \mathbf{y}_t - m_{spe} \tag{3.12}$$

$$m_{spe} = \frac{\sum_t \mathbf{w}_t \mathbf{y}_t}{\sum_t \mathbf{w}_t} \tag{3.13}$$

Energy Content in Bark Scale

The Bark scale is a psycho-acoustic scale proposed by Zwicker in [93]. The scale ranges from 1 to 25. The division is performed according to the concept of the critical bands in the human auditory system. The conversion between frequency measured in Hertz and Bark scale is given by Equation 3.14. The Bark scale frequency bands are almost linear below 1 kHz, while from frequencies superior to 1 kHz the scale grows exponentially, which yields a perceptual filter-bank.

$$f_{Bark} = 13 \cdot arctan(0.00076 f_{Hz}) + 3.5 arctan\left(\left(\frac{f_{Hz}}{7500}\right)^2\right) \tag{3.14}$$

In this work the log-energy of the speech signal distributed in the 25 critical bands is calculated. The process to compute these energies consists in calculate the short time Fourier transform (STFT) of the speech signal, separate the corresponding spectrum in 25 frequency bands according to the Bark scale, and calculate the log-energy of each band [94].

3.2.4 Noise Measures

The noise measures have been used commonly to quantify the turbulent noise in the vocal tract. This kind of measures have been classically used for the assessment of voice quality [52], which is affected due to the emotional state of the speaker [95]. In this study have been implemented classical noise measures such as the harmonic to noise ratio (HNR) and the normalized noise energy (NNE), which allow to evaluate the loss of harmonic structure in speech when the speaker is over threat conditions. The glottal noise excitation ratio (GNE) is also considered to measure the influence of the emotions in the vocal folds vibration pattern.

Harmonic to Noise Ratio

The HNR is the relationship between the energy of the harmonic content of the speech signal with the additive noise produced in the vocal tract. This feature can be considered a measure for the degree of periodicity of a voiced signal. The method for the estimation is based on the proposed by Yumoto in [96].

Normalized Noise Energy

The NNE is the relationship between the energy of noise with the total energy of the speech signal, both measured in dB, and can be used to measure the degree of hoarseness in the speech signal. The process for the estimation was proposed by Kasuya in [97].

Glottal to Noise Excitation Ratio

The GNE quantifies the relationship between the vocal excitation due to the vocal folds vibration with the excitation produced by the noise turbulent in vocal tract. This feature was proposed by Michaelis in [98]. The GNE is more robust than the other noise measures due to its estimation does not require the calculation of the fundamental period.

3.3 Non-Linear Dynamics Analysis

The speech production model involves some non-linear process such as the non-linear pressure flow in glottis and the non-linearity that occurs in the vocal fold collision [52]. These processes can not be characterized using classical measures. In order to resolve this problem, the non-linear dynamics (NLD) analysis has been established as a mathematical alternative for the analysis of this kind of process. The NLD analysis describes the temporal evolution of a system through a multiple dimension space on which the speech signal is reconstructed.

The use of NLD or complexity measures in speech processing tasks has been increased in the last years and have proved to be useful in the analysis of voice quality and the voice pathology detection [52–54], also have been used in speech emotion recognition [55–57]. Particularly, fear and anger have shown more complexity than neutral speech recordings. The reason could be refereed to that in such emotions, people tend to use more fricative sounds than in neutral state. Those fricative sounds are more noisy and complex than the voiced sounds [55]. Features related to the NLD include the correlation dimension (CD), the largest Lyapunov exponent (LLE), the Hurst exponent, the Lempel-Ziv complexity, and several entropy measures.

3.3.1 Embedding Process and Phase Space

The NLD analysis begins with the reconstruction of the state space of the speech signal. In this way, topological features of the phenomenon can be analyzed. The state space is known as phase space or attractor. The most common technique for reconstruction of the attractor is the method proposed by Takens in [99]. A time series $x[n]$ can be represented in a new space defined by Equation 3.15. τ is the time delay and m is the embedding dimension. τ can be determined by calculating the first minimum of the mutual information and the false neighbors method can be used to estimate m [100].

$$\mathbf{X}[k] = \{\mathbf{x}[k], \mathbf{x}[k+\tau], \mathbf{x}[k+2\tau], \cdots, \mathbf{x}[k+(m-1)\tau]\} \tag{3.15}$$

As illustration, Figure 3.3 shows a sinusoidal signal and to its reconstructed attractor, and Figure 3.4 shows the attractor generated by a speech signal of vowel /a/.

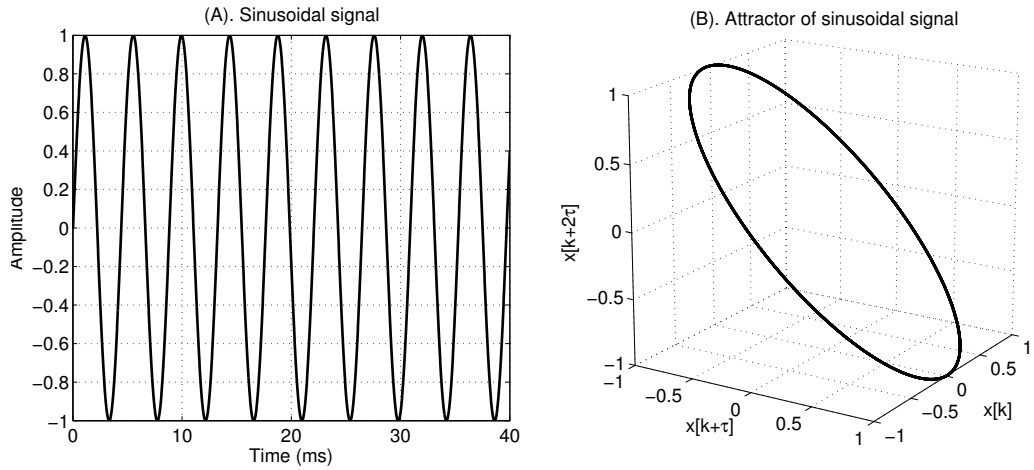

Fig. 3.3 Attractor of sinusoidal signal

3.3.2 Correlation Dimension

This feature allows the estimation of the exact space that is occupied by the reconstructed vector in the phase space. The CD is an indicator about the complexity and dimensionality of speech signal and can be related to the perturbation measures [55]. To estimate the CD, the correlation sum (CS) is defined using the Equation 3.16. Where Θ is the Heaviside step

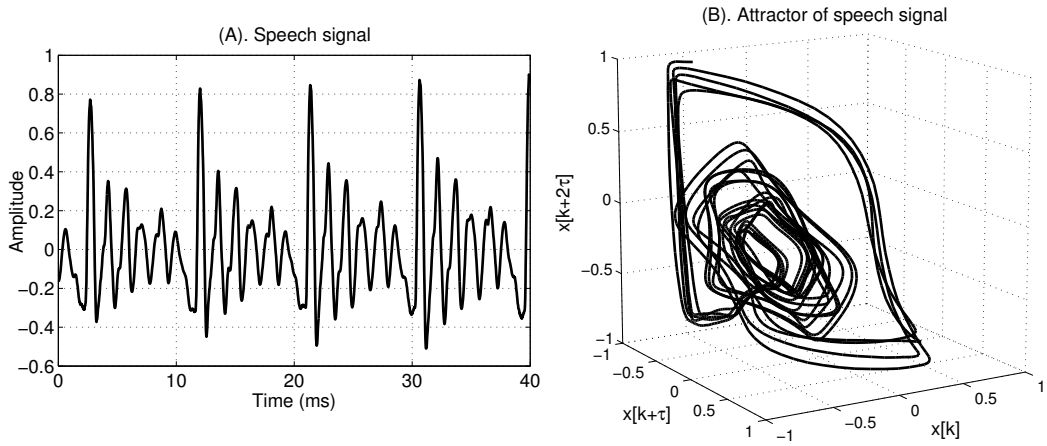

Fig. 3.4 Attractor of speech signal

function and ε is related to the radius of a hyper-sphere in which the points $\mathbf{x}_i$ and $\mathbf{x}_j$ may be inside.

$$CS(\varepsilon) = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta(\varepsilon - |\mathbf{x}_i - \mathbf{x}_j|) \tag{3.16}$$

In [101] is considered that for small values of ε, $CS(\varepsilon)$ can be calculated according to Equation 3.17, and the CD can be estimated using Equation 3.18. In order to estimate CD, it is necessary to plot $log(CS(\varepsilon))$ versus $log(\varepsilon)$. The slope of the resulting line, when a linear regression is performed, corresponds to the CD [55].

$$CS(\varepsilon) = \lim_{\varepsilon\to 0} \varepsilon^{CD} \tag{3.17}$$

$$CD = \lim_{\varepsilon\to 0} \frac{log(CS(\varepsilon))}{log(\varepsilon)} \tag{3.18}$$

3.3.3 Largest Lyapunov Exponent

This measure quantifies the exponential divergence of neighbor trajectories in a phase space. In other words, this feature is an indicator about the aperiodicity of a speech signal [55]. After the phase space reconstruction, the nearest neighbor of every point in the trajectory is located. The nearest neighbor $\mathbf{x}_{\hat{j}}$ minimizes the euclidean distance $d_j(0)$ to the point $\mathbf{x}_j$. In the time series the data must be separated a distance larger than the signal average period in order to guarantee that the neighbors data are in different trajectories in the phase space. The LLE is estimated as the mean separation rate between the nearest neighbors according to

Equation 3.19. λ_1 corresponds to LLE, $d(t)$ is the mean divergence in the time instant t, and C is a constant used for normalization.

$$d(t) = Ce^{\lambda_1 t} \tag{3.19}$$

3.3.4 Hurst Exponent

This feature expresses the long term dependence of a time series. The HE is defined according to the asymptotic behaviour of the rescaled range of a time series as a function of a time interval. This feature was introduced by Hurst in [102]. The method for estimation consist of dividing the time series into intervals of size L and calculating the average ratio between the range R with the standard deviation S of the time series. HE can be estimated as the slope of the curve, as can be observed in Equation 3.20.

$$E\left[\frac{R(L)}{S(L)}\right] = CL^{HE} \tag{3.20}$$

HE can be used to represent the emotional state of speech according to the arousal level of the signal as follows [56]:

- $0 < HE < 0.5$ represents high arousal emotions such as anger or happiness.
- $HE \approx 0.5$ represents neutral speech.
- $0.5 < HE < 1$ represents low arousal emotions such as sadness or boredom.

3.3.5 Lempel Ziv Complexity

This feature establishes a measure about the degree of disorder of spatio-temporal patterns in a time series [103]. The signal is transformed into binary sequences according to the difference between consecutive samples, and the LZC reflects the rate of new patterns in the sequence, and ranges from 0 (deterministic sequence) to 1 (random sequence). LZC distribution shows values nearer to 1 for fear and anger speech, than in case of neutral speech [55].

3.3.6 Entropy Measures

Entropy describes the complexity of a system. Two different entropy measures are used in this work: the non-normalized Shannon entropy calculated using Equation 3.21, and the log-energy entropy calculated using Equation 3.22. b is the number of bins used to estimate

the probability density function of the signal, and $p(j)$ is the probability of the j-th bin in the histogram created.

$$S_E = -\sum_{j=1}^{b} \left[p(j)^2 \cdot log \left| p(j)^2 \right| \right] \tag{3.21}$$

$$LE_E = -\sum_{j=1}^{b} log \left| p(j)^2 \right| \tag{3.22}$$

3.4 Parametric Non-stationary Analysis

The speech production process involves several physiological aspects such as turbulent noise caused by an air escape through the glottis and the laryngeal tensions involved in breathy and whisper phonation, which may carry important paralinguistic information related to the emotion of the speaker [104]. These processes produce a non-stationary behavior in speech signal that cannot be characterized properly using the conventional acoustic features due to the assumption of local stationarity [105]. In order to model these phenomena, non-stationary modeling should be considered. The non-stationary analysis allows both to evaluate the time-dependence and to represent the spectral evolution of the signal [106]. Non-stationary models can be classified as parametric and non-parametric [106].

Parametric methods are based on parametrized representations of the time dependent auto-regressive moving average (TARMA) models which are able to represent abrupt changes in the spectral evolution of the signals [106]. Such methods can be classified into three approaches according to the “structure" of their parameters: (1) the unstructured parameter evolution methods which are characterized by low parsimony and slow tracking on the dynamics, (2) the stochastic parameter evolution methods characterized by slow and medium tracking of dynamics, and (3) the deterministic parameter evolution characterized by high parsimony and fast or slow tracking depending on the estimated parameters [106]. Figure 3.5 summarizes the classification of parametric and non-parametric methods for non-stationary signal modeling.

TARMA models have been applied on the modeling and simulation of earthquake ground motion [107], modeling and detection of damage in mechanical structures with time-dependent dynamics [106, 108], and modeling of speech and other bio-signals [105, 106, 109]. These previous attempts have demonstrated the usefulness of TARMA models as representations of non-stationary processes and makes them very appealing for the automatic classification of emotions from speech. In this work we have focus on the smoothness priors TARMA (SP-TARMA) to perform the parametric non-stationary analysis.

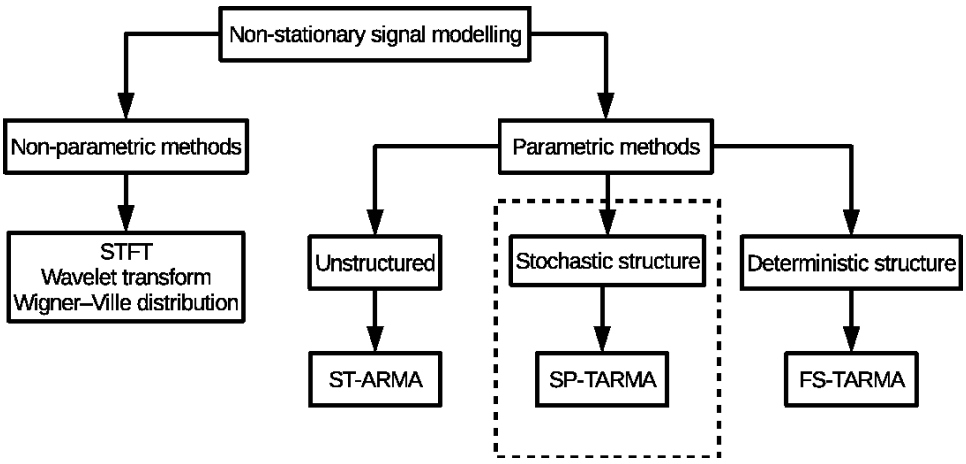

Fig. 3.5 Classification of methods for non-stationary signal modeling. ST-TARMA: short time TARMA; SP-TARMA: smoothness prior TARMA; FS-TARMA: functional series TARMA.

3.4.1 SP-TARMA Models

A TARMA(n_a, n_c) model is defined by Equation 3.23, which includes the auto-regressive (AR) and the moving average (MA) components. n_a and n_c are the orders of the AR and MA models. $\mathbf{e}[n]$ is an unobservable "innovations" sequence with zero mean, and time-dependence variance $\sigma_e^2[n]$, and $\mathbf{a}_i[n]$, $\mathbf{c}_i[n]$ are the parameters of AR, and MA models [106].

$$\mathbf{x}[n] + \underbrace{\sum_{i=1}^{n_a} \mathbf{a}_i[n] \cdot \mathbf{x}[n-i]}_{AR\ part} = \mathbf{e}[n] + \underbrace{\sum_{i=1}^{n_c} \mathbf{c}_i[n] \cdot \mathbf{e}[n-i]}_{MA\ part} \tag{3.23}$$

Stochastic parameter evolution TARMA imposes an stochastic structure in the time-dependence of the parameters. In this case the evolution of the parameters $\mathbf{a}_i$, $\mathbf{c}_i$ is subjected to stochastic smoothness constraints. The constraints are referred to smoothness priors TARMA (SP-TARMA). In this case the model is referred to SP-TARMA (n_a, n_c, k). Where k denotes the order of the difference equations that describe the evolution of the parameters as is shown by Equations 3.24 and 3.25. $\mathbf{w}_{ai}[n]$ and $\mathbf{w}_{ci}[n]$ are Gaussian sequences with possibly time-dependent variances. B is the back-shift operator, which operates at $B^k\mathbf{a}[n] = \mathbf{a}[n-k]$.

$$\Delta^k \mathbf{a}[n] = (1-B)^k \mathbf{a}[n] = \mathbf{w}_{ai}[n] \tag{3.24}$$

$$\Delta^k \mathbf{c}[n] = (1-B)^k \mathbf{c}[n] = \mathbf{w}_{ci}[n] \tag{3.25}$$

Thus the smoothness constraints for $k = 1$ and $k = 2$ take the form

$$k = 1: \ (1-B)^1\mathbf{a}[n] = \mathbf{a}[n] - \mathbf{a}[n-1]$$

$$k = 2: \ (1-B)^2\mathbf{a}[n] = (1-2B+B^2)\mathbf{a}[n] = \mathbf{a}[n] - 2\mathbf{a}[n-1] + \mathbf{a}[n-2]$$

The orders of a SP-TARMA model are determined by two possible criteria, the Akaike information criterion or the Bayesian information criterion. Both are based on the superposition of the negative log-likelihood function of the model and penalize the complexity of the model in order to discourage the over-fitting of the model [106]. The orders of the model are such that minimize the criteria. In this work, the Bayesian information criterion is used to select n_a and n_c, and the minimum residual sum squares (RSS) is used to select the parameter k. As illustration, Figure 3.6 shows the TARMA models extracted both for a voiced and an unvoiced segment. Note that for the case of the voiced segment there is no a time-dependence of the parameters of the model, which indicate the stationarity od the segment, contrary to the unvoiced segment, where a time-dependence of the parameters $\mathbf{a}[n]$ and $\mathbf{c}[n]$ is observed.

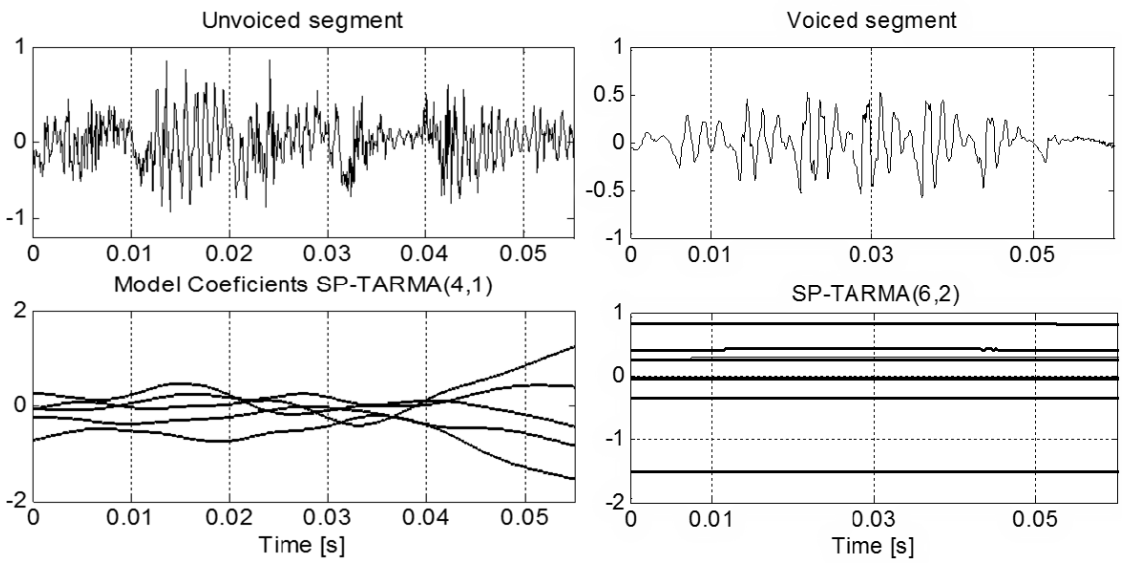

Fig. 3.6 Unvoiced segment with its TARMA model (left) and Voiced segment with its TARMA model (right)

3.5 Wavelet Analysis

The STFT is a time-frequency representation based on the analysis of short-time fixed length frames. Such fixed length does not allow analyze the details and fast changes in non-stationary signals. In this way, the wavelet transform (WT) is introduced as an alternative to allow the representation, decomposition, and reconstruction of signals that present abrupt changes in their spectral evolution. The WT allows a time-frequency multi-resolution analysis (MRA) based on the decomposition of the signal into time-variable length frames according to the frequency changes in the signal. Figure 3.7 shows the mean difference between the time-frequency representation of a signal using the STFT, and the discrete version of the WT (DWT).

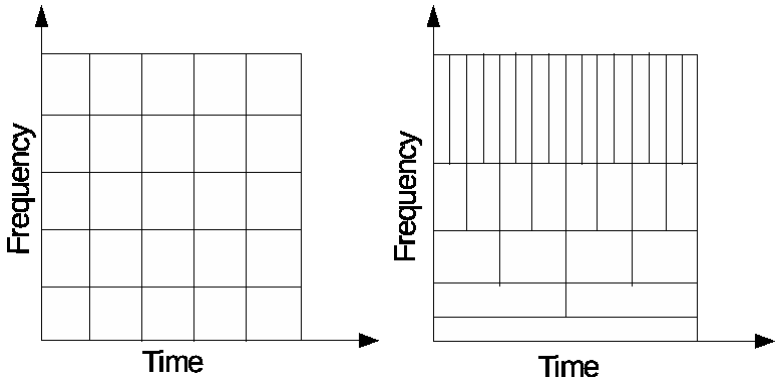

Fig. 3.7 Time-frequency representation of STFT (left) and DWT (right)

Unlike the Fourier analysis, where the base functions for the decomposition are sinusoid signals, in the WT the base functions $\psi_{s,u}(t)$ are small waves of limited duration known as Wavelets, whose energy is located around a fixed point. These waves are scaled and translated in order to create a complete base of the decomposition space. Formally, the WT of a signal $x(t)$ is given by Equation 3.26. Where s defines the scale, and u the translation.

$$WT(u,s) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) exp \left[-j\omega_0 \left(\frac{t-u}{s} \right) \right] dt \tag{3.26}$$

3.5.1 Discrete Wavelet Transform

In this case, the scale and translation parameters s and u are discrete, and the WT correspond to a series of wavelet coefficient called the discrete wavelet transform (DWT). The discretization of s is given by a exponential sampling as $s = s_0^j$. The discretization of u depends of the

value of scale as $u = ku_0 s_0^j$. Based on this discretization, the wavelets functions translated, and scaled are given by Equation 3.27.

$$\psi_{j,k}(t) = s_0^{-j/2}\psi(S_0^{-j}t - ku_0) \tag{3.27}$$

If the scale and translation parameters are sampled in power of 2 ($s_0 = 2$), is obtained a better representation, and the functions are called dyadic wavelets. In this case the signal $x(t)$ is represented as a series of *approximation* coefficients $\mathbf{a}_j[n]$ related to the lower frequencies, and *detail* $\mathbf{d}_j[n]$ related to the higher frequencies in multiple resolutions. The details coefficients represent the information from a high resolution to the lower resolution, and the approximation corresponds to the lowest resolution in frequency. The DWT is represented by the set of detail coefficients in all resolutions and the approximation coefficients in the lowest resolution. The method to calculate these coefficients was developed by Mallat in [110]. The method consist in the application of a series of discrete filters $\mathbf{h}[n]$ and $\mathbf{g}[n]$ called conjugate mirror filters, which must satisfy the condition $\mathbf{g}[n] = (-1)^{1-n}\mathbf{h}[1-n]$. Figure 3.8 illustrates the decomposition of DWT.

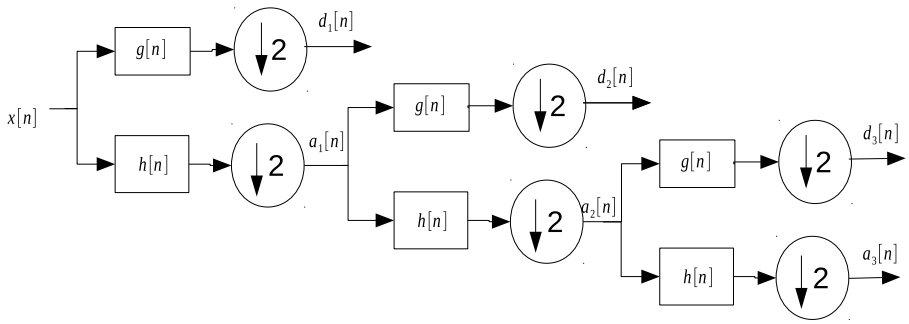

Fig. 3.8 Discrete wavelet transform decomposition

3.5.2 Wavelet Packet Transform

In the DWT decomposition the signal is decomposed in two frequency bands represented by the approximation and detail coefficients, respectively. Only the approximation is used for further decomposition. Hence, the DWT provides a left recursive binary tree structure, as can be observed in Figure 3.8. In the wavelet packet tree (WPT), both the approximation and detail coefficients are decomposed in two sub-bands, which provide a balanced binary tree structure, as can be observed in Figure 3.9. In this case the wavelet coefficients are

represented by $W_{j,k}$, where j corresponds to the decomposition level, and k is the number of the node in each decomposition level.

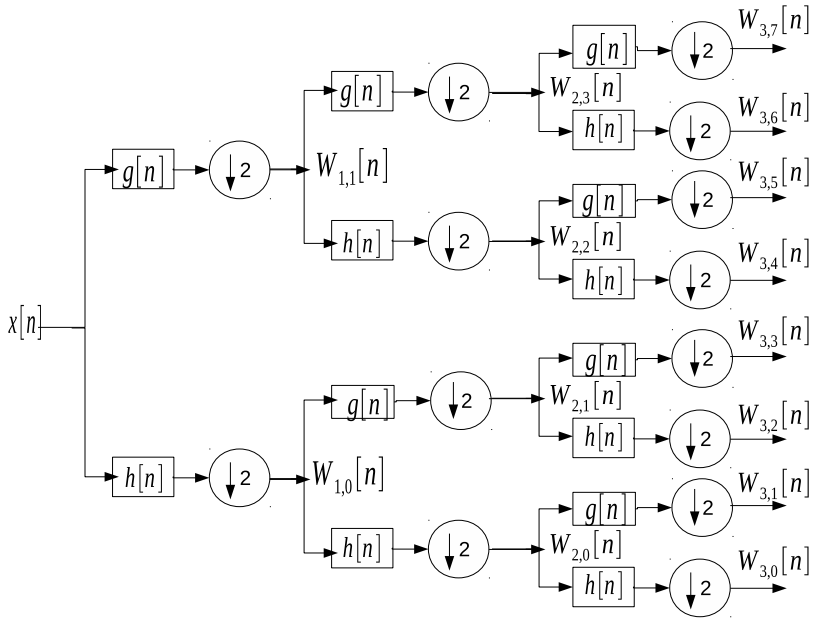

Fig. 3.9 Wavelet packet transform decomposition

3.5.3 Wavelet Perceptual Packet

The WPT can be adjusted to approximate the human auditory system [111] using the concept of critical bands of the Bark scale, forming a perceptual filter-bank called wavelet perceptual packet (WPP). The Bark scale can be used to obtain a wavelet representation formed by 17 critical bands. The structure of the tree of WPP consider 17 decomposition in 5 levels, as can be observed in Figure 3.10.

3.5.4 Bionic Wavelet Transform

The bionic wavelet transform (BWT) was developed based on a model of the active auditory system [112], which made it appropriate for different speech processing tasks. This transform has been widely used for the design of cochlear implants, and SE algorithms [113].

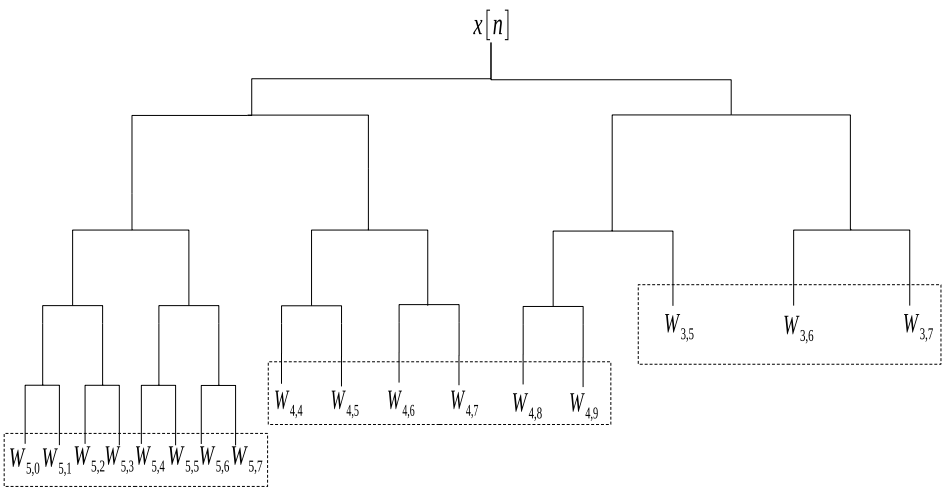

Fig. 3.10 Wavelet perceptual packet decomposition

Formally, the BWT is a time adaptive wavelet transform based on the Morlet wavelet designed specially to model speech signals. The BWT is defined according to Equation 3.28 [112].

$$BWT(u,s) = \int_{-\infty}^{\infty} x(t) \frac{1}{\lambda(u+\Delta u)\sqrt{s}} \psi^* \left(\frac{t-u}{s\lambda(u+\Delta u)} \right) exp\left[-j\omega_0 \left(\frac{t-u}{s} \right) \right] dt \quad (3.28)$$

The main difference between the BWT from Equation 3.28 and the WT from Equation 3.26 is the introduction of the time-adaptive parameter $\lambda(u+\Delta u)$. The envelope of the BWT mother function can be adjusted by this parameter.

The function $\lambda(u+\Delta u)$ is derived from the active auditory model, and it is expressed according to Equation 3.29. Where α is a saturation constant, and β and γ are the gains of the model. In this study, it is considered $\alpha = 0.8$, $\beta = 0.87$, and $\gamma = 0.45$, as in related works [112, 113].

$$\lambda(u+\Delta u) = \frac{1}{1-\alpha\frac{\beta}{\beta+|BWT(u,s)|}} \cdot \frac{1}{1+\gamma\left|\frac{\partial}{\partial t}BWT(u,s)\right|} \quad (3.29)$$

3.5.5 Synchro-squeezing Wavelet Transform

The synchro-squeezing wavelet transform (SSWT) was defined to incorporate the wavelet transform and auditory nerve-models into a tool that could be used for speech processing tasks [114].

The main objective of Synchro-squeezzing is to “sharpen" a time-frequency representation $TF(t,f)$ by “re-allocating" the value of the representation in the point (t,f) into a different point (t',f') according to the local behaviour of $TF(t,f)$ [115]. For the case of the conventional $WT(u,s)$, the transform will be spread over a region around the harmonic components in the time-scale plane [115]. In the SSWT, the aim is to re-allocate the $WT(u,s)$ to get a concentrated time-frequency representation of the signal, from which instantaneous frequency lines can be extracted. The $SSWT(u,f)$ is estimated based on the representation of the $WT(u,s)$ using Equation 3.30 [114]. The MATLAB implementation of the algorithm for Shynchro-squeezzing is freely available in [116].

$$SSWT(u,f_i) = (\Delta f)^{-1} \sum_{s_k:|f(s_k,u)-f_i|\leq \Delta f/2} WT(u,s_k)s_k^{-3/2}(\Delta s)_k \tag{3.30}$$

As illustration, Figure 3.11 shows the difference between the conventional WT and SSWT using a Morlet wavelet mother for the signal:

$$x(t) = cos[2\pi(0.1t^{2.6} + 3sin(2t) + 10t)] + e^{-0.2t}cos[2\pi(40 + t^{1.3})t]$$

Figure 3.12 shows the same difference for a voiced segment from a speech signal. Note in both figures that the instantaneous frequencies are better determined in the SSWT representation.

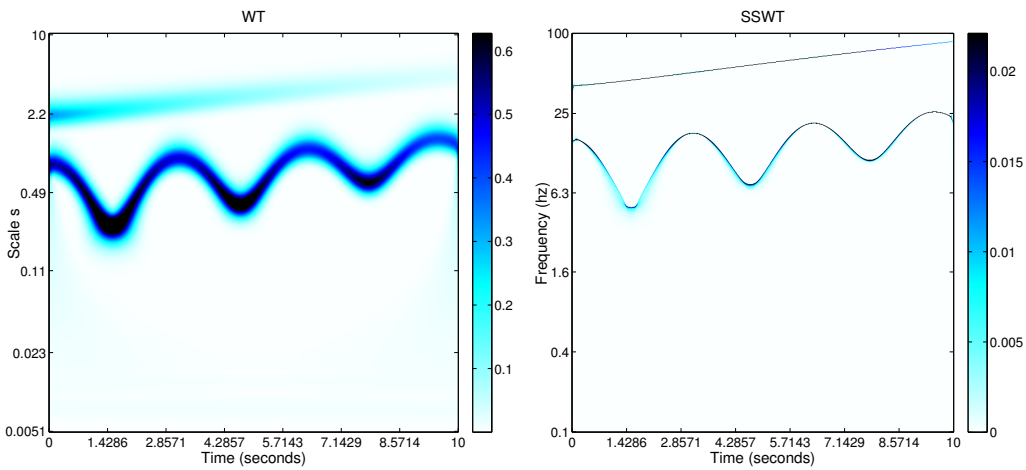

Fig. 3.11 Wavelet transform and Synchro-squeezzing wavelet transform for synthetic signal

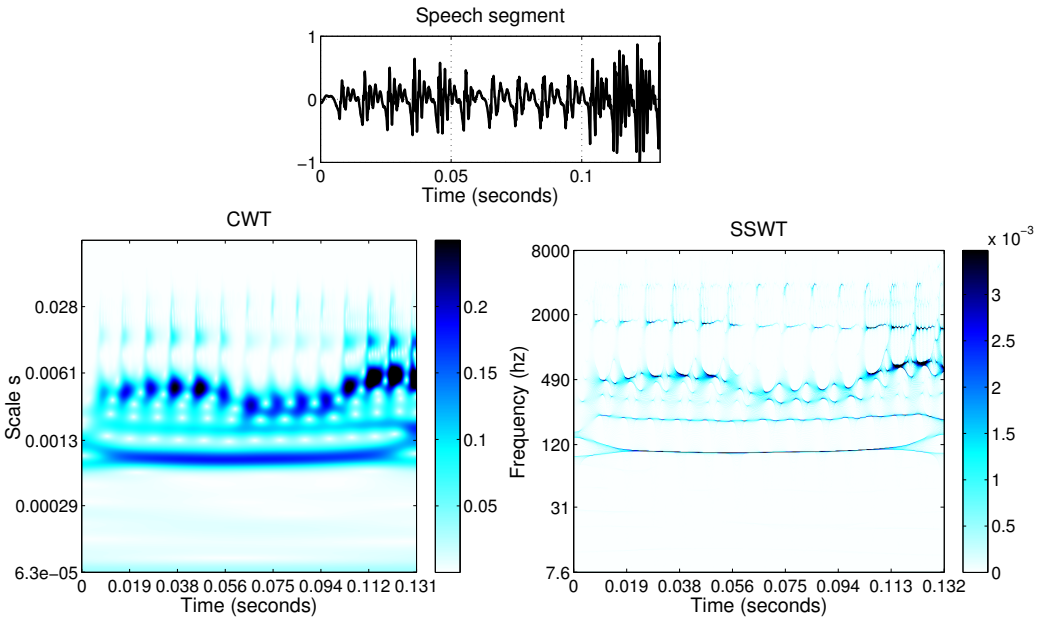

Fig. 3.12 Wavelet transform and Synchro-squeezzing wavelet transform for a speech segment .

Besides the re-allocation of SSWT, other advantage of this transform is its robustness to white noise and other perturbations [116], which made it able to analyze speech signal recorded in non-controlled conditions.

3.5.6 Features estimated from Wavelet Transform

Different set of features are estimated using each transformation described in the previous subsections. For the case of decompositions such as DWT, WPT, and WPP, different features are estimated in each one of the decomposition signals, including:

- MFCC
- log-Energy
- LPC
- NLD measures
- Statistical functions such as mean, standard deviation, skewness, and kurtosis.

The estimation of features for CWT, BWT, and SSWT consists in divide the time-frequency representation in 22 frequency regions separated according to the Bark scale.

Each region corresponds to sub-band frequencies from 0 to 8 kHz. For each sub-band the energy content is calculated using Equation 3.31. Where u_k is the sample index and f_i is the frequency index of the time frequency representation. f_i is calculated with the Bark scale using Equation 3.14.

$$E[i] = log\left|\frac{1}{N}\sum_{f_i}\sum_{u_k}^{N}|SSWT(u_k, f_i)|^2\right| \tag{3.31}$$

For the feature estimation, the speech segments are re-sampled to 16 kHz in order to avoid results dependent from the sampling frequency from the databases. Then, the representations are calculated for frames of 40 ms of length with time shift of 20 ms, and the energy content is estimated. Figure 3.13 summarizes the process.

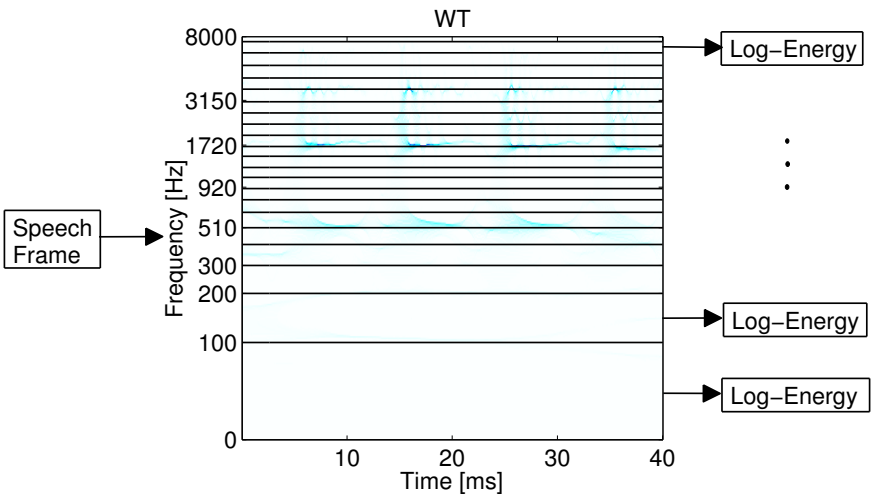

Fig. 3.13 Feature estimation process for CWT, BWT, and SSWT

3.6 Summary

This chapter described the main concepts and features proposed and used in this study for the recognition of emotions from speech. The features are divided into four categories: (1) Acoustic features, which include features derived from prosody, perturbation, spectral, cepstral, and noise measures. (2) Non-linear dynamics features which involves measures as the correlation dimension, the largest Lyapunov exponent, the Hurst exponent, the Lempel Ziv complexity, and entropy measures. (3) Features derived from the parametric non-stationary analysis using the time dependent ARMA models. Finally (4), features computed from the Wavelet transform including decompositions from the wavelet packet transform, wavelet

perceptual packet, time-frequency representations such as the continuous wavelet transform, the bionic wavelet transform, and the synchro-squeezing wavelet transform.

The next chapter will describe the concepts related to the noisy and telephony conditions that can corrupt the quality of the speech signals and the performance of any computational system for the speech analysis.

Chapter 4

Non-Controlled Noise Conditions

For real world applications, the effect produced by the telephone channels when the speech signal is transmitted and the background noise must be considered. These effects decrease the quality of the speech signal and the recognition capability of the models. There exist several studies that consider these scenarios by adding noise to the recordings or by simulating telephone conditions [69].

In this work, different noisy environments are considered for the analysis such as the street noise and the cafeteria babble. The evaluated conditions do not consider the noise influences on the speaking style such as Lombard effect, but it already forms a reasonable basis for the analysis of speech signal in non-controlled noise conditions and covers scenarios as microphone mismatch, cellular/phone channels and voice coding effects.

4.1 Noise Addition

It is important to consider the effect of background noise for recognition of emotions from speech to develop suitable solutions for real world applications. For example, for the case of monitoring threatening calls when the caller is in a public phone on the street.

The noisy environments considered in this work include the addition of street noise and cafeteria babble to the original noise-free speech recordings. The noisy signals are real and were captured with an omnidirectional microphone and a professional audio card. The change of the power spectral density (PSD) with frequency for each kind of noise along with additive white Gaussian noise (AWGN) is shown in Figure 4.1. Note that AWGN does not change with the frequency, while cafeteria and street decrease after 1 kHz. Note also that the street noise exhibits more power in the low frequency zone (under 200 Hz) while the cafeteria noise exhibits higher power values between 400 Hz and 1 kHz.

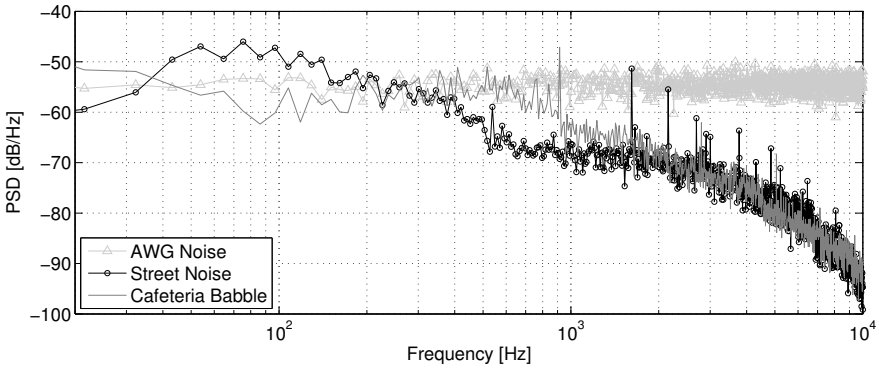

Fig. 4.1 PSD of Cafeteria, Street,and AWG noises

4.1.1 Cafeteria Babble

This kind of noise is featured by the presence of external voices produced by other speakers, which produce interference with the speech signal to evaluate. The noise contains also frequency components widely distributed in all the spectrum and exhibits a highly time-variability. The Cafeteria babble may contain also several impulsive noises produced among others by clink of dishes, cough, or laugh.

4.1.2 Street Noise

This kind of noise is produced in the principal or centric avenues of a city. The noise is produced by cars and other distant objects, which cause that the street noise contains frequency components specially distributed in the low frequency zone, although the noise produced by nearby cars may produce components in high frequency. The street noise may content impulsive noise produced by the horn of the cars.

4.2 Telephony Codecs and Channels

Besides the effect of the background noise, the effect of different telephony codecs and channels must be evaluated due to the speech recordings may be recorded by different sources and with different conditions in terms of the sampling frequency and the number of quantization bits. For example the recordings from an emergency call-center, from a mobile phone, or from a video-conference using Skype or Google Hangouts.

In this study different codecs are considered including mobile codecs derived from the adaptive multi-rate (AMR), codecs for voice over IP channels such as g.722, and codecs used for transmission on Internet such as SILK and Opus.

4.2.1 Adaptive Multi-Rate Narrowband (AMR-NB)

This codec is defined by the European telecommunications standards institute (ETSI) and the 3rd generation partnership project (3GPP) in [117]. This codec is widely used for the global system for mobile communications (GSM) and for the universal mobile telecommunications system (UMTS) phone networks. The codec uses a multi-rate algebraic code excited linear prediction (MR-CELP) scheme with a range of bit-rates from 4.75 kbps to 12.20 kbps. The different bit-rates are achieved by changing the number of samples encoded and the number of bits used to encode each sample. The codec includes a voice activity detector, a comfort noise generator, and an error concealment mechanism. The version used is the VisualOn implementation for the encoder and the OpenCORE implementation for the decoder, both found in the Libav open source library [118].

4.2.2 Adaptive Multi-Rate Wideband (AMR-WB)

This codec is defined also by ETSI and the 3GPP standard in [119]. The main difference between AMR-NB and AMR-WB consists in the bit-rates used for the transmission. For this case the bit rate ranges from 6.60 kbps to 23.85 kbps.

4.2.3 GSM Full Rate

A model of a mobile telephone based on full rate GSM 06.10 standard is considered. This model is widely used for compression of speech signals in mobile communications. The codec decreases the transmission bit-rate to 12.2 Kbps, in a bandwidth of 4 kHz according to the GSM standard [120].

4.2.4 G.722

This codec is defined by the international telecommunications union (ITU) in the recommendation G.722 [121]. This codec was developed with the aim of encode wide-band audio signals within 64 kbps to improve the speech quality. The coding scheme uses a sub-band adaptive pulse code modulation (SB-ADPCM). For this case the wide-band signal is split by filters in a lower sub-band (from 0 to 4 kHz) and a higher sub-band (from 4 kHz to 8 kHz). Each one is quantized independently. This codec is commonly used for VoIP applications where high bandwidth is available such as local area networks.

4.2.5 G.726

This codec is defined also by the (ITU) in the recommendation G.726 [122] and it is primarily intended to be used for international trunks. The coding scheme take as input a signal with a bit-rate of 64 kbps narrow-band and converts it to one of four different bit-rates (40, 32, 24 or 16 kbps) using the ADPCM scheme.

4.2.6 SILK

This is the codec used in the popular VoIP software Skype®. SILK can encode either narrow, medium, wide, and super wide bands speech signals with bit-rates in a range from 6 to 40 kbps. The coding scheme includes among others voice activity detection, pitch analysis, linear prediction analysis and noise shaping analysis blocks. The codec implements packet loss concealment and discontinuous transmission mechanisms. The code released to the public in [123] was compiled and used in this work.

4.2.7 Opus

This codec is defined by the Internet engineering task force (IETF) in [124]. Opus is based on the SILK codec and also supports variable bit-rates in a range from 8 kbps to 40 Kbps, but can extend beyond to improve the quality. This codec supports variable sampling rates. This work considers an average bit-rate of 64 Kbps and the default settings of the codec. Such bit-rate may be used in applications with high-speed Internet connection. Opus codec has been implemented in the WebRTC framework which is becoming a standard for Internet based multimedia communications [125].

4.3 Summary

Different kinds of background noise and telephony codecs were described in this chapter. The main aim of consider the background noise and codecs is to decrease the quality of the speech utterances and evaluate the recognition capability of the different algorithms under these non-controlled noise conditions in order to develop real world applications. Street and Cafeteria noises were considered, and different codecs used for mobile, VoIP, and Internet communications were evaluated.

The next chapter describes the methodology followed in this study to evaluate the proposed approaches. It include the description of the experiments, the feature sets used, and the classification and cross-validation schemes.

Chapter 5

Methodology

5.1 Experiments

Four different experiments were performed using the recordings from each database. The experiment number one considers the discrimination between high and low arousal emotions. The experiment number two is the detection of positive and negative valence emotions. The experiment three is the recognition of the fear-type emotions in Berlin, enterface05, and SAVEE databases. The fear-type emotions consider anger, anxiety, and disgust. In Berlin and SAVEE are also included the neutral state. Finally, the experiment number four is the multi-class recognition considering all emotions from the databases: seven emotions in Berlin, six emotions in enterface05, five emotions in FAU-Aibo, four emotions in IEMOCAP, and seven emotions in SAVEE. Table 5.1 lists the emotions considered in the four experiments addressed in this study.

5.2 Feature Sets

This section describes the feature sets computed using the features derived from acoustics, non-linear and wavelet based measures.

5.2.1 Acoustic Feature Sets

Three feature sets are derived from the acoustic analysis. (1) The features calculated using the OpenEAR toolkit, (2) the set formed by prosody measures derived from duration of voiced and unvoiced frames, the F_0, and the energy content, and (3) the feature set formed by spectral and noise measures.

Table 5.1 Four experiments

Database	2-class Arousal	2-class Valence	multi-class Fear-type	multi-class All emotions
Berlin	*High*: Fear, Disgust, Happiness, Anger. *Low*: Boredom, Neutral, Sadness.	*Positive*: Neutral Happiness. *Negative*: Boredom, Anger Sadness, Fear, Disgust.	Anger, Disgust Fear, Neutral	Fear, Disgust Happiness, Neutral Boredom, Sadness Anger
Enterface05	*High*: Fear, Disgust, Happiness, Anger. Surprise *Low*: Sadness.	*Positive*: Surprise Happiness. *Negative*: Anger Sadness, Fear, Disgust.	Anger, Disgust Fear	Fear, Disgust Happiness, Anger Surprise, Sadness
FAU-Aibo		Negative vs Idle		Anger, Emphatic Neutral, Positive Rest.
IEMOCAP	*High*: Fear, Disgust, Happiness, Anger. Surprise, Excitation, Frustration *Low*: Sadness, Neutral.	*Positive*: Surprise, Neutral Happiness, Excitation *Negative*: Anger, Frustration, Sadness, Fear, Disgust.		Anger, Sadness Happiness, Anger
SAVEE	*High*: Fear, Disgust, Happiness, Anger, Surprise. *Low*: Neutral, Sadness.	*Positive*: Neutral Happiness, Surprise. *Negative*: Anger, Sadness Fear, Disgust.	Anger, Disgust Fear, Neutral	Fear, Disgust Happiness, Neutral Surprise, Sadness Anger

OpenEAR

The feature set is formed by the 384 measures used as baseline in the "2009 INTERSPEECH emotional challenge" [34]. The feature set is formed by 16 descriptors and their derivatives. 12 statistical functions are calculated for each descriptor. The descriptors include the zero crossing rate (ZCR), the root mean square (RMS) energy, the F_0, the HNR, and 12 MFCC.

The 12 statistics functions include the mean, standard deviation, kurtosis, skewness, minimum value, maximum value, relative position of minimum, relative position of maximum, range, the slope of a linear regression, the offset of a linear regression, and the mean square error (MSE) of the regression. Finally the feature set is formed by $16 \times 2 \times 12 = 384$ features per utterance. Table 5.2 summarizes the features calculated.

Prosody

The feature set is formed by 38 measures related to duration, F_0, and energy. The duration features include 6 relationship measures between the duration of voiced, unvoiced, and silence segments. For the contours of the F_0, the energy and their derivatives, 8 functions are calculated: the mean, maximum value, minimum value, range, standard deviation, skewness,

Descriptors (16×2)	**statistic functions** (12)
ZCR	mean
RMS Energy	standard deviation
F_0	kurtosis, skewness
HNR	max value, min value, relative position, range
MFCC 1-12	slope, offset, MSE linear regression
Δs	

Table 5.2 OpenEAR features

kurtosis, and median. Thus the feature vector is formed by $6+8\times 4=38$ measures per utterance. Table 5.3 summarizes the computed features.

Descriptors	**statistic functions**
Duration	$sil/(v+u)$, v/u, $u/(v+u)$, $v/(v+u)$, v/sil, u/sil
F_0, ΔF_0	mean, max value, min value, range, std, skewness, kurtosis, median
Energy, ΔEnergy	mean, max value, min value, range, std, skewness, kurtosis, median

Table 5.3 Prosody features

Spectral and Noise

Different feature sets are formed for voiced and unvoiced segments. For the case of voiced segments 12 MFCC, HNR, GNE, and NNE are calculated for windows of 40 ms forming a 15-dimensional feature vector per window. For the case of unvoiced segments, the feature set is formed by 12 MFCC and the energy content in 25 frequency bands separated according to Bark scale, forming a 37-dimensional feature vector per window.

5.2.2 Non-Linear Dynamics

The NLD features are calculated only for voiced segments. The feature set is formed by a 4-dimensional feature vector that includes the CD, LLE, HE, and LZC. Additionally, another feature vector is formed by the merge of the NLD measures with the spectral and noise measures.

5.2.3 TARMA Models

Two different set of features are estimated from SP-TARMA models to characterize the non-stationary behavior of unvoiced segments. The first one considers statistic functions calculated on the model coefficients $\mathbf{a}[n]$, and $\mathbf{c}[n]$. The orders of $\mathbf{a}[n]$, and $\mathbf{c}[n]$ must be

estimated to construct the model. In this work, the Bayesian information criterion is used to select $\mathbf{n_a}$ and $mathbfn_c$, and the minimum residual sum squares (RSS) is used to select the parameter $mathbfk$. Figure 5.1 shows the order estimation. The chosen values are $\mathbf{n_a} = \mathbf{5}$, $\mathbf{n_c} = \mathbf{3}$, and $\mathbf{k} = \mathbf{1}$. After estimate the order of the model, the feature set is formed by 7 functions calculated on the coefficients: mean, standard deviation, kurtosis, skewness, max value, min value, and log-energy, forming a 56-feature vector per each unvoiced segment ($7 \times n_a + 7 \times 3 = 56$).

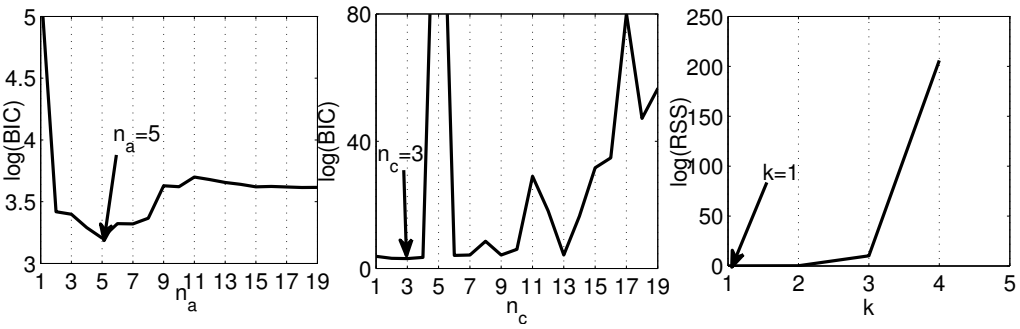

Fig. 5.1 Estimation of the order of SP-TARMA models

The second set of features is formed by 12 MFCCs calculated from the signals that are predicted by the SP-TARMA model. The MFCCs estimated from the model predictions may have reduced noise content, compared to the estimates obtained from the raw signal.

5.2.4 Wavelet Packet Transform and Multi-Resolution Analysis

Four different approaches were performed using WPT to find a suitable feature extraction scheme. The aproach number one considers the wavelet decomposition in different resolution levels from the first to the seventh, and their combinations. Superior levels were not considered due to the high number of features that would be introduced to the feature set and to the results obtained until such level. For each level, three frequency bands were considered: the low frequency decomposition signals formed by the first half of the nodes, the high frequency decomposition signals, and the combination of all nodes of the level. Figure 5.2 illustrates the division of the nodes for this experiment. The log-energy of each decomposition is computed and the Daubechies3 wavelet function is used.

The approach number two considers different wavelet decomposition using the Daubechies3 wavelet function to characterize both voiced and unvoiced segments. The decompositions were selected according to a forward selection criterion, where a different node was sequentially added and evaluated in a classification stage, starting from the lower frequency node in level one ($W_{1,0}$), to the last node of the level seven ($W_{7,127}$). Figure 5.3 shows the

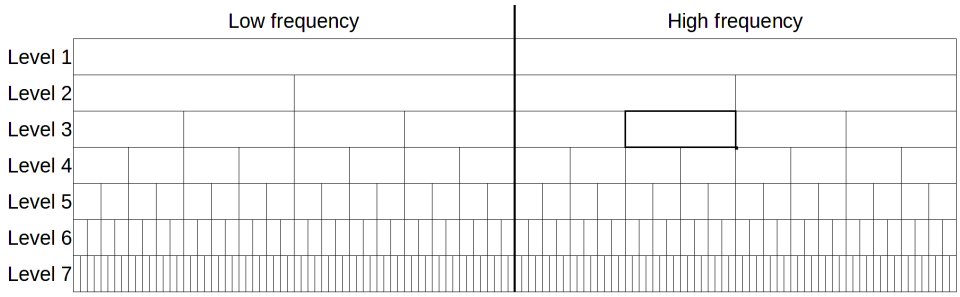

Fig. 5.2 WPT in different decomposition levels

decomposition considered both for voiced (up) and unvoiced segments (bottom), respectively. The chosen packets are focused on the low frequency components of the spectrum of the speech, which contain suitable information about the emotional content [80, 126].

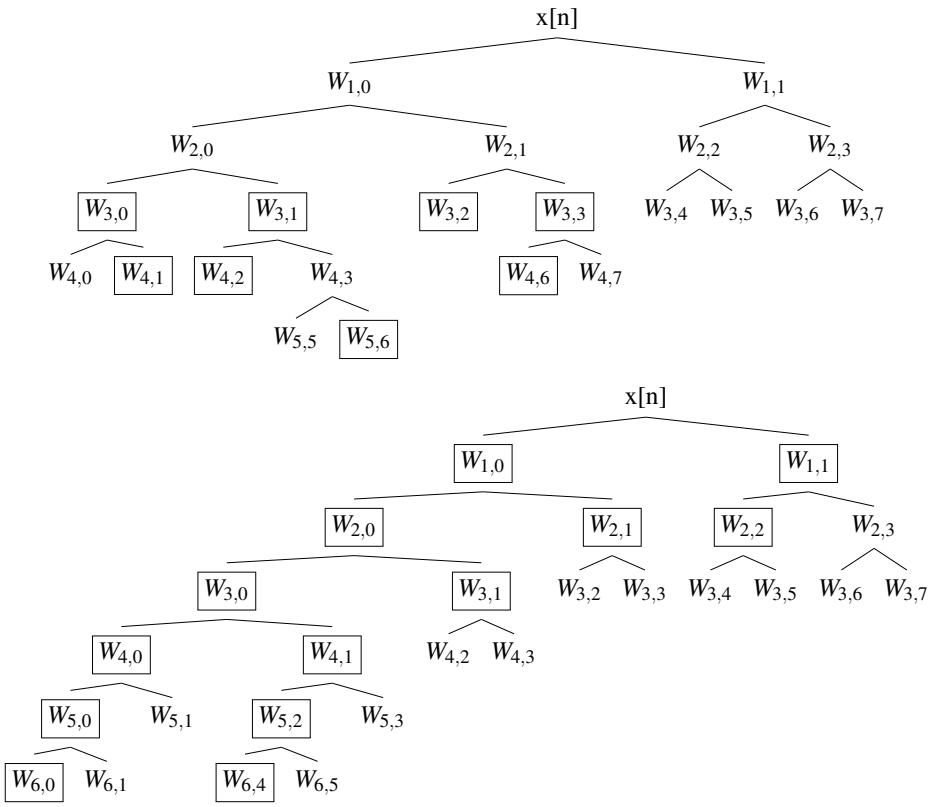

Fig. 5.3 WPT used for voiced segments (up) and unvoiced segments (bottom). $W_{x,y}$ indicates the wavelet decomposition in level x, in node y

The approach number three considers the multi-resolution decomposition according to the wavelet perceptual packets (WPP) which map the Bark scale to the wavelet domain. The decomposition is formed by 17 nodes from levels 3, 4, and 5.

Finally, the approach number four considers different measures calculated on the wavelet decompositions from Figure 5.3. The measures include the log-energy, MFCC, statistic functions, and NLD measures.

5.2.5 Wavelet Based Time-Frequency Representations

For the case of the time-frequency representations derived from the wavelet transform that includes the CWT, the BWT, and the SSWT, the feature vector is formed by 22 measures related to the energy content in 22 sub-bands separated according to the Bark scale in a range from 0 to 16 kHz.

5.3 Classification and Validation Methods

All the feature sets extracted from the speech signals are modeled by GMM supervectors, which are derived from a GMM adapted from a universal background model (UBM). The strategy is based on the combination of SVMs and GMMs. The theoretical background is explained in Appendix B.

The methodology is formed by a two-stage strategy. (1) The features estimated from voiced and unvoiced frames are transformed into a GMM supervector representation. The supervectors derived from voiced segments are classified separately of those estimated from unvoiced segments using a SVM with a radial base function (Gaussian) kernel. (2) A fusion scheme based on the scores obtained from the first classification scheme are used as new features to train a second classification stage using another SVM with a Gaussian kernel. Figure 5.4 illustrates the proposed classification scheme.

The metric used for the evaluation of the methodology is the unweighted average recall (UAR) instead of the weighted average recall (WAR). WAR is related to the global accuracy and it is preferred for the cases when the distribution of the classes in the databases is balanced. For the cases when the distribution of classes are high unbalanced the UAR measure is used, which can be defined as the average of the ratio of the true positives per class.

In all databases, a speaker independent cross-validation strategy based on LOSO is followed: in Berlin and IEMOCAP databases the data is divided into ten groups according to the ten speakers in each database. The utterances of a single speaker are used for test set and the utterances of the other nine speakers are used to train the classifier. The procedure is repeated for all the speakers from the database. In SAVEE it is performed the same cross-validation strategy with the four speakers from the database. In enterface05 the 44

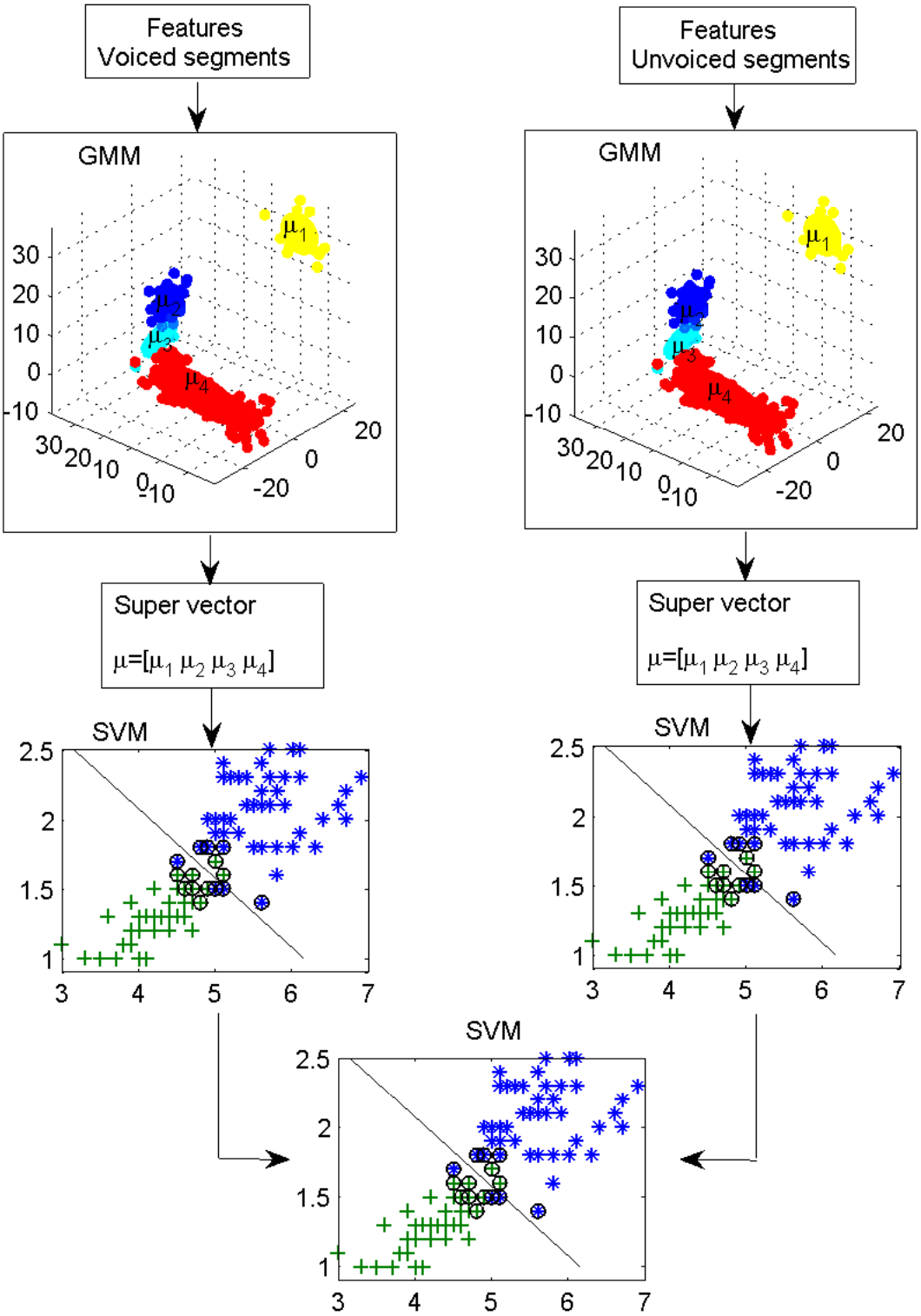

Fig. 5.4 Classification scheme

speakers are separated into 11 groups formed by 4 different speakers, and the validation is performed with those 11 groups. Finally, in FAU-Aibo the same strategy followed in the "2009 INTERSPEECH emotional challenge" is followed, where the recordings of the children from one of the participating schools are used for train, and the recordings from the other school are used for test.

5.4 Non-Controlled Noise Conditions

5.4.1 Additive Noise

This scenario is tested with two kinds of additive environmental noise: street noise and cafeteria babble. The evaluation of the noise includes SNR levels of 0, 3, and 6 dB, as it was published in [80]. The effect of two different speech enhancement techniques is also evaluated. The first one is the method proposed in [127]. This method will be refereed to logMMSE. The second one method is based on the subspace approach proposed in [128] based on the Karnuhen-Loeve Transform (KLT) decomposition. This method will be refereed to as KLT. For a detailed explanation of the speech enhancement techniques follow the appendix A.

5.4.2 Natural Environment Noise

With the aim of characterize the emotions for utterances corrupted by noise added in a non-artificial way, the recordings of the databases were re-captured in presence of two different noisy environments: street and office. For the re-capturing process we consider a high quality studio monitor B2030A from Behringer to reproduce the recordings and a professional microphone shure SM63 with a professional audio card M-Audio fast track C400. The specifications of the devices are shown in Table 5.4 and the scheme of the experiment is illustrated in Figure 5.5.

Device	Reference
Studio monitor	B2030A [129]
Microphone	Shure SM63 [130]
Audio card	Fast track C400 [131]

Table 5.4 Devices for re-capturing the databases in noisy environments

Fig. 5.5 Scheme for re-capture the databases in noisy conditions

5.4.3 Audio Codecs

The codecs used in this study compress the speech signals to reduce the bit-rate and thus to make a more efficient use of the network resources such as the bandwidth. A total of seven audio codecs were considered: the GSM, AMR-NB, and AMR-WB used for mobile networks. The G.722 and G.726 used for VoIP networks. The SILK codec used by Skype, and the Opus codec used in the WebRTC framework. The description of such codecs was explained with details in Section 4.2.

The speech signals from the databases were encoded and decoded using each codec. The encoder and decoder for AMR-NB, GSM, AMR-WB, G.722 and G.726 codecs were implemented with the Libav library. For the SILK codec, the code released to the public in [132] was compiled for the implementation, and for the case of the Opus encoder the version implemented in the package *opus-tools* found in the Debian repositories was used.

The settings of the encoders were left with their default values, except for the case of the bit-rate for the AMR-NB, and AMR-WB codecs, which was set to different fixed bit-rates. Both Opus and SILK codecs use Variable Bit-Rates (VBR).

5.5 Summary

The methodology followed in this study was described in this chapter. The four experiments developed in this work were explained: (1) the high-low arousal detection, (2) the positive-negative valence detection, (3) the classification of the fear-type emotions, and (4) the classification of the multiple emotions from the data-sets. The feature sets proposed in this study based on acoustics, non-linear dynamics, and wavelet based measures are also

explained, followed by the classification and cross-validation strategies. Finally the non-controlled noise conditions considering the background noise and the audio codecs are explained. The next chapter shows and discuss the main results obtained in this work considering each features set and the non-controlled noise conditions.

Chapter 6

Results and Discussion

6.1 Classification of Noise-Free Speech Signals

The next subsections include the main results obtained for each experiment described in Section 5.1: the high vs low arousal detection, the positive vs negative valence, the fear-type emotions, and all emotions.

6.1.1 Results Experiment 1: High vs Low Arousal Emotions

Table 6.1 shows the results obtained with all the feature sets related to the acoustics, NLD, and wavelet based measures both for voiced, unvoiced, and the fusion of the scores for the detection of high-low arousal emotions in Berlin, enterface05, IEMOCAP and SAVEE databases.

In Berlin database the highest result is obtained with the 384 features from OpenEAR followed by the obtained with the spectral+noise+NLD. For the case of SAVEE database the highest result is obtained with the feature sets derived from the WPT analysis for voiced segments, followed by the prosody measures. In enterface05 the highest result correspond to the time frequency representations based on the wavelet transform e.g., CWT, BWT, and SSWT followed by openEAR. Finally in IEMOCAP the highest result is obtained also with the features based on the wavelet based time frequency representations.

In general, higher results are obtained with features extracted from voiced segments rather than unvoiced, but there are some cases where the features extracted from unvoiced segments exceed those obtained with voiced as for the case of spectral+noise features in SAVEE and IEMOCAP. The fusion scheme is useful to improve the results in features extracted from the wavelet measures specially in SAVEE and IEMOCAP.

Table 6.1 Results for classification of high vs low arousal emotions

Feature set	Segments	# Feat.	Berlin	SAVEE	enterface-05	IEMOCAP
OpenEAR	all signal	384	$\boxed{97.3 \pm 3.0}$	83.3 ± 8.8	81.0 ± 2.0	75.5 ± 3.8
Prosody	all signal	38	91.8 ± 4.3	86.5 ± 5.1	76.3 ± 5.3	72.0 ± 4.2
Spectral+noise	Voiced	15 per window	94.9 ± 7.1	74.8 ± 8.7	80.2 ± 1.9	72.2 ± 4.9
	Unvoiced	37 per window	82.2 ± 7.5	80.8 ± 7.4	78.4 ± 1.1	75.1 ± 3.2
	Fusion		91.5 ± 9.8	81.3 ± 4.2	80.1 ± 3.4	71.3 ± 5.1
NLD	Voiced	4 per window	94.5 ± 6.0	78.9 ± 3.2	77.4 ± 1.1	71.0 ± 6.2
Spectral+noise+NLD	Voiced	19 per window	96.9 ± 4.4	77.7 ± 10.1	80.2 ± 1.6	72.3 ± 6.4
	Unvoiced	37 per window	82.2 ± 7.5	80.6 ± 7.3	77.9 ± 1.5	75.1 ± 3.2
	Fusion		92.7 ± 6.0	82.9 ± 6.3	79.7 ± 1.9	72.0 ± 4.1
SP-TARMA stat.	Unvoiced	56 per segment	85.5 ± 6.4	70.5 ± 3.2	78.9 ± 0.8	64.0 ± 3.1
SP-TARMA MFCC	Unvoiced	12 per window	90.5 ± 7.6	82.3 ± 4.1	78.5 ± 1.3	70.8 ± 4.7
WPP energy	Voiced	17 per window	93.1 ± 5.1	81.0 ± 10.7	80.9 ± 2.5	69.5 ± 5.5
	Unvoiced	17 per window	82.1 ± 5.5	74.0 ± 7.9	80.7 ± 1.0	73.5 ± 3.4
	Fusion		91.4 ± 7.4	83.3 ± 8.9	80.9 ± 2.5	74.1 ± 3.2
WPT energy	Voiced	8 per window	92.7 ± 4.2	85.4 ± 6.1	80.7 ± 5.0	73.5 ± 4.6
	Unvoiced	13 per window	80.6 ± 6.5	71.3 ± 7.2	79.0 ± 1.7	70.0 ± 6.7
	Fusion		90.9 ± 5.5	83.3 ± 5.9	80.1 ± 2.1	72.3 ± 4.1
WPT MFCC	Voiced	96 per window	94.6 ± 5.0	88.5 ± 6.4	79.6 ± 1.3	70.8 ± 4.1
	Unvoiced	156 per window	79.9 ± 5.0	79.6 ± 9.4	78.2 ± 1.0	71.4 ± 5.0
	Fusion		93.5 ± 6.0	82.9 ± 6.8	78.9 ± 2.0	72.1 ± 4.0
WPT stat.	Voiced	24 per window	93.2 ± 5.2	88.9 ± 5.7	79.8 ± 1.0	73.6 ± 5.4
	Unvoiced	36 per window	79.7 ± 4.3	74.0 ± 5.0	77.6 ± 0.6	69.4 ± 5.8
	Fusion		89.1 ± 6.6	85.6 ± 4.9	78.3 ± 2.6	74.1 ± 4.1
WPT energy+MFCC+stat	Voiced	128 per window	95.7 ± 4.0	$\boxed{89.0 \pm 5.7}$	80.7 ± 5.0	74.6 ± 4.0
	Unvoiced	208 per window	82.2 ± 5.7	82.3 ± 9.9	78.3 ± 1.2	73.0 ± 6.2
	Fusion		93.2 ± 5.0	87.3 ± 3.9	78.9 ± 2.0	75.4 ± 3.2
WPT NLD	Voiced	32 per window	94.4 ± 5.0	85.8 ± 2.4	78.9 ± 1.2	69.5 ± 3.2
	Unvoiced	26 per window	82.5 ± 5.8	71.0 ± 5.1	78.5 ± 0.9	70.2 ± 3.4
	Fusion		93.0 ± 4.6	85.6 ± 4.5	78.6 ± 2.2	71.3 ± 3.2
CWT Energy	Voiced	22 per window	95.7 ± 5.6	82.5 ± 9.1	81.2 ± 2.2	74.4 ± 3.8
	Unvoiced	22 per window	89.1 ± 8.7	79.8 ± 8.1	79.6 ± 1.2	75.1 ± 2.7
	Fusion		93.3 ± 8.3	87.3 ± 7.4	80.8 ± 2.5	76.4 ± 3.9
BWT Energy	Voiced	22 per window	95.6 ± 5.5	82.3 ± 8.0	81.5 ± 1.7	74.3 ± 4.2
	Unvoiced	22 per window	89.6 ± 8.5	80.4 ± 7.2	79.8 ± 1.5	74.8 ± 2.8
	Fusion		94.0 ± 6.6	84.6 ± 7.1	$\boxed{81.9 \pm 2.2}$	76.1 ± 4.0
SSWT Energy	Voiced	22 per window	95.8 ± 5.5	84.4 ± 8.3	81.1 ± 1.7	75.7 ± 4.7
	Unvoiced	22 per window	89.2 ± 8.4	79.5 ± 6.7	80.4 ± 1.4	75.6 ± 2.9
	Fusion		95.0 ± 5.5	81.8 ± 5.7	80.2 ± 2.9	$\boxed{77.2 \pm 3.6}$

Note that the addition of NLD measures to the spectral+noise feature set improves the results in Berlin and SAVEE, which may indicates that the NLD can be considered complementary measures to the acoustic analysis.

6.1.2 Results Experiment 2: Positive vs Negative Valence Emotions

Table 6.2 shows the results obtained with all the feature sets for the discrimination between positive and negative emotions in Berlin SAVEE, enterface05, FAU-Aibo and IEMOCAP databases.

Table 6.2 Results for classification of positive vs negative valence emotions

Feature set	Segments	# Feat.	Berlin	SAVEE	enterface-05	FAU-Aibo	IEMOCAP
OpenEAR	all signal	384	87.2 ± 2.4	72.5 ± 5.7	81.4 ± 3.6	62.0	59.0 ± 3.2
Prosody	all signal	38	81.2 ± 6.2	67.7 ± 6.8	66.0 ± 5.5	62.6	57.5 ± 2.4
Spectral+noise	Voiced	15 per window	83.5 ± 4.3	64.8 ± 4.9	75.1 ± 2.5	68.6	55.9 ± 3.0
	Unvoiced	37 per window	73.7 ± 5.3	64.2 ± 1.6	71.4 ± 2.1	62.9	54.3 ± 3.0
	Fusion		80.2 ± 3.6	65.4 ± 2.7	72.4 ± 4.4	67.8	58.5 ± 3.6
NLD	Voiced	4 per window	79.5 ± 4.4	61.3 ± 2.5	70.7 ± 1.9	66.7	53.8 ± 4.2
Spectral+noise+NLD	Voiced	19 per window	82.9 ± 5.8	66.7 ± 4.0	74.9 ± 2.4	69.6	57.0 ± 3.1
	Unvoiced	37 per window	73.7 ± 5.3	62.9 ± 3.5	71.4 ± 2.1	62.9	54.3 ± 3.0
	Fusion		79.8 ± 5.8	67.3 ± 5.1	73.8 ± 5.1	68.8	59.5 ± 3.4
SP-TARMA stat.	Unvoiced	56 per segment	73.5 ± 6.4	60.1 ± 3.2	68.7 ± 1.4	56.4	58.5 ± 2.7
SP-TARMA MFCC	Unvoiced	12 per window	73.9 ± 6.1	62.4 ± 3.1	69.3 ± 1.4	61.6	57.0 ± 1.9
WPP energy	Voiced	17 per window	79.0 ± 6.8	64.2 ± 3.5	72.6 ± 2.8	68.5	54.0 ± 2.7
	Unvoiced	17 per window	76.5 ± 5.7	63.1 ± 4.6	73.0 ± 2.5	63.9	56.8 ± 2.7
	Fusion		78.7 ± 6.4	67.5 ± 4.4	71.5 ± 3.6	69.0	58.1 ± 3.2
WPT energy	Voiced	8 per window	78.7 ± 5.8	66.0 ± 4.0	72.0 ± 1.1	66.9	57.9 ± 2.2
	Unvoiced	13 per window	73.0 ± 6.4	61.5 ± 2.4	72.5 ± 2.8	64.7	53.5 ± 2.9
	Fusion		78.5 ± 4.8	62.7 ± 7.5	72.1 ± 3.4	67.5	58.5 ± 2.1
WPT MFCC	Voiced	96 per window	77.5 ± 5.5	70.0 ± 8.9	75.4 ± 4.0	65.2	58.5 ± 2.7
	Unvoiced	156 per window	73.9 ± 5.0	65.2 ± 3.5	70.6 ± 1.6	65.0	59.0 ± 3.6
	Fusion		75.6 ± 6.0	71.0 ± 8.5	71.8 ± 3.1	66.5	59.1 ± 2.0
WPT stat.	Voiced	24 per window	77.6 ± 4.1	67.1 ± 7.5	70.5 ± 2.2	69.8	57.3 ± 1.9
	Unvoiced	36 per window	74.6 ± 6.4	64.8 ± 7.0	69.2 ± 1.3	62.9	54.1 ± 3.0
	Fusion		77.7 ± 6.9	67.0 ± 5.0	68.5 ± 1.9	67.7	56.3 ± 2.5
WPT energy+MFCC+stat	Voiced	128 per window	81.2 ± 3.3	70.7 ± 9.8	75.9 ± 3.0	68.2	57.3 ± 2.5
	Unvoiced	208 per window	75.0 ± 5.3	64.6 ± 4.2	72.6 ± 1.6	65.0	56.1 ± 5.5
	Fusion		75.7 ± 4.5	70.0 ± 8.4	72.5 ± 3.6	67.7	59.1 ± 1.9
WPT NLD	Voiced	32 per window	79.6 ± 6.2	65.3 ± 6.2	72.7 ± 1.9	67.5	56.1 ± 2.1
	Unvoiced	26 per window	75.4 ± 5.9	61.5 ± 3.4	72.6 ± 3.0	62.2	53.1 ± 2.2
	Fusion		78.7 ± 5.1	64.8 ± 5.5	71.2 ± 2.8	68.0	57.3 ± 2.7
CWT Energy	Voiced	22 per window	80.0 ± 3.7	64.4 ± 5.0	74.6 ± 1.7	66.5	54.5 ± 3.8
	Unvoiced	22 per window	76.3 ± 5.4	63.8 ± 3.2	73.4 ± 2.6	56.7	57.5 ± 2.3
	Fusion		78.2 ± 4.2	66.7 ± 3.5	74.4 ± 2.0	67.5	58.4 ± 4.7
BWT Energy	Voiced	22 per window	80.0 ± 3.7	63.8 ± 6.3	74.2 ± 2.0	68.4	54.6 ± 3.6
	Unvoiced	22 per window	76.4 ± 6.7	63.8 ± 4.5	73.6 ± 2.7	61.8	57.6 ± 2.1
	Fusion		78.0 ± 5.5	64.6 ± 5.9	73.5 ± 4.2	68.7	58.1 ± 3.2
SSWT Energy	Voiced	22 per window	81.7 ± 4.6	64.2 ± 4.8	75.6 ± 2.9	70.3	56.2 ± 4.0
	Unvoiced	22 per window	76.9 ± 6.0	63.1 ± 3.4	74.3 ± 2.8	60.5	58.3 ± 1.9
	Fusion		78.5 ± 3.8	65.4 ± 5.3	73.8 ± 3.6	69.3	59.5 ± 3.3

In Berlin the highest result is obtained with the features from OpenEAR, as in the experiment for high vs low arousal classification. In this case there is a difference around 5% between the results from openEAR and the obtained with spectral+noise+NLD. For the case of SAVEE and enterface05 the highest result is obtained also with the features from OpenEAR followed by the obtained using the WPT energy+MFCC+stat. In FAU-Aibo the highest results are obtained with the features derived from the wavelet measures, the SSWT energy, and the WPT stat. Finally in IEMOCAP there is not differences between the results obtained with OpenEAR, spectral+noise+NLD and the wavelet based measures.

For the WPT based measures, the highest results are obtained when the energy, statistics functionals and MFCC are merged together and computed on the wavelet coefficients. For the time-frequency representations based on wavelets the highest results are obtained when the SSWT is used, but there is not a high difference relative to the other representations.

Finally, note that the results obtained for the classification of positive vs negative valence emotions are lower than those obtained for the detection of high vs low arousal. This fact can explained because most of the feature sets used are based on energy measures, which are widely useful to discriminate emotions in different arousal planes such as happiness vs sadness, but not to classify emotions in the same arousal plane and with different valence, such as anger vs happiness.

6.1.3 Results Experiment 3: Fear-type Emotions

Table 6.3 shows the results obtained for the classification of the fear-type emotions e.g., anger, disgust, and fear in Berlin, enterface05, and SAVEE. In Berlin and SAVEE the neutral state is also included.

For this experiment the highest result in Berlin is obtained with the features from OpenEAR, as in the previous experiments, followed by the features based on the SSWT and the spectral+noise+NLD. Note that the WPT energy+MFCC+stat provides the highest results in SAVEE, followed by the prosody measures, which also produce the lowest result in enterface05, where the highest result is obtained also with the 384 features from OpenEAR. In general the features extracted from voiced segments provides higher results than the obtained with the features extracted from unvoiced. There are some cases where the fusion highly improves the results relative to the separately classification of voiced and unvoiced segments, such as the spectral+noise, SSWT, CWT feature sets in SAVEE.

6.1.4 Results Experiment 4: Multiple Emotions

This experiment consists in the classification of multiple emotions from the databases: seven in Berlin, six in enterface05, seven in SAVEE, five in FAU-Aibo, and four in IEMOCAP. Table 6.4 shows the results.

As in the three previous experiments, the highest result in Berlin and enterface05 is obtained with the OpenEAR toolkit. For the case of Berlin, the highest results is followed by the results obtained with the SSWT energy feature set, with a difference around 10%, while for enterface05 the second highest result is obtained with the CWT. IN SAVEE the highest result is obtained with the WPT energy+MFCC+stat. In FAU-AIbo the highest results is 38.9% using the spectral+noise+NLD feature set. Finally in IEMOCAP the SSWT

Table 6.3 Results for classification of fear-type emotions

Feature set	Segments	# Feat.	Berlin	SAVEE	enterface-05
OpenEAR	all signal	384	91.4 ± 5.0	65.2 ± 18.1	78.2 ± 5.5
Prosody	all signal	38	75.6 ± 7.0	69.5 ± 15.7	52.7 ± 3.7
Spectral+noise	Voiced	15 per window	88.5 ± 9.0	54.5 ± 8.8	71.6 ± 4.9
	Unvoiced	37 per window	68.5 ± 8.6	54.5 ± 9.7	57.1 ± 6.2
	Fusion		85.6 ± 6.0	68.5 ± 10.8	67.0 ± 7.2
NLD	Voiced	4 per window	81.1 ± 10.5	62.2 ± 5.7	63.2 ± 4.1
Spectral+noise+NLD	Voiced	19 per window	88.3 ± 9.8	58.5 ± 14.3	70.2 ± 5.7
	Unvoiced	37 per window	68.5 ± 8.6	53.5 ± 8.4	57.1 ± 6.2
	Fusion		83.0 ± 9.7	64.5 ± 13.7	66.8 ± 5.7
SP-TARMA stat.	Unvoiced	56 per segment	66.5 ± 6.7	61.5 ± 5.1	54.0 ± 4.9
SP-TARMA MFCC	Unvoiced	12 per window	70.3 ± 8.3	64.3 ± 7.8	59.7 ± 5.0
WPP energy	Voiced	17 per window	81.7 ± 6.2	57.8 ± 12.7	65.0 ± 4.7
	Unvoiced	17 per window	59.5 ± 23.0	56.5 ± 3.8	66.4 ± 6.5
	Fusion		81.3 ± 5.8	56.1 ± 11.5	65.9 ± 6.3
WPT energy	Voiced	8 per window	77.8 ± 6.1	65.5 ± 9.7	66.7 ± 5.1
	Unvoiced	13 per window	56.5 ± 22.7	48.1 ± 3.5	61.8 ± 6.4
	Fusion		74.3 ± 9.7	64.8 ± 13.2	66.0 ± 5.7
WPT MFCC	Voiced	96 per window	80.8 ± 6.0	69.5 ± 11.3	65.9 ± 6.1
	Unvoiced	156 per window	64.9 ± 25.1	48.2 ± 3.5	64.9 ± 6.6
	Fusion		79.6 ± 4.0	69.5 ± 12.5	68.0 ± 7.3
WPT stat.	Voiced	24 per window	82.3 ± 7.6	65.8 ± 16.1	65.3 ± 4.9
	Unvoiced	36 per window	54.1 ± 21.5	51.5 ± 5.5	61.2 ± 4.3
	Fusion		77.9 ± 10.7	66.2 ± 14.8	67.3 ± 5.8
WPT energy+MFCC+stat	Voiced	128 per window	84.0 ± 5.7	70.8 ± 13.8	70.9 ± 4.6
	Unvoiced	208 per window	69.0 ± 26.6	59.5 ± 14.5	65.4 ± 4.4
	Fusion		82.8 ± 6.7	72.2 ± 12.3	71.0 ± 9.0
WPT NLD	Voiced	32 per window	81.7 ± 6.5	64.2 ± 10.2	65.8 ± 4.6
	Unvoiced	26 per window	57.2 ± 22.4	48.5 ± 2.0	64.7 ± 5.4
	Fusion		84.7 ± 8.1	64.5 ± 12.9	67.4 ± 6.3
CWT Energy	Voiced	22 per window	87.5 ± 8.5	58.5 ± 19.9	69.8 ± 4.4
	Unvoiced	22 per window	80.0 ± 6.0	56.5 ± 6.4	69.0 ± 4.3
	Fusion		83.9 ± 6.5	67.9 ± 11.4	71.7 ± 4.6
BWT Energy	Voiced	22 per window	86.1 ± 10.0	63.5 ± 20.6	70.3 ± 5.4
	Unvoiced	22 per window	78.6 ± 6.6	57.2 ± 6.0	67.6 ± 4.4
	Fusion		86.1 ± 5.7	68.2 ± 10.8	70.6 ± 7.2
SSWT Energy	Voiced	22 per window	88.3 ± 7.0	61.5 ± 13.4	70.1 ± 5.8
	Unvoiced	22 per window	79.7 ± 6.3	55.8 ± 7.2	69.3 ± 4.4
	Fusion		89.6 ± 6.0	68.9 ± 9.4	73.5 ± 6.4

energy features produce the highest result followed by the spectral+noise+stat and the WPT energy+MFCC+stat with a lower difference. Note also that the features extracted from the voiced segments produce the highest results and the fusion scheme is useful to improve the results in several cases, specially in IEMOCAP and database.

For the case of TARMA based features, the highest results are obtained by considering the MFCCs calculated from the model predictions instead of the measures calculated directly on the coefficients. These results exhibit an improvement to the study published in [133], where only the fear-type emotions were considered. The evaluation of TARMA models must be evaluated more deeply considering other time structures such as syllables and the

Table 6.4 Results for classification of multiple emotions

Feature set	Segments	# Feat.	Berlin	SAVEE	enterface-05	FAU-Aibo	IEMOCAP
OpenEAR	all signal	384	$\boxed{80.4 \pm 8.0}$	49.4 ± 17.6	$\boxed{63.2 \pm 6.7}$	32.5	57.2 ± 2.8
Prosody	all signal	38	64.8 ± 6.5	47.7 ± 11.8	31.9 ± 4.4	36.8	50.8 ± 5.0
Spectral+noise	Voiced	15 per window	68.9 ± 11.0	38.3 ± 10.8	49.9 ± 5.4	34.3	47.0 ± 5.6
	Unvoiced	37 per window	43.2 ± 6.4	35.4 ± 7.5	33.4 ± 3.8	35.8	52.5 ± 3.6
	Fusion		65.9 ± 11.0	44.2 ± 6.8	49.0 ± 5.3	34.8	56.4 ± 3.3
NLD	Voiced	4 per window	62.9 ± 7.5	40.6 ± 2.7	40.6 ± 5.0	31.8	48.1 ± 5.2
Spectral+noise+NLD	Voiced	19 per window	69.2 ± 10.2	41.7 ± 11.6	49.0 ± 3.9	$\boxed{38.9}$	49.6 ± 6.5
	Unvoiced	37 per window	43.2 ± 6.4	35.4 ± 7.5	33.8 ± 3.0	29.3	52.5 ± 3.6
	Fusion		62.6 ± 11.1	42.7 ± 8.5	47.8 ± 5.3	34.3	56.4 ± 2.9
SP-TARMA stat.	Unvoiced	56 per segment	46.1 ± 6.4	34.3 ± 4.4	33.4 ± 2.7	22.5	43.1 ± 3.0
SP-TARMA MFCC	Unvoiced	12 per window	47.3 ± 8.0	36.5 ± 5.1	38.0 ± 4.6	27.5	46.3 ± 5.0
WPP energy	Voiced	17 per window	62.0 ± 7.1	40.4 ± 8.8	43.9 ± 4.0	30.1	50.1 ± 3.2
	Unvoiced	17 per window	44.0 ± 17.4	34.0 ± 6.7	41.8 ± 2.8	29.0	48.8 ± 4.0
	Fusion		61.8 ± 6.9	39.2 ± 9.5	48.1 ± 5.5	32.5	52.2 ± 3.1
WPT energy	Voiced	8 per window	63.0 ± 7.2	44.8 ± 10.0	44.0 ± 4.0	30.5	50.3 ± 3.7
	Unvoiced	13 per window	41.0 ± 17.0	32.7 ± 5.0	38.4 ± 6.1	30.0	44.6 ± 4.5
	Fusion		59.2 ± 5.9	43.1 ± 8.5	48.1 ± 5.3	32.3	51.2 ± 2.4
WPT MFCC	Voiced	96 per window	63.1 ± 6.0	48.1 ± 11.5	44.0 ± 3.5	36.0	54.5 ± 2.2
	Unvoiced	156 per window	46.8 ± 18.5	42.3 ± 14.2	38.7 ± 4.3	34.8	53.6 ± 5.0
	Fusion		60.9 ± 8.5	48.5 ± 11.0	46.7 ± 4.8	35.4	55.5 ± 2.1
WPT stat.	Voiced	24 per window	61.5 ± 5.1	43.8 ± 13.7	43.0 ± 3.3	29.0	49.1 ± 5.4
	Unvoiced	36 per window	39.9 ± 16.2	30.6 ± 4.6	35.7 ± 4.4	33.0	45.4 ± 4.4
	Fusion		58.6 ± 6.2	48.8 ± 11.5	41.8 ± 4.4	32.4	50.1 ± 2.2
WPT energy+MFCC+stat	Voiced	128 per window	65.0 ± 3.7	50.2 ± 12.5	49.2 ± 3.0	$\boxed{38.0}$	56.1 ± 2.3
	Unvoiced	208 per window	48.8 ± 18.9	41.5 ± 11.9	38.9 ± 3.8	28.6	50.1 ± 9.3
	Fusion		66.1 ± 4.8	51.7 ± 14.0	48.7 ± 6.1	38.6	57.1 ± 4.3
WPT NLD	Voiced	32 per window	62.4 ± 7.0	45.8 ± 9.6	42.2 ± 3.2	30.2	50.1 ± 3.6
	Unvoiced	26 per window	40.9 ± 15.9	30.6 ± 4.6	40.8 ± 3.9	26.8	49.3 ± 5.7
	Fusion		64.0 ± 7.4	45.6 ± 9.3	44.3 ± 3.9	31.5	52.5 ± 4.5
CWT Energy	Voiced	22 per window	61.3 ± 8.3	40.6 ± 13.5	48.4 ± 4.7	35.5	46.7 ± 6.0
	Unvoiced	22 per window	54.7 ± 6.6	39.4 ± 5.8	45.7 ± 4.0	34.0	51.3 ± 3.6
	Fusion		66.6 ± 6.5	43.8 ± 9.0	51.3 ± 5.6	34.5	55.9 ± 5.0
BWT Energy	Voiced	22 per window	63.7 ± 9.1	41.2 ± 14.9	48.4 ± 4.4	32.1	46.6 ± 5.3
	Unvoiced	22 per window	55.5 ± 7.4	39.8 ± 4.3	44.9 ± 4.3	30.2	51.2 ± 3.9
	Fusion		66.5 ± 7.7	47.3 ± 10.3	49.7 ± 4.3	34.1	55.2 ± 5.7
SSWT Energy	Voiced	22 per window	64.0 ± 8.0	42.7 ± 11.1	48.0 ± 3.5	32.6	48.7 ± 5.0
	Unvoiced	22 per window	55.0 ± 8.2	39.6 ± 6.2	45.9 ± 3.6	21.8	52.0 ± 2.9
	Fusion		69.3 ± 7.6	45.4 ± 12.1	48.8 ± 5.8	31.4	$\boxed{58.2 \pm 4.1}$

transitions between voiced and unvoiced segments. Other measures from TARMA models must be also implemented and analyzed.

An additional experiment was performed in Berlin database for the classification of the seven emotions using features derived from the WPT in different decomposition levels from the first to the seventh considering three sets of decomposition packets: (1) the all packets from each level, which are related to the complete spectrum of the signal, (2) the lowest half packets, which represent the low frequency components of the spectrum, and (3) the highest frequency packets. Figure 6.1 shows the results of this analysis both for voiced (up), and unvoiced segments (bottom).

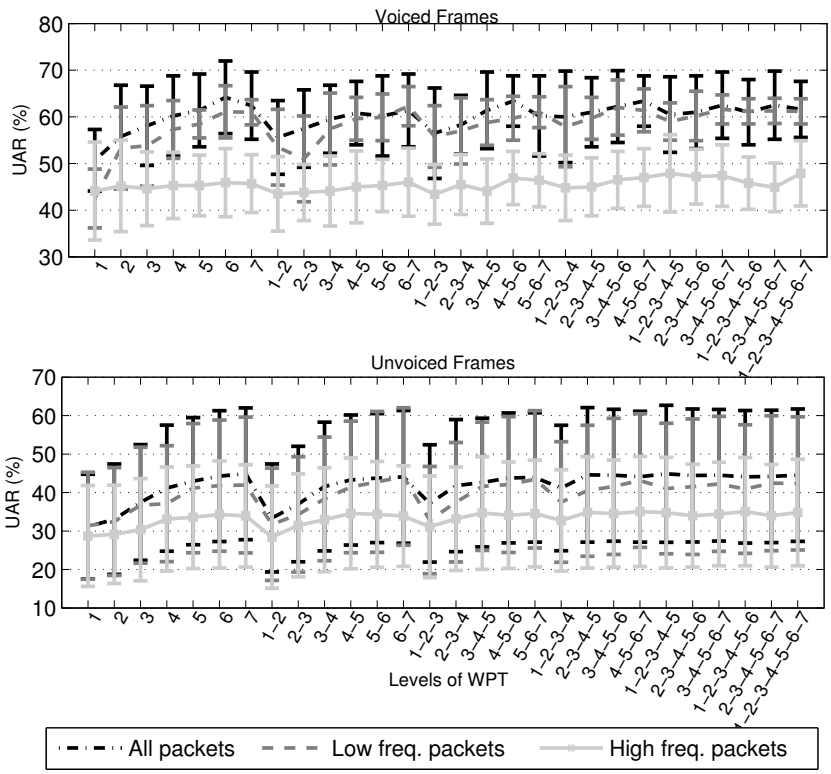

Fig. 6.1 Results for WPT in different levels. Voiced frames (up), Unvoiced frames (bottom)

The results indicate that the highest results are obtained when all frequency packets are considered, but when only the low frequency packets are used, the results are close to those obtained with all coefficients in several levels. The lowest results are obtained when only the higher frequency nodes are classified. These results allow to conclude that the low frequency packets contain the most suitable information to characterize emotions from speech, which allow to reduce to the half the number of features without considering the high frequency nodes. According to the decomposition levels, the highest results are obtained considering the levels 3, 4, 5, and 6. Levels 1 and 2 produce a low resolution in frequency and a small number of features, which causes that those nodes may not suitable to increase the results. On the other hand the seventh level provides a high resolution in frequency and a low resolution in time, which produces a high number of features that increase the complexity of the system. For that case the improvement in the results is low relative to the number of added features.

6.1.5 Summary and Comparisons

Table 6.5 shows the comparison between the results obtained with the proposed features relative to those reported in the state of the art both for arousal, valence, and classification of

all emotions in Berlin, enterface05, IEMOCAP, and FAU-Aibo. In FAU-Aibo and IEMOCAP the same constellations of train and test sets were also considered, which make that the results be directly comparable. The results obtained for the discrimination between high-low arousal emotions, and positive-negative valence emotions are similar to the reported in the state of the art using the same databases. For the classification of all emotion the results obtained in Berlin and enterface05 databases does not equal those reported in the related works. However, in FAU-Aibo and IEMOCAP, we obtain results close to the reported in the state of art.

Table 6.5 Comparison of the results obtained with the state of the art

Source	**# Feat.**	**Arousal**	**Valence**	**All**
Berlin database				
[37]	384	96.0%	80.0%	80.0%
[39]	6552	97.4%	87.5%	81.9%
[49]	42	–	–	77.9%
[56]	12 per frame	–	–	68.1%
[134]	88	97.8%	86.7%	86.0%
OpenEAR	384	97.3%	87.2%	80.4%
Acoustic+NLD	19 per frame	96.9%	82.9%	69.2%
WPT	128 per frame	95.7%	81.2%	66.1%
SSWT	22 per frame	95.8%	81.7%	69.3%
enterface05 database				
[37]	384	76.0%	65.0%	68.0%
[39]	6552	80.8%	79.7%	61.1%
[40]	1582	–	–	69.3%
[49]	42	–	–	53.9%
OpenEAR	384	81.0%	81.4%	63.2%
Acoustic+NLD	19 per frame	80.2%	74.9%	49.0%
WPT	128 per frame	79.7%	75.9%	49.2%
SSWT	22 per frame	81.1%	75.6%	48.0%
IEMOCAP database				
[38]	384	–	–	56.3%
[43]	513	–	–	56.7%
[41]	1584	–	–	63.1%
OpenEAR	384	75.5%	59.0%	57.2%
Acoustic+NLD	37 per frame	75.1%	59.5%	56.4%
WPT	128 per frame	75.4%	59.1%	57.1%
SSWT	22 per frame	77.2%	59.5%	58.2%
FAU-Aibo database				
[38]	384	–	–	39.9%
[42]	384	–	64.2%	–
[50]	1584	–	–	44.2%
[134]	88	–	76.5%	43.1%
OpenEAR	384	–	62.0%	32.5%
Acoustic+NLD	19 per frame	–	69.6%	38.9%
WPT	128 per frame	–	68.2%	38.0%
SSWT	22 per frame	–	70.3%	32.6%

6.2 Results in Signals Corrupted by Additive Noise

For this experiment, we consider the speech recordings from Berlin, enterface05, and FAU-Aibo databases corrupted by two environmental additive noise conditions: Street and cafeteria

babble. The two SE algorithms (KLT and logMMSE) are also considered. The performance of four different feature sets are evaluated with the noisy signals: (1) the OpenEAR features, (2) the spectral+noise+NLD features, (3) the WPT energy+MFCC+stat, and (4) the features derived from SSWT. The next subsections include the results for the four experiments.

6.2.1 Results Experiment 1: High vs Low Arousal Emotions

Figures 6.2, 6.3, 6.4, and 6.5 show the results obtained with the four feature sets. The figures show the results obtained in Berlin and enterface05 databases. The red line in each figure indicates the result obtained with the original (noise-free) recordings. The horizontal axis indicates the results at different SNR levels, and each bar shows the results of the noisy and enhanced signals.

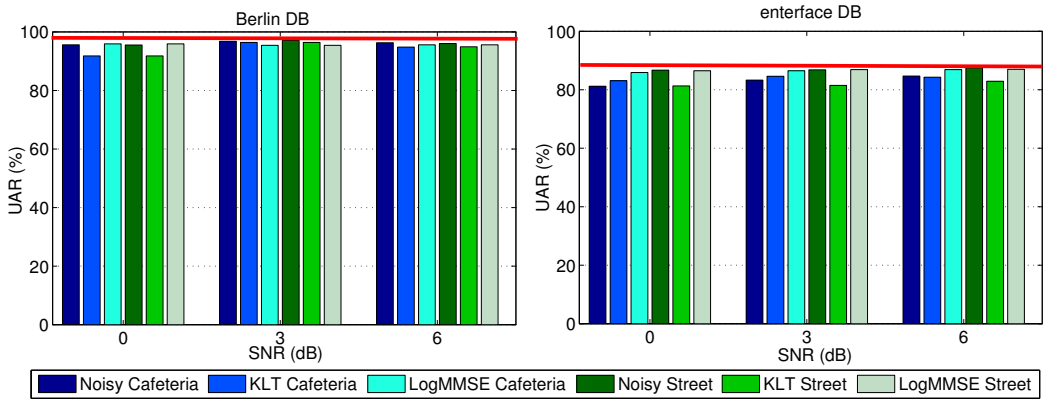

Fig. 6.2 Classification of Arousal considering OpenEAR features

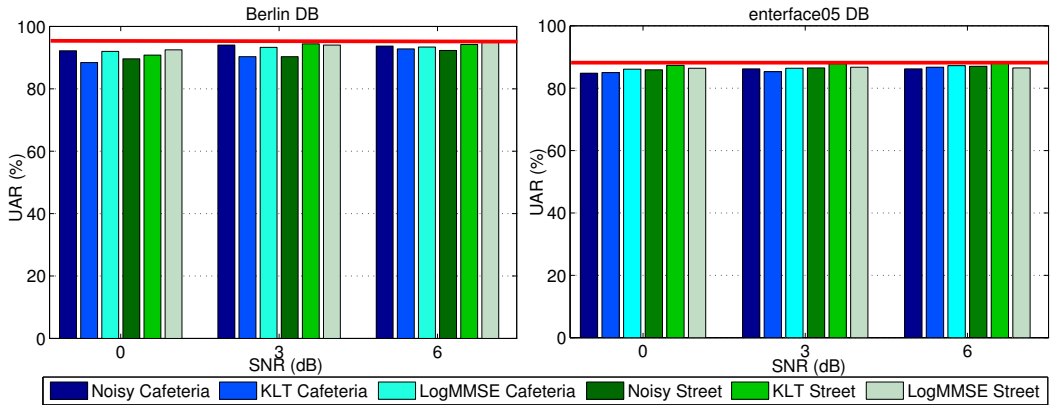

Fig. 6.3 Classification of Arousal considering spectral+noise+NLD features

For this experiment note that there is not a high difference in the results between those obtained with the noisy and enhanced signals for the four feature sets.

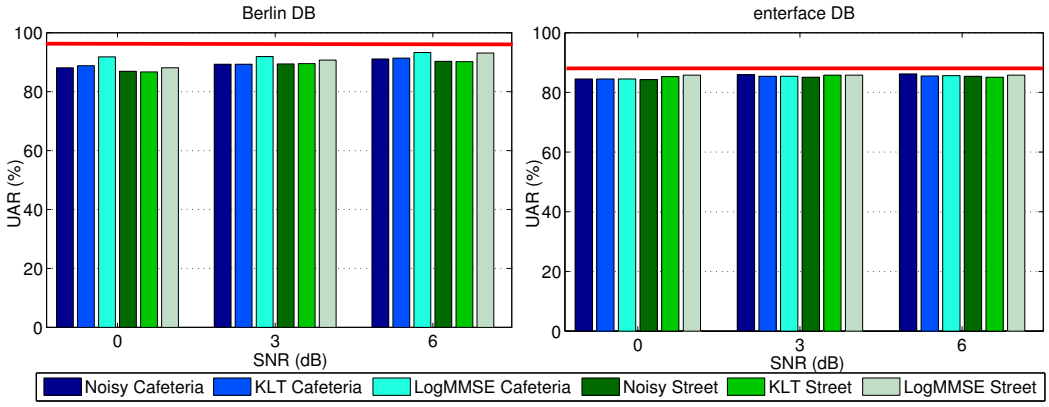

Fig. 6.4 Classification of Arousal considering WPT based features

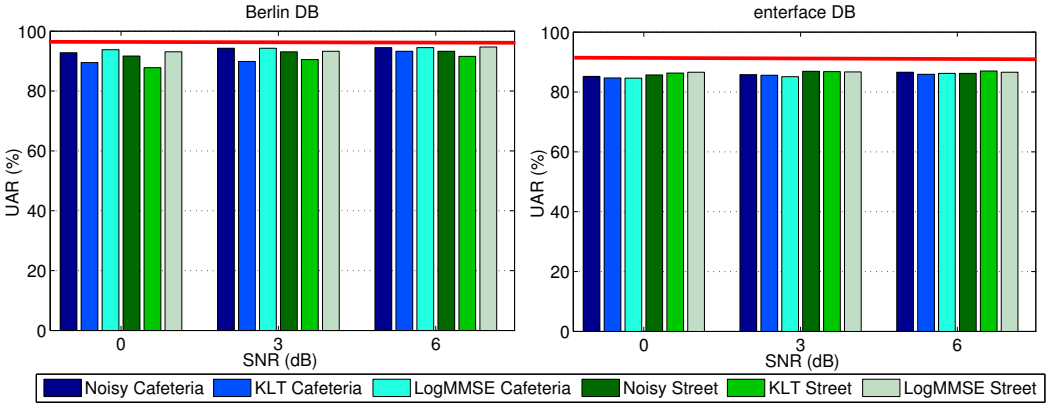

Fig. 6.5 Classification of Arousal considering SSWT based features

6.2.2 Results Experiment 2: Positive vs Negative Valence Emotions

Figures 6.6, 6.7, 6.8, and 6.9 contain the results for the evaluation of openEAR, spectral+noise+NLD, WPT based and SSWT based features in the noisy conditions for the valence detection.

For the case of OpenEAR, there is a reduction in the results due to the noise in most of cases in the three databases, except with the street noise in Berlin. Note that the degradation produced by the the cafeteria noise is more critical than the produced by the street noise, in the three databases, and that in enterface05 the logMMSE improves the results

For the case of spectral+noise+NLD features there is a high difference between the results obtained with the noisy and the noise-free signals, specially in Berlin and FAU-Aibo. Note also that in FAU-Aibo non of the SE algorithms improve the results relative to the noisy signals.

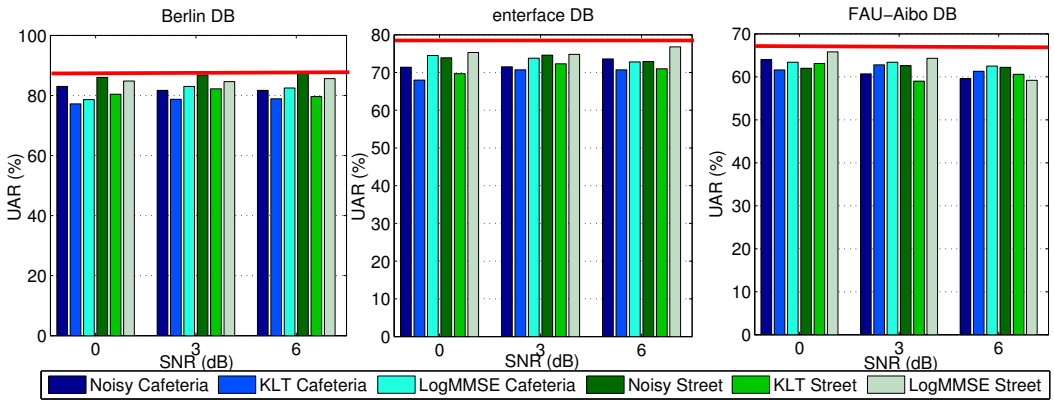

Fig. 6.6 Classification of Valence considering OpenEAR features

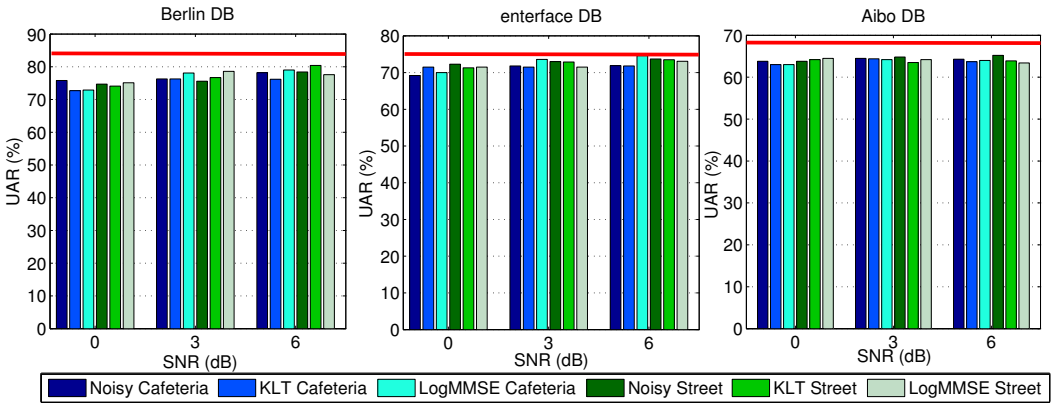

Fig. 6.7 Classification of Valence considering spectral+noise+NLD features

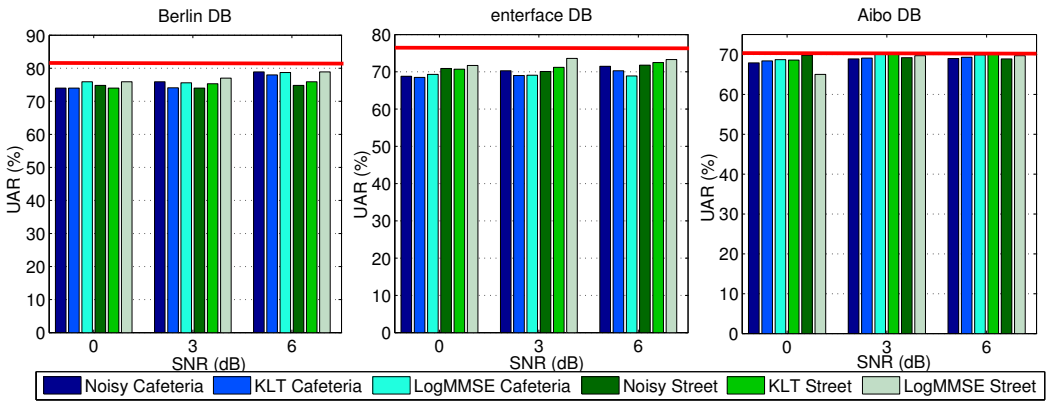

Fig. 6.8 Classification of Valence considering WPT based features

The logMMSE improves the results for the features based on the WPT for the signals affected by the street noise both in the Berlin and, the enterface05.

In general for all feature sets, in FAU-Aibo database the SE algorithms do not improve the results relative to the obtained with the noisy signals. The hypothesis for such case is that both SE techniques need a silence part of fixed duration for the characterization of background noise, and in FAU-Aibo database, there is not provided such silence segment due to the way that were segmented the recordings of the database.

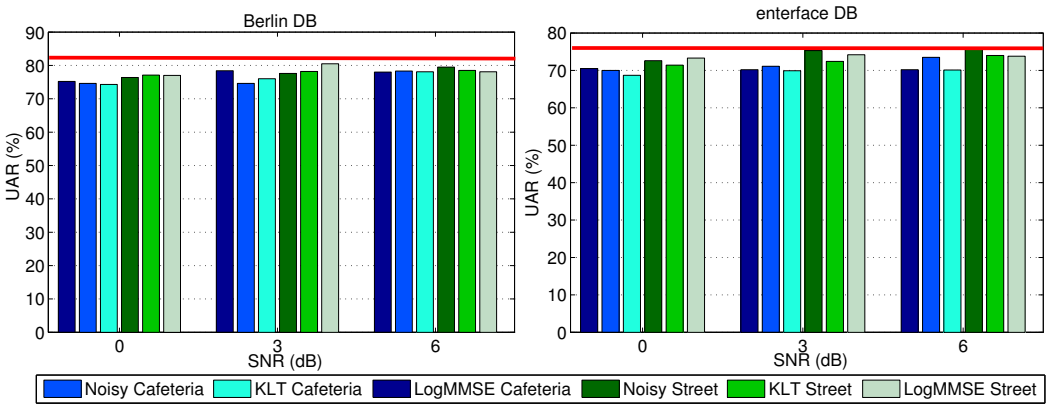

Fig. 6.9 Classification of Valence considering SSWT based features

Finally for the SSWT based measures note that the signals affected by cafeteria produce the lowest results in all cases.

6.2.3 Results Experiment 3: Fear-Type Emotions

Figures 6.10, 6.11, 6.12, and 6.13 contain the results of the classification of the fear-type emotions for openEAR, spectral+noise+NLD, WPT and SSWT based features, respectively.

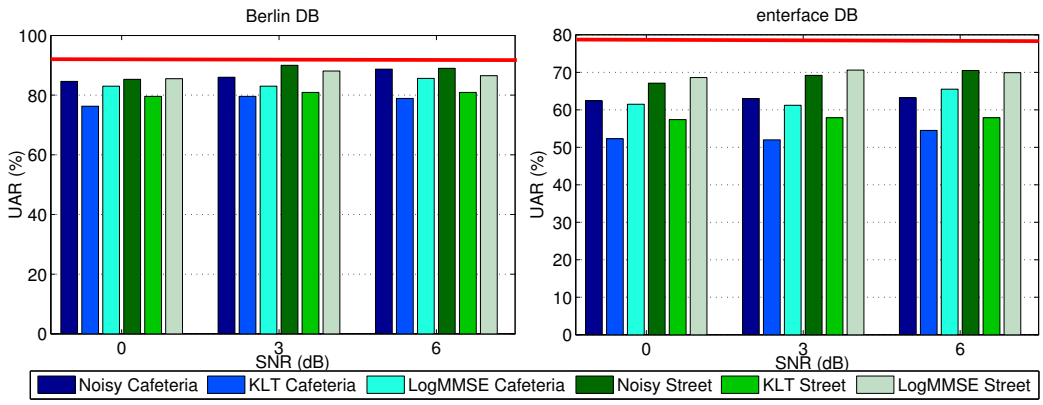

Fig. 6.10 Classification of Fear-type emotions considering OpenEAR features

Note that in enterface05 when the OpenEAR feature set is used there is a high difference between the results of the classification of the noisy signals, and those obtained with the

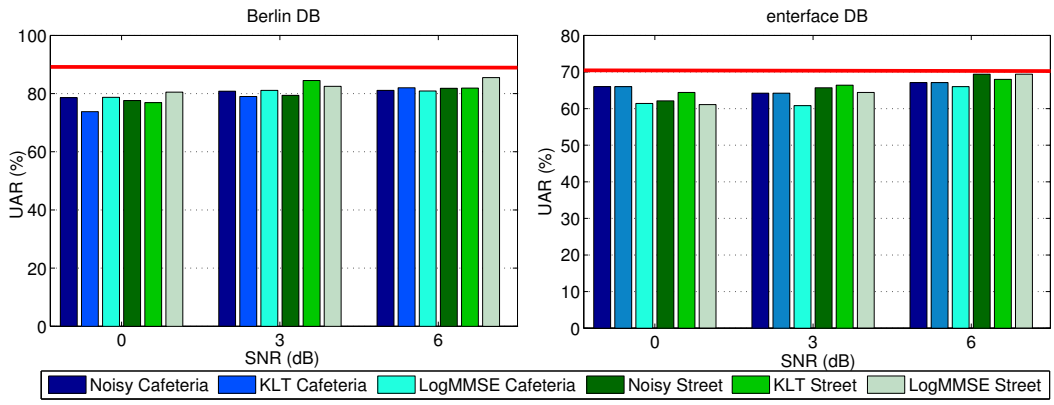

Fig. 6.11 Classification of Fear-type emotions considering spectral+noise+NLD features

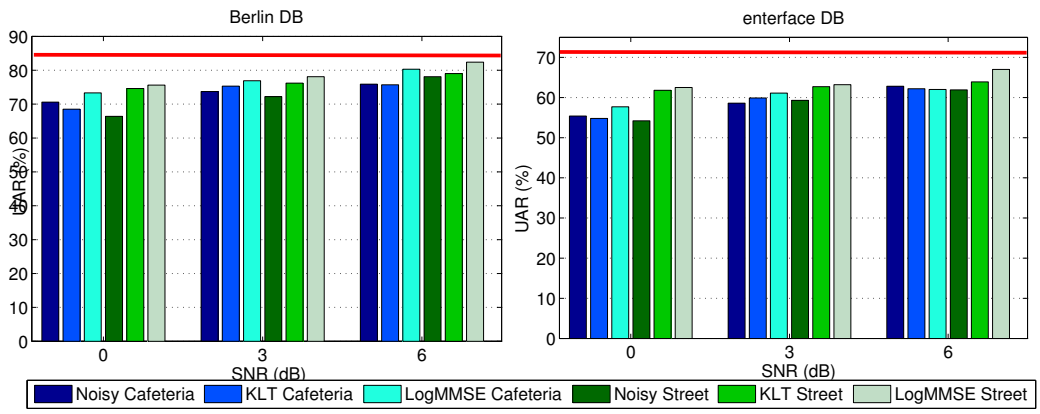

Fig. 6.12 Classification of Fear-type emotions considering WPT based features

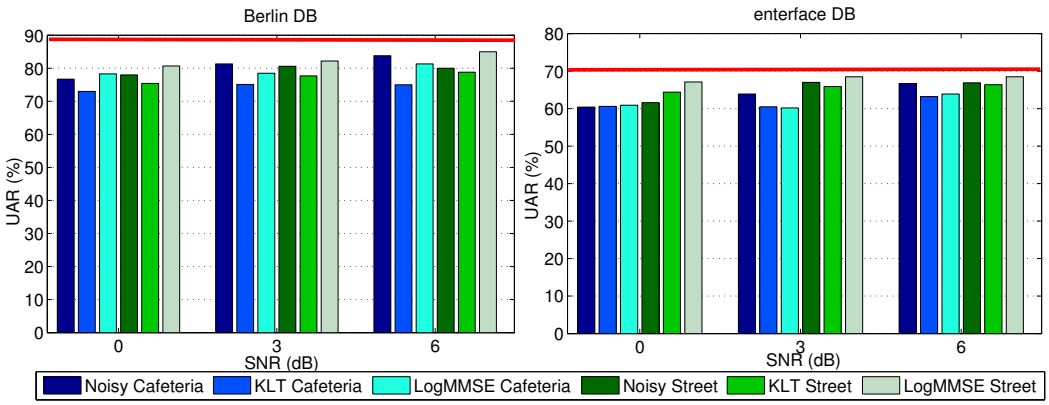

Fig. 6.13 Classification of Fear-type emotions considering SSWT based features

original recordings. Note also the cafeteria babble noise produces the lower results, and the KLT algorithm reduces the performance of the classifier in all cases.

For the case of WPT and SSWT based measures the logMMSE algorithm improves the results in almost all cases, specially in street noise; for example in the WPT in enterface05 when SNR = 6 dB, and in SSWT in Berlin in all cases and enterface05 when SNR = 0 dB.

6.2.4 Results Experiment 4: Multiple Emotions

Figures 6.14, 6.15, 6.16, and 6.17 show the results of the OpenEAR, spectral+noise+NLD, WPT, and SSWT for the classification of all emotions from the databases.

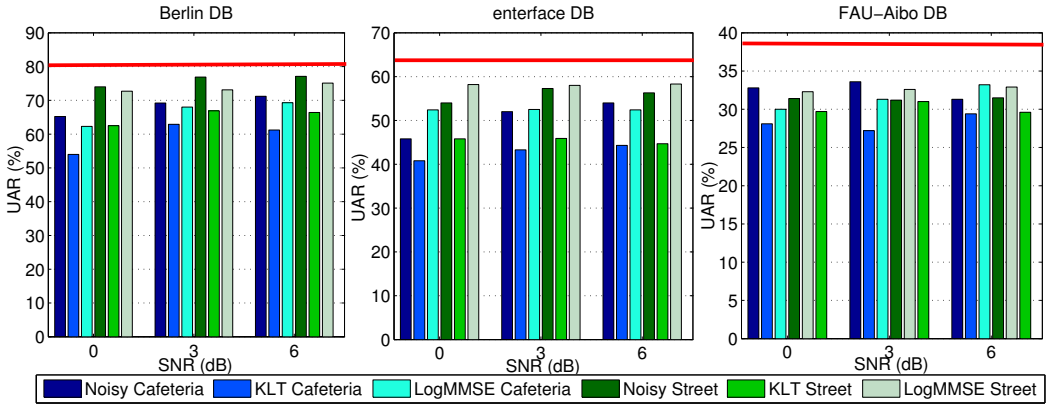

Fig. 6.14 Classification of All emotions considering OpenEAR features

Note that with OpenEAR the logMMSE improves the results in enterface05, specially in the signals affected by street noise. In FAU-Aibo the speech enhancement algorithms do not improve the results due to the reasons previously described.

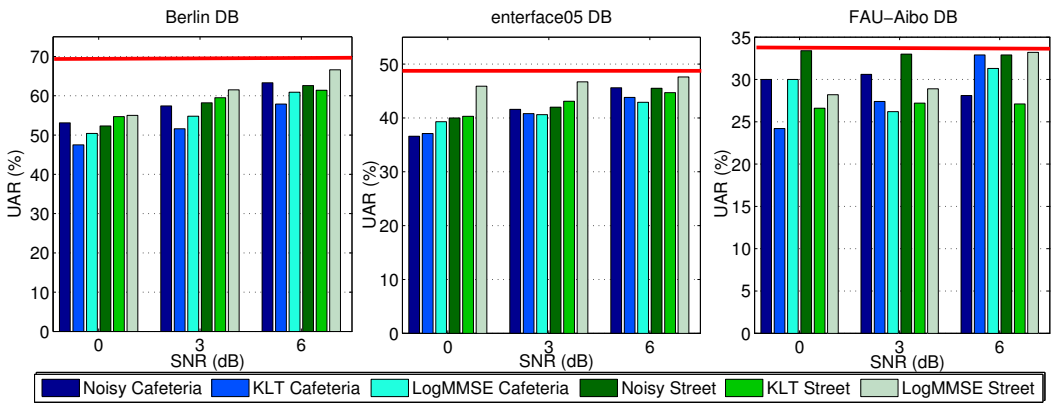

Fig. 6.15 Classification of All emotions considering spectral+noise+NLD features

For the case of spectral+noise+NLD features, note that the SE techniques improve the results relative to the obtained with the noisy signals, both in Berlin and enterface05,

specially the logMMSE algorithm evaluated in street noise, as can be observed in Berlin when SNR $= 6\,dB$, and in enterface05 when SNR $= 0\,dB$, and SNR $= 3\,dB$.

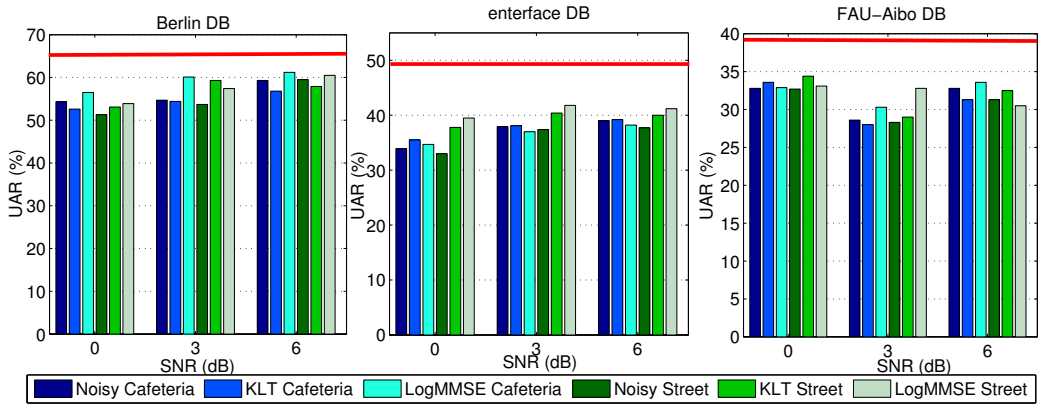

Fig. 6.16 Classification of All emotions considering WPT based features

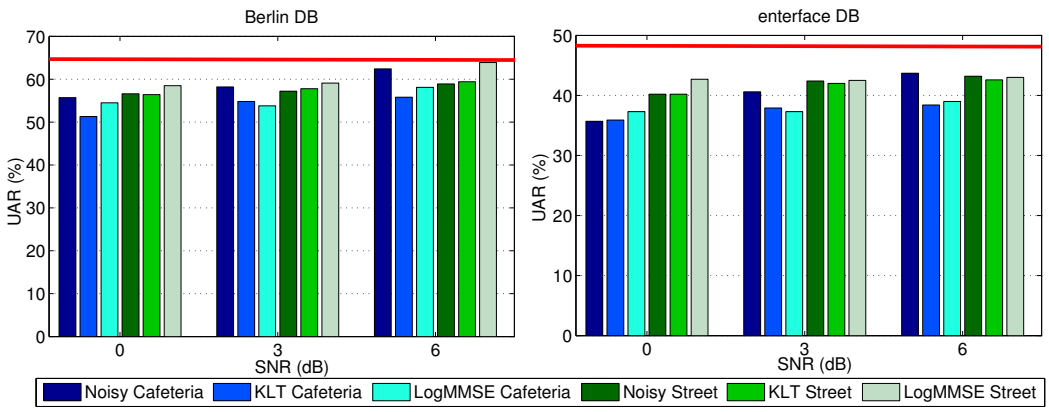

Fig. 6.17 Classification of All emotions considering SSWT based features

Finally, for the wavelet based measures, the logMMSE also improves the results for the case of street noise in Berlin and enterface05 databases.

6.3 Results in Signals Recorded in Noisy Environments

In this case the recordings of Berlin database were re-captured in two kinds of environmental noise: street and office. The recordings were also processed by the KLT and logMMSE algorithms. Table 6.6 shows the results obtained for the four experiments and considering four feature sets: (1) OpenEAR, (2) spectral+noise+NLD measures, (3) The WPT based features calculated on the selected decompositions, and (4) the SSWT based features.

Table 6.6 Results for Berlin DB re-captured in noisy environments

Recordings	**OpenEAR**	**Acoustic+NLD**	**WPT MRA**	**SSWT**
High-Low arousal detection				
Street Noise	95.8 ± 6.1	95.5 ± 6.7	96.5 ± 3.2	95.5 ± 6.2
Office Noise	96.4 ± 4.0	94.9 ± 6.2	96.3 ± 4.5	96.7 ± 4.0
KLT Street	95.1 ± 5.0	95.1 ± 6.8	96.7 ± 3.2	96.4 ± 5.6
KLT Office	95.6 ± 5.1	94.4 ± 7.0	97.1 ± 3.4	96.7 ± 4.5
logMMSE Street	95.7 ± 4.0	94.7 ± 5.8	96.4 ± 2.9	96.2 ± 4.6
logMMSE Office	96.1 ± 3.7	95.2 ± 5.2	96.0 ± 3.9	96.0 ± 3.3
Original	97.3 ± 3.0	96.9 ± 4.4	95.7 ± 4.0	95.8 ± 5.5
Positive-Negative valence detection				
Street Noise	86.0 ± 2.5	81.9 ± 6.9	79.3 ± 4.7	82.3 ± 4.5
Office Noise	88.0 ± 3.3	83.5 ± 5.7	80.2 ± 4.9	81.3 ± 5.7
KLT Street	85.3 ± 4.2	83.1 ± 5.7	80.4 ± 4.2	81.8 ± 5.3
KLT Office	86.5 ± 4.0	83.1 ± 6.8	80.1 ± 6.3	81.6 ± 3.3
logMMSE Street	86.3 ± 5.0	82.6 ± 7.2	78.3 ± 6.2	81.3 ± 6.7
logMMSE Office	82.6 ± 4.2	82.2 ± 4.8	78.4 ± 5.9	81.8 ± 5.7
Original	87.2 ± 2.4	82.9 ± 5.8	81.2 ± 3.3	81.7 ± 4.6
Fear-Type emotion classification				
Street Noise	91.6 ± 5.3	87.3 ± 9.8	84.8 ± 7.1	85.6 ± 9.1
Office Noise	91.3 ± 3.7	86.9 ± 9.4	87.0 ± 6.1	87.2 ± 9.2
KLT Street	85.4 ± 6.4	87.0 ± 8.0	85.5 ± 6.8	85.5 ± 9.1
KLT Office	89.9 ± 5.3	87.9 ± 9.8	85.3 ± 5.4	85.2 ± 7.9
logMMSE Street	86.9 ± 5.7	85.2 ± 7.8	81.5 ± 5.9	82.8 ± 10.3
logMMSE Office	87.1 ± 7.4	85.5 ± 8.9	81.3 ± 4.2	84.3 ± 6.0
Original	91.4 ± 5.0	88.3 ± 9.8	84.7 ± 5.7	88.3 ± 7.0
All emotion classification				
Street Noise	78.1 ± 4.4	67.3 ± 9.0	64.7 ± 5.0	63.3 ± 4.0
Office Noise	76.5 ± 6.3	66.8 ± 9.3	66.6 ± 4.4	65.4 ± 7.5
KLT Street	75.5 ± 7.9	68.1 ± 10.0	65.3 ± 4.8	63.7 ± 6.9
KLT Office	73.7 ± 7.0	69.0 ± 10.2	66.5 ± 4.9	63.5 ± 7.7
logMMSE Street	75.8 ± 6.0	66.2 ± 7.0	62.0 ± 5.7	59.9 ± 6.2
logMMSE Office	74.9 ± 5.7	66.0 ± 7.6	64.3 ± 5.9	61.9 ± 6.4
Original	80.4 ± 8.0	69.2 ± 10.2	65.0 ± 3.8	64.0 ± 8.0

The results indicate that there is not a great effect produced by the re-capturing process. For the classification of low-high arousal emotions, the highest difference between the results obtained with the noisy and the noise-free signals is 2%, which is found with the spectral+noise+NLD feature set in the recordings affected by office noise. For the discrimination between positive-negative valence emotions the highest reduction is also of 2% with the WPT based measures, in signals affected by street noise. For the classification of the fear-type emotions, the lowest result is produced also by street noise, but using the SSWT based features. Finally for the classification of all emotions, the highest difference between the results of noise-free and noisy signals is of 4%, which is obtained with the features computed with OpenEAR. Note also that the two algorithms for SE do not improve the results relative to the obtained with the noisy recordings, contrary to the case of the additive noise, when the logMMSE improved the results. The hypothesis to explain this behavior is that when signals are re-captured, the loss of quality is less critical than when the recordings are affected by additive artificial noise. This fact reflect that the SE algorithms only provide an improvement in cases when the SNR is lower than a certain threshold. In other words when the speech

recordings are affected by a critical background noise, instead of the normal noise captured in an office environment.

6.4 Results in Signals Compressed by Telephony Codecs

These environments include the evaluation of seven codecs for speech compression, which are commonly used in VoIP, and mobile telephone channels, the codecs are: AMR-NB, AMR-WB, GSM, G.722, G.726, SILK, and Opus. The detailed description was performed in Section 4.2. Different bit-rates of the AMR-NB, and AMR-WB were also considered. The results are compared to those obtained with the original, and down-sampled to 8 kHz recordings.

Tables 6.7 and 6.8 contains the results for the evaluation of telephony codecs both for the Berlin and enterface05 databases, respectively. the four experiments, and the four feature sets evaluated.

For the case of Berlin database for the classification of high vs low arousal emotions, note that the results obtained with all the feature sets were not affected by the presence of the codecs. For the discrimination between positive and negative emotions the results with OpenEAR features are decreased in up to 7%, while the highest decreasing in the results with the other feature sets is 5% for the spectral+noise+NLD, 7% with the WPT-AMR, and 1% with the SSWT. For the fear-type emotion classification the highest difference between the results of compressed and original recordings is: 9% using openEAR, 6% using the spectral+noise+NLD, 14% considering the WPT-AMR, and 6% using the SSWT based features. Finally, for the classification of all emotions, the highest difference are is of 6% using OpenEAR, 7% considering the spectral+noise+NLD, 11% using the WPT-AMR based measures, and 3% with the SSWT based features.

Note that in all cases the highest reduction is produced by the AMR-NB codecs, with different bit rates, and the codecs that less decrease the performance are AMR-WB, and Opus. Note also that the SSWT based features are the less affected by the telephony codec compression, while the features extracted from signals compressed by WPT-AMR are those which exhibit the highest reduction due to the compression effect.

For the enterface05 database there is also a significant effect produced by the codecs for the classification of fear-type and all emotions, specially in AMR-NB, and G.726 codecs.

Table 6.7 Results in telephony codecs for Berlin database

Recording	Bit-rate [Kbps]	OpenEAR	Acoustic+NLD	WPT-MRA	SSWT
High-Low arousal detection					
Original	256	97.3±3.0	94.9±7.1	95.7±4.0	95.8±5.5
Down-sampled	128	95.7±3.9	94.5±6.8	94.1±5.0	94.9±4.3
AMR-NB	4.75	95.9±3.9	94.8±5.7	91.8±5.6	92.8±4.4
AMR-NB	7.95	95.3±4.0	94.3±6.7	92.3±3.3	94.5±5.1
GSM	12.2	96.6±4.2	93.6±6.2	91.5±2.9	94.3±5.2
AMR-WB	6.6	96.9±5.2	93.7±7.0	91.7±3.0	95.6±4.4
AMR-WB	23.85	96.8±6.2	93.9±7.0	92.5±3.7	95.5±5.4
G.722	64	97.2±5.6	93.4±6.7	93.6±3.7	95.5±5.8
G.726	16	96.9±2.9	93.3±5.9	91.4±3.1	94.2±4.7
SILK	64*	96.9±6.6	94.3±5.7	93.2±2.4	95.9±5.5
Opus	25*	98.7±3.3	94.2±5.8	93.6±3.8	95.7±5.3
Positive-Negative valence detection					
Original	256	87.2±2.4	83.5±4.3	81.2±3.3	81.7±4.6
Down-sampled	128	83.3±4.7	81.0±5.2	77.2±5.4	82.2±5.5
AMR-NB	4.75	82.6±7.3	78.9±4.7	74.8±6.9	80.5±5.7
AMR-NB	7.95	80.8±4.8	79.4±4.4	75.1±6.9	81.5±4.6
GSM	12.2	80.6±5.0	80.9±5.7	72.8±7.0	82.4±5.8
AMR-WB	6.6	86.0±3.5	81.3±4.5	74.1±7.2	82.1±4.7
AMR-WB	23.85	87.1±5.1	81.4±4.8	73.6±5.8	81.0±4.8
G.722	64	82.4±3.7	82.2±3.3	75.3±6.3	81.9±3.8
G.726	16	81.6±4.5	81.2±4.9	73.3±6.5	81.6±4.9
SILK	64*	83.7±2.6	79.0±5.5	76.8±6.7	82.0±5.0
Opus	25*	85.2±5.3	79.5±4.1	78.4±7.9	82.7±4.5
Fear-Type emotion classification					
Original	256	91.4±5.0	88.5±11.0	84.0±5.7	88.3±7.0
Down-sampled	128	82.1±6.4	85.0±9.1	81.8±5.6	84.6±5.7
AMR-NB	4.75	83.1±6.0	83.5±10.7	70.4±23.5	82.8±8.0
AMR-NB	7.95	84.9±5.9	84.6±12.4	78.9±5.4	83.6±5.8
GSM	12.2	81.9±4.5	82.0±8.8	81.3±6.5	82.0±5.5
AMR-WB	6.6	83.9±8.1	85.7±8.4	82.0±6.8	87.1±6.6
AMR-WB	23.85	86.9±7.1	84.0±9.5	81.3±10.0	85.3±8.4
G.722	64	87.7±4.2	86.1±8.0	76.1±26.1	87.0±5.5
G.726	16	82.2±4.8	80.0±7.5	77.1±15.6	83.7±5.7
SILK	64*	87.6±7.1	84.1±7.9	76.4±25.8	87.4±7.1
Opus	25*	90.2±6.1	85.0±8.3	76.5±25.8	86.5±6.4
All emotion classification					
Original	256	80.4±8.0	68.9±11.0	65.0±3.7	64.0±8.0
Down-sampled	128	74.4±6.4	65.5±8.1	61.9±7.2	65.2±7.0
AMR-NB	4.75	74.7±6.2	64.7±7.8	54.9±17.3	62.5±6.0
AMR-NB	7.95	74.0±5.4	64.6±8.0	61.6±4.1	62.8±5.0
GSM	12.2	74.3±5.6	64.5±6.8	61.0±6.5	63.6±6.6
AMR-WB	6.6	78.8±6.7	65.5±10.9	63.6±4.8	61.3±6.3
AMR-WB	23.85	77.5±6.3	62.4±9.0	63.9±4.9	64.7±9.6
G.722	64	79.8±6.6	64.5±7.0	58.2±19.4	66.6±7.9
G.726	16	75.8±7.6	64.5±7.3	59.9±5.7	62.1±7.4
SILK	64*	76.8±5.8	65.4±9.0	59.3±20.3	63.3±7.0
Opus	25*	77.5±5.1	61.8±8.8	59.2±20.2	64.7±6.3

* Mean bit rate

6.5 Summary

The main results obtained with the proposed approaches were described in this section. The different feature sets based on acoustics, NLD, and wavelet measures were tested in four experiments: (1) detection of high vs low arousal emotions, (2) classification of positive vs negative valence emotions, (3) recognition of fear-type emotions, and (4) the classification of

Table 6.8 Results in telephony codecs for enterface05 database

Recording	Bit-rate [Kbps]	OpenEAR	Acoustic+NLD	WPT-MRA	SSWT
High-Low arousal detection					
Original	256	81.0±2.0	80.2±1.9	80.7±5.0	81.1±1.7
Down-sampled	128	80.3±4.0	79.4±0.9	77.8±1.5	79.4±1.9
AMR-NB	4.75	82.2±3.6	79.2±1.0	77.9±1.0	79.3±1.3
AMR-NB	7.95	83.1±3.7	80.0±1.4	77.7±1.1	79.6±1.8
GSM	12.2	82.3±3.6	79.0±1.3	77.2±0.7	79.6±1.5
AMR-WB	6.6	83.3±2.7	80.0±1.5	77.2±1.0	80.5±1.4
AMR-WB	23.85	82.7±3.9	79.5±1.8	77.6±0.9	80.5±1.8
G.722	64	82.5±3.9	79.5±1.3	78.4±1.4	80.5±1.2
G.726	16	82.0±3.7	79.5±1.7	76.6±0.5	79.8±1.9
SILK	64*	82.6±3.9	78.5±1.2	78.9±1.6	81.1±1.7
Opus	25*	83.5±4.1	78.8±0.9	81.3±1.9	79.8±1.9
Positive-Negative valence detection					
Original	256	81.4±5.5	75.1±2.5	75.9±3.0	75.6±2.9
Down-sampled	128	78.6±3.3	74.2±2.8	69.6±1.7	75.0±2.4
AMR-NB	4.75	77.8±3.8	73.8±2.2	68.5±1.0	73.7±2.4
AMR-NB	7.95	77.0±4.6	73.6±2.3	70.4±2.0	74.8±2.3
GSM	12.2	77.6±3.3	73.9±2.5	69.6±1.4	74.7±3.3
AMR-WB	6.6	79.4±3.8	75.0±3.1	71.9±1.9	75.2±1.9
AMR-WB	23.85	77.6±4.2	74.5±1.7	71.6±2.8	75.4±1.8
G.722	64	78.7±3.9	74.3±2.4	71.8±1.6	75.9±2.9
G.726	16	78.5±3.0	73.1±2.1	68.5±1.0	74.6±3.0
SILK	64*	78.5±4.5	74.5±2.5	73.0±3.2	76.6±3.5
Opus	25*	77.9±3.9	72.2±1.4	71.4±3.0	75.5±1.3
Fear-Type emotion classification					
Original	256	78.2±5.5	71.6±4.9	70.9±4.6	70.1±5.8
Down-sampled	128	76.0±4.8	70.2±4.9	59.2±4.3	69.5±5.2
AMR-NB	4.75	75.8±5.0	70.0±6.2	58.7±4.7	67.7±4.6
AMR-NB	7.95	75.8±3.8	68.1±4.9	57.0±4.6	70.0±4.7
GSM	12.2	76.7±4.4	69.0±5.5	58.0±3.0	70.3±4.2
AMR-WB	6.6	77.5±4.0	73.3±3.7	59.9±4.7	70.8±4.1
AMR-WB	23.85	79.0±3.5	70.4±5.3	63.0±6.4	72.3±6.4
G.722	64	79.0±5.1	71.7±4.5	62.9±5.4	71.0±5.6
G.726	16	76.6±4.6	69.7±4.4	58.7±5.0	67.7±5.3
SILK	64*	77.6±4.0	70.7±6.2	65.5±3.9	72.1±6.7
Opus	25*	76.6±4.7	71.9±6.0	62.5±4.1	69.9±5.9
All emotion classification					
Original	256	63.2±6.7	49.9±5.4	49.2±3.0	48.0±3.5
Down-sampled	128	55.5±6.0	45.8±4.0	35.4±3.0	45.3±4.2
AMR-NB	4.75	54.6±5.7	45.5±4.9	34.8±3.7	44.7±3.6
AMR-NB	7.95	55.0±4.8	45.0±4.1	35.0±4.3	47.0±4.1
GSM	12.2	55.0±4.5	46.0±5.1	34.7±3.1	45.7±4.6
AMR-WB	6.6	56.7±5.0	48.8±4.3	37.5±4.4	47.9±3.4
AMR-WB	23.85	58.1±4.0	47.7±3.5	39.6±4.7	48.4±3.5
G.722	64	57.4±6.1	47.5±4.1	40.3±4.9	50.1±3.6
G.726	16	54.1±4.8	45.4±4.0	33.5±3.4	44.7±3.6
SILK	64*	58.2±4.3	46.9±5.2	42.5±5.1	49.6±3.8
Opus	25*	57.0±4.3	45.6±4.0	41.4±3.8	48.4±4.4

* Mean bit rate

multiple emotions. The experiments were performed in different acoustic conditions: the noise-free speech signals, the signals corrupted by additive noise, the re-captured recordings in noisy environments, and the signals compressed by telephony codecs. For the noisys conditions, the performance of two SE algorithms is also tested. According to results, all feature sets are more robust for the discrimination between high and low arousal emotions than for the detection of positive vs negative valence. The results obtained for the 2-classes

classification experiments are close to the reported in the state of the art. The proposed approaches are also suitable for the recognition of fear-type emotions from speech. The next chapter describes the main conclusions derived from this study and the future work.

Chapter 7

Conclusion and Future Work

This research work aims in the development of a methodology for the automatic recognition of emotions from speech signals in non-controlled noise conditions. For that purpose, different feature sets were proposed. Measures derived from acoustic, non-linear, and wavelet analysis were computred to characterize the emotions from five different databases widely used in the state of art: Berlin, enterface05, SAVEE, FAU-Aibo, and IEMOCAP. The non-controlled noise conditions were tested considering four scenarios: (1) the original noise-free recordings, (2) the signals corrupted by two additive environmental noises, which were recorded in a street and a cafeteria, (3) the re-captured signals in two natural noisy environments as street and office, and (4) the recordings compressed by seven codecs used for the transmission through different telephone channels. A classification scheme based on the combination of GMM and SVM was used for the analysis.

For the original noise-free recordings, all the feature sets selected are more suitable for the recognition of emotions according to the arousal dimension rather than the valence domain. As consequence, there is a strong need for the definition of features which are useful to discriminate between different emotions which are similar according to the arousal dimension and different in the valence such as happiness and anger.

The results obtained for the discrimination between high vs arousal emotions, and positive vs negative valence emotions are similar to the reported in the state of the art, with the advantage that some of the proposed approaches consider a less number of features than the used in related works, using the same databases. For the fear-type emotion classification different results are obtained according to the feature set used and the database. The highest results are obtained in Berlin, followed by SAVEE and enterface05. Finally for the classification of all emotion, the results obtained in Berlin, enterface05, and SAVEE databases do not equal those reported in the state of art. However, in FAU-Aibo and IEMOCAP databases, we obtain results close to the reported in related works for that multi-class analysis.

According to the different feature sets used in this study, the highest results are achieved by considering the features derived from the acoustic analysis, the wavelet multi-resolution decomposition, and the synchro-squeezed wavelet transform. On the other hand, the NLD measures improve the results obtained when they are combined with the spectral features, indicating that the NLD analysis provides suitable information to characterize emotions from speech signals, and it can be used as complement to the conventional acoustic analysis.

The use of TARMA models to characterize emotions in speech is also proposed to analyze the non-stationary processes produced in the speech signals. The results show be promising for such analysis. However, a deeper study related to such methods and models must be addressed to find the most suitable features in order to improve the results.

The multi-resolution analysis using the WPT proves to be useful to characterize the emotional content in speech. According to the results, the analysis of the low frequency zone of the spectrum characterized considering the WPT produces the same results than the obtained considering all spectrum, which allow to reduce the number of features in half. Also the highest results are obtained for levels three, four, and five.

The time-frequency representations considered in this study include different versions of the wavelet transform, as the conventional continuous representation of the wavelet transform, the bionic wavelet transform, and the synchro-squeezed wavelet transform. The best results are achieved using the SSWT in most of cases, indicating that the re-allocating method that sharpens the frequency components of the spectrum to a narrower band provides to be useful to characterize emotional speech.

Most of the characterization approaches proposed in this work are evaluated separately for voiced and unvoiced segments. Then a fusion scheme is used to combine both feature sets. For the most of cases, higher results are obtained with the features derived from voiced segments instead of the calculated from unvoiced segments. The fusion scheme shows to be useful in cases when both feature sets produce similar results in the separately classification.

For the analysis of signals in non-controlled noise conditions, the performance of two algorithms for speech enhancement was evaluated to determine if they are suitable to improve the classification results when the speech signals are corrupted by noise with different SNR levels. The first algorithm is based on a statistical characterization of noise (logMMSE), and the other is based on the subspace decomposition of the speech signals using a transformation based on PCA (KLT).

The effect of two noise environments is evaluated in this study by adding to the original signals the recordings of background noise from a cafeteria babble, and a street. The results obtained for such environments indicate that the reduction in the classification rate is more critical for the cafeteria babble than for the street noise. For the most of cases the logMMSE

algorithm is able to improve the results respect to those achieved with the noisy signals. The suitability of KLT algorithm to improve the classification rate is not clear due to some contradictory results.

An additional experiment was performed to evaluate the effect of non-additive noise conditions. For that case the recordings were re-captured under two different noisy environments using a professional audio monitor, an omnidirectional microphone, and a professional audio card. The noisy environments consider the noise produced in a street and in an office with an air conditioning unit. The results indicate that there is no a significant effect in the classification rates due to the effect of these noisy conditions. The reduction is lower than the produced in the additive noise conditions. The SNR for the re-captured signals is around 8 dB compared to the evaluated in the noisy additive environments (0 dB, 3 dB, and 6 dB). The speech enhancement algorithms do not improve the classification rate with respect to the corrputed signals.

For the future work, other speech enhancement approaches might be evaluated to guarantee a better performance in the classification of noisy speech signals. More naturalistic non-controlled noise scenarios must be considered, which allow evaluate the noise influences on the speaking style such as the Lombard effect, and non-additive noise.

The effect of compression of the speech signal with state-of-the-art codecs is also evaluated in this study. The evaluation of such effects is performed independently of other distortion of the signal when transmitting through a communications channel. The codecs were selected based on their relevance in modern communications systems such as VoIP and mobile phone networks. The results indicate that the compression does not produce a considerable degradation of the results. However, the bit-rate of the codecs used also plays a relevant role on the classification results. Lower bit-rates tend to decrease the results. Future work should address other distortions generated on the speech signal by the communications channel and how they affect the performance of automatic recognition of emotions systems, i.e., loss of packages and delays in VoIP channels.

Appendix A

Speech Enhancement

A.1 Statistical model based (logMMSE)

These methods were proposed in [127]. The algorithm is based on finding an estimator of the noise-free speech signal $x(t)$ that minimizes the mean square error calculated between the log-spectrum of the noise-free speech signal and the estimator. With such a purpose, the authors find an estimator obtained directly from the amplitude spectrum of the noisy observable signal $y(t)$, multiplied by a non-linear gain function which depends only of the a priori SNR estimated with the first 120 ms of the noisy speech signal and updated in each silence part.

Let $X_k = A_k e^{j\alpha k}$, N_k, and $Y_k = R_k e^{j\beta k}$ denote the $k-$th component of the spectrum of the clean, the noise, and the observable signals. The main objective is to find the estimator $\widehat{A_k}$ that minimizes the distortion measure expressed in Equation A.1. The estimator $\widehat{A_k}$ can be calculated according to the Equation A.2

$$E\left\{\left[log(A_k) - log(\widehat{A_k})\right]^2\right\} \tag{A.1}$$

$$\widehat{A_k} = exp\left\{E\left[log\left(A_k\right)|Y_k\right]\right\} \tag{A.2}$$

The desired estimator $\widehat{A_k}$ can be calculated according to Equation A.3. The procedure to derive such Equation is explained with detail in [127]. The term ξ_k is referred to the a priori SNR of the noisy signal, and v_k is defined according to Equation A.4.

$$\widehat{A_k} = \frac{\xi_k}{1+\xi_k} exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}R_k \tag{A.3}$$

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\lambda_k \tag{A.4}$$

Where the term λ_k satisfy the relationship $\frac{1}{\lambda_k} = \frac{1}{\lambda_x} + \frac{1}{\lambda_n}$. λ_x, and λ_n are the variance of the signal and the noise, respectively. Figure A.1 shows an example of the performance of this method. Figure includes the spectrogram of a signal corrupted by two different kinds of noise (cafeteria and street) with SNR $= 0dB$ before (left) and after (right) the speech enhancement. Note the improvement in the quality, specially for the case of the noisy street.

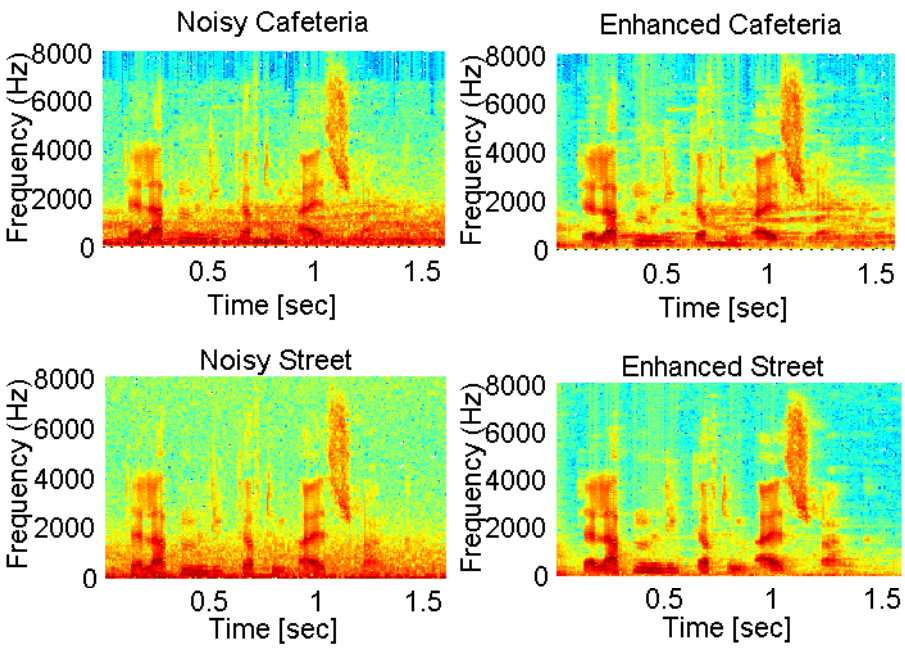

Fig. A.1 Performace of logMMSE technique for Speech enhancement

A.2 Subspace decomposition based (KLT)

This is a non-parametric technique based on the decomposition of the vector space of a noisy signal into two subspaces: one for the noise-free speech signal and other for the background noise. The decomposition can be performed by applying the Karnuhen-Loève Transformation (KLT) to the noisy signal [135]. Such transformation is related to the principal component analysis (PCA). For the analysis it is assumed that the noise is additive and uncorrelated with the noise-free signal. The methods based on subspace decomposition can minimize the speech distortion while keeping the residual noise below a present threshold [128, 135].

A generalized sub-space approach was proposed in [128] with the purpose of remove colored noise. The clean signal is estimated by nulling the signal components of the noise subspace and retaining the components related to the clean signal subspace.

The method focuses on finding a linear minimum mean-square error estimator of the clean signal: $\widehat{x} = H\mathbf{y}$. The authors propose an optimum H obtained from a matrix that can simultaneously diagonalize the covariance matrices of both the clean signal and the noise. H is computed according to Equation A.5. Where V is the matrix of eigenvector of the joint covariance matrix Σ of the noise and speech signals, which is estimated from the noisy signal; and Q is a diagonal matrix whose elements are computed from the positive eigenvalues of Σ.

$$H = R_n V Q V^T \tag{A.5}$$

The algorithm is as follows:

1. Compute the covariance matrix R_y of the noisy signal, and estimate the matrix $\Sigma = R_n^{-1} R_y - I$. The noise covariance matrix R_n must be calculated considering noise samples collected during speech absent frames.

2. Perform the eigenvalue decomposition of Σ as $\Sigma V = V \Lambda_x$.

3. Estimate the dimension M of the clean speech signal subspace as the number of eigenvalues of Σ which are higher than zero.

4. Estimate the value of μ to control the trade-off between the speech distortion and the residual of noise according to Equation A.6.

$$\mu = 4.2 - SNR_{dB}/6.25 \tag{A.6}$$

5. Compute the gain matrix G, and the transformation matrix H using Equations A.7, A.8 and A.9.

$$g_{kk} = \begin{cases} \frac{\lambda_x^{(k)}}{\lambda_x^{(k)} + \mu_k}, & k \leq M \\ 0, & k > M \end{cases} \tag{A.7}$$

$$G = diag(g_{11}, g_{22}, \ldots g_{MM}) \tag{A.8}$$

$$H = R_n V \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} V^T \tag{A.9}$$

6. Estimate the enhanced speech signal as $\widehat{x} = Hy$

As in previous method, Figure A.2 shows an example of the performance of this method both for cafeteria and street noise.

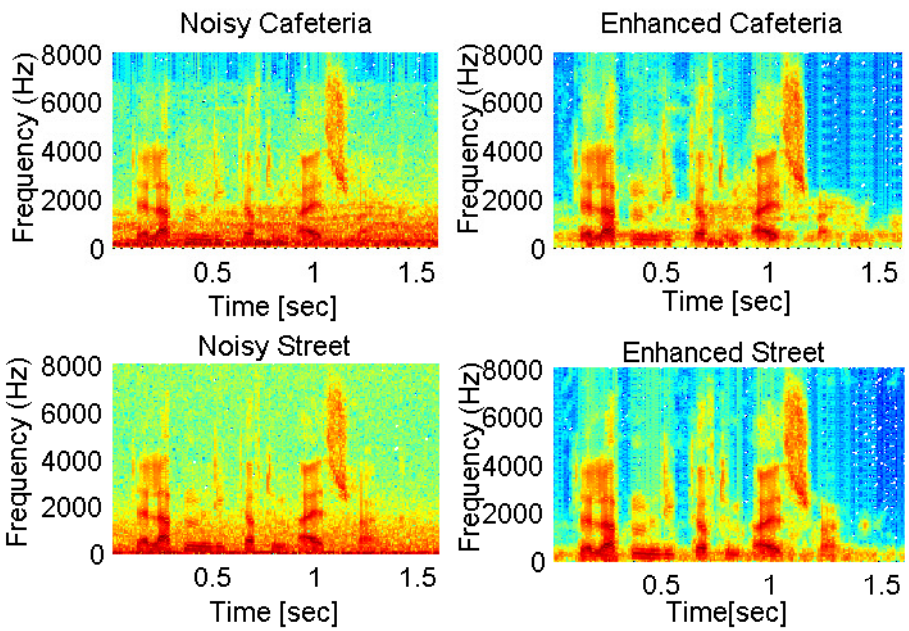

Fig. A.2 Performace of KLT technique for Speech enhancement

Appendix B

Classification

B.1 Gaussian Mixtures Models

A GMM can be defined as a probabilistic model represented by the linear combination of several multivariate Gaussian components. The model is expressed according to its probability density function using Equation B.1. Where M is the number of Gaussian components, P_j corresponds to the prior probability of the $j-$th component, and $\mathcal{N}$ is a multivariate Gaussian density function with mean vector μ_j, and covariance matrix Σ_j.

$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{M} P_j \mathcal{N}(x|\mu_j, \Sigma_j) \tag{B.1}$$

Training GMM consist of estimating the parameters $\Theta = \{P, \mu, \Sigma\}$ from a training set. The most common method for the estimation is the expectation maximization (EM) algorithm [136]. The UBM is trained using the Expectation Maximization (EM) algorithm [136] using recordings from all classes, *i.e.* emotions from the trainig set. Then the specific GMM for each class is adapted using the maximum a posteriori (MAP) rule. Finally, given a sample $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_T]$, where $\mathbf{x}_i$ is the feature vector extracted from the frame i, the decision of which class belongs each speech sample is taken evaluating the maximum log-likelihood (LL) of the model of each emotion. The log-likelihood is calculated according to Equation B.2

$$LL(\mathbf{X}|\Theta) = \frac{1}{T} \sum_{i=1}^{T} log(p(\mathbf{x}|\Theta)) \tag{B.2}$$

B.2 GMM Supervector

In this case, first a GMM-UBM is created, considering all classes, and all speakers. Then, the mean vectors μ_j of the UBM are adapted and merged toguether for each speaker utterance using the MAP rule in order to create the GMM supervector for each utterance [137]. A GMM supervector is a vectorial representation of the parameters of each one of these models. Figure B.1 shows the process of the creation of the GMM supervectors. Finally the supervectors are used as new features to train a discriminative classifier as a SVM. This method leads to a "hybrid" classification strategy, where the generative GMM-UBM model is used to create new feature vectors for the discriminative SVM.

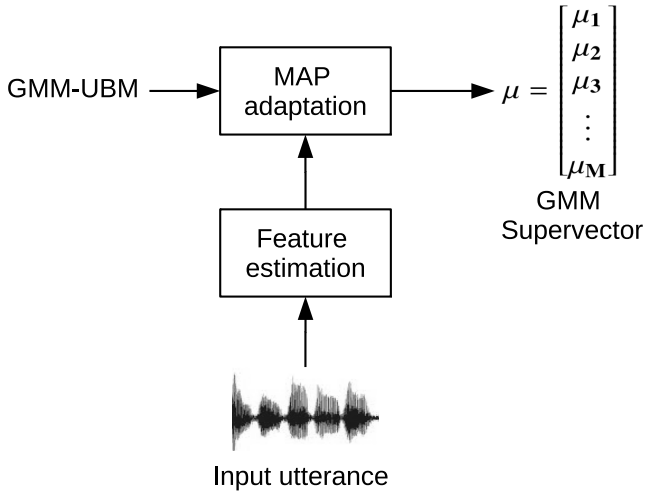

Fig. B.1 GMM Supervector construction

Appendix C

Publications

The following articles have been published during the development of this study.

1. J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth. Non-linear dynamics characterization from wavelet packet transform for automatic recognition of emotional speech. Smart Innovation, Systems and Technologies, 48 pp. 199–207, 2016.

2. J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, L. D. Avendaño, and E. Nöth. Time dependent ARMA for automatic recognition of fear-type emotions in speech. Lecture Notes in Artificial Intelligence, 9302, pp. 110–118, 2015.

3. J.C. Vásquez-Correa, T. Arias-Vergara, J.R. Orozco-Arroyave, J.F. Vargas-Bonilla, J.D. Arias-Londoño and E. Nöth. Automatic Detection of Parkinson's Disease from Continuous Speech Recorded in Non-Controlled Noise Conditions. Proceedings of the 16th Anual conference of the international speech and communication association (INTERSPEECH), Dresden, 2015.

4. J. C. Vásquez-Correa, N. García, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth. Emotion recognition from speech under environmental noise conditions using wavelet decomposition. In Proceedings of the 49th IEEE International Carnahan Conference on Security Technology (ICCST), Taipei, 2015.

5. N. García, J.C. Vásquez-Correa, J.F. Vargas-Bonilla, J.R. Orozco-Arroyave, J.D. Arias-Londoño. Automatic Emotion Recognition in Compressed Speech Using Acoustic and Non-Linear Features. Proceedings of the 20th Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), Bogotá, 2015.

6. J. C. Vásquez-Correa, N. García, J. F. Vargas-Bonilla, J. R. Orozco-Arroyave, J. D. Arias-Londoño, and O. L. Quintero-Montoya. Evaluation of wavelet measures on automatic detection of emotion in noisy and telephony speech signals. In Proceedings of the 48th IEEE International Carnahan Conference on Security Technology (ICCST), Rome, 2014.

7. J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla and E. Nöth. New Computer Aided Device for Real Time Analysis of Speech of People with Parkinson's Disease. Revista Facultad de Ingeniería Universidad de Antioquia, N. 72 pp. 87-103, 2014.

8. N. García, J.C. Vásquez-Correa, J.F. Vargas-Bonilla, J.R. Orozco-Arroyave, J.D. Arias-Londoño. Evaluation of the effects of speech enhancement algorithms on the detection of fundamental frequency of speech. Proceedings of the 19th Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), Armenia, 2014.

Nomenclature

Roman Symbols

a	Approximation coefficients in discrete wavelet transform
B	Back-shift operator
C	Constant for normalization
d	Detail coefficients in discrete wavelet transform
E	Expectation operator
F_0	Fundamental frequency of speech
L	Time interval for Hurst exponent
m	Embedding dimension of attractor
s	Scale parameter in wavelet transform
u	Translation parameter in wavelet transform

Greek Symbols

α, β, γ	Parameters for bionic wavelet transform
λ_1	Largest Lyapunov Exponent
ψ	Mother Wavelet function
τ	Time delay for reconstruction of attractor
Θ	Heaviside step function

Acronyms / Abbreviations

AR Auto-regressive

ARMA Auto Regressive Moving Average

BWT Bionic Wavelet Transform

CD Correlation Dimension

CS Correlation Sum

dB Decibel

DWT Discrete Wavelet Transform

FFT Fast Fourier Transform

GMM Gaussian mixture model

GNE Glottal to Noise Excitation ratio

HE Hurst exponent

HNR Harmonic to Noise Ratio

LEE Log-energy entropy

LLE largest Lyapunov exponent

LOGSO Leave one group speaker out

LOSO Leave one speaker out

LZC Lempel-Ziv complexity

MA Moving Average

MFCC Mel frequency cepstral coefficients

NLD Non-Linear Dynamics

NNE Normalized Noise Energy

SE Speech Enhancement

SSWT Synchro-squeezing Wavelet Transform

STFT Short time Fourier Transform

SVM Support vector machine

TARMA Time dependent Auto Regressive Moving Average

UAR Unweighted average recall

VAD Voice Activity Detection

WPT Wavelet Packet Transform

WT Wavelet Transform

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[2] F. Weninger, M. Wöllmer, and B. Schuller. Emotion recognition in naturalistic speech and language - a survey. In *Emotion Recognition: A Pattern Analysis Approach*, pages 237–267. 2015.

[3] P. Gupta and N. Rajput. Two-stream emotion recognition for call center monitoring. In *Proceedings of the Anual conferenece of the international speech and communication association (INTERSPEECH)*, pages 2241–2244, 2007.

[4] Nemesysco. nemesysco - Voice analysis technologies, http://www.nemesysco.com/index.html.

[5] QA5. QA5 technologies, http://www.qa5system.com/index.html.

[6] European Union. Semaine - The sensitive agent project, http://www.semaine-project.eu/.

[7] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008.

[8] L. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2011.

[9] K. E. B. Ooi, L. A. Low, M. Lech, and N. Allen. Early prediction of major depression in adolescents using glottal wave characteristics and Teager energy parameters. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4613–4616, 2012.

[10] Y. Yang, C. Fairbairn, and J. F. Cohn. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4(2):142–150, 2013.

[11] HUMAINE Project. AAAC. Emotion-research, http://emotion-research.net/wiki/Databases, 2012.

[12] C. E. Williams and K. N. Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, 1972.

[13] K. R. Scherer. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. *Proceedings of the Anual conferenece of the international speech and communication association (INTERSPEECH)*, pages 1517–1520, 2005.

[15] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[16] R Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.

[17] S. Haq and P. J. B. Jackson. Multimodal emotion recognition. *Machine audition: principles, algorithms and systems, IGI Global, Hershey*, pages 398–423, 2010.

[18] T. Bänziger, M. Mortillaro, and K. R. Scherer. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161, 2012.

[19] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, and C. Sedogbo. The SAFE corpus: illustrating extreme emotions in dynamic situations. In *Proceedings of the International conference on Language Resources and Evaluation (LREC)*, pages 76–79, 2006.

[20] J. A. Bachorowski. Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57, 1999.

[21] R. Fernandez and R. W. Picard. Modeling drivers speech under stress. *Speech Communication*, 40(1):145–159, 2003.

[22] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech communication*, 40(1):117–143, 2003.

[23] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface'05 audio-visual emotion database. In *Proceedings of International Conference on Data Engineering Workshops*, pages 8–15, 2006.

[24] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[25] L. Devillers and L. Vidrascu. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Proceedings of the Anual conferenece of the international speech and communication association (INTERSPEECH)*, 2006.

[26] C. Vaudable and L. Devillers. Negative emotions detection as an indicator of dialogs quality in call centers. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5109–5112, 2012.

[27] S. Steidl. *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Germany, 2009.

[28] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom. Getting started with SUSAS: a speech under simulated and actual stress database. In *Eurospeech*, volume 97, pages 1743–46, 1997.

[29] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8, 2013.

[30] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.

[31] C. Busso, S. Lee, and S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):582–596, 2009.

[32] B. Stasiak and K. Rychlicki-Kicior. Fundamental frequency extraction in speech emotion recognition. *Multimedia Communications, Services and Security*, 287:292–303, 2012.

[33] V. Sethu, E. Ambikairajah, and J. Epps. On the use of speech parameter contours for emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–14, 2013.

[34] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 emotion challenge. In *Proceedings of the Anual conferenece of the international speech and communication association (INTERSPEECH)*, pages 312–315, 2009.

[35] J. Přibil and A. Přibilová. Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–22, 2013.

[36] F. Eyben, M. Wöllmer, and B. Schuller. OpenEAR-introducing the munich open-source emotion and affect recognition toolkit. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction*, pages 1–6, 2009.

[37] F. Eyben, A. Batliner, and B. Schuller. Towards a standard set of acoustic features for the processing of emotion in speech. In *Proceedings of Meetings on Acoustics*, volume 9, pages 1–12, 2010.

[38] C. C. Lee, E. Mower, C. Busso, S. Lee, and Sh. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9):1162–1171, 2011.

[39] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5688–5691, 2011.

[40] W. Zheng, M. Xin, X. Wang, and B. Wang. A novel speech emotion recognition method via incomplete sparse least square regression. *IEEE Signal Processing Letters*, 21(5):569–572, 2014.

[41] R. Xia, J. Deng, B. Schuller, and Y. Liu. Modeling gender information for emotion recognition using denoising autoencoder. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 990–994, 2014.

[42] J. Deng, Z. Zhang, F. Eyben, and B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, 2014.

[43] S. Mariooryad and C. Busso. Compensating for speaker or lexical variabilities in speech for emotion recognition. *Speech Communication*, 57:–12, 2014.

[44] S. Wu, Tiago H. Falk, and W. Y. Chan. Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785, 2011.

[45] S. Haq, P. J. B. Jackson, and J. Edge. Speaker-dependent audio-visual emotion recognition. In *AVSP*, pages 53–58, 2009.

[46] M. Bejani, D. Gharavian, and N. M. Charkari. Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. *Neural Computing and Applications*, 24(2):399–412, 2014.

[47] Y. Kim, H. Lee, and E. M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3687–3691, 2013.

[48] M. C. Sezgin, B. Gunsel, and K. K. Gunes. Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1):1–21, 2012.

[49] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli. Hybrid deep neural network-hidden markov model (DNN-HMM) based speech emotion recognition. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 312–317, 2013.

[50] Y. Attabi and P. Dumouchel. Anchor models for emotion recognition from speech. *IEEE Transactions on Affective Computing*, 4(3):280–290, 2013.

[51] E. Yuncu, H. Hacihabiboglu, and C. Bozsahin. Automatic speech emotion recognition using auditory models with binary decision tree and SVM. In *Proceedings of the International Conference on Pattern Recognition*, pages 773–778, 2014.

[52] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Dominguez. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Transactions on Biomedical Engineering*, 58(2):370–379, 2011.

[53] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Díaz-de María. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1186–1195, 2009.

[54] G Vaziri, F. Almasganj, and M. S. Jenabi. On the fractal self-similarity of laryngeal pathologies detection: the estimation of Hurst parameter. In *International Conference on Information Technology and Applications in Biomedicine*, pages 383–386, 2008.

[55] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, and J. R. Orozco-Arroyave. Nonlinear dynamics characterization of emotional speech. *Neurocomputing*, 132:126–135, 2013.

[56] L. Zao, D. Cavalcante, and R. Coelho. Time-frequency feature and AMS-GMM mask for acoustic emotion classification. *IEEE Signal Processing Letters*, 21(5):620–624, 2014.

[57] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth. Non-linear dynamics characterization from wavelet packet transform for automatic recognition of emotional speech. *Smart Innovation, Systems and Technologies*, 48:199–207, 2016.

[58] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. D. O'Shaughnessy. Amplitude modulation features for emotion recognition from speech. In *Proceedings of the Anual conferenece of the international speech and communication association (INTERSPEECH)*, pages 2420–2424, 2013.

[59] A. B. Kandali, A. Routray, and T. K. Basu. Vocal emotion recognition in five native languages of assam using new wavelet features. *International Journal of Speech Technology*, 12(1):1–13, 2009.

[60] V. N Degaonkar and S. D. Apte. Emotion modeling from speech signal based on wavelet packet transform. *International Journal of Speech Technology*, 16(1):1–5, 2013.

[61] S. Ntalampiras and N. Fakotakis. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Transactions on Affective Computing*, 3(1):116–125, 2012.

[62] L. He, M. Lech, J. Zhang, X. Ren, and L. Deng. Study of wavelet packet energy entropy for emotion classification in speech and glottal signals. In *Fifth International Conference on Digital Image Processing*, 2013.

[63] K. Wang, N. An, and L. Li. Speech emotion recognition based on wavelet packet coefficient model. In *Proceedings of the International Symposium on Chinese Spoken Language Processing*, pages 478–482, 2014.

[64] Y. Huang, A. Wu, G. Zhang, and Y. Li. Speech emotion recognition based on coiflet wavelet packet cepstral coefficients. In *Pattern Recognition*, pages 436–443. 2014.

[65] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. *Speech Prosody*, pages 276–289, 2006.

[66] B. Schuller, G. Rigoll, M. Grimm, K. Kroschel, T. Moosmayr, and G. Ruske. Effects of in-car noise-conditions on the recognition of emotion within speech. *Fortschritte der Akustik*, 33(1):305–306, 2007.

[67] A. Tawari and M. Trivedi. Speech emotion analysis in noisy real-world environment. In *Proceedings of the International Conference on Pattern Recognition*, pages 4605–4608, 2010.

[68] J. Pohjalainen and P. Alku. Automatic detection of anger in telephone speech with robust auto-regressive modulation filtering. In *Proceedigns of International Conference on Acoustic, Speech, and Signal Processing (ICASSP), pages=7537–7541, year=2013*.

[69] J. Pohjalainen and P. Alku. Multi-scale modulation filtering in automatic detection of emotions in telephone speech. In *Proceedigns of International Conference on Acoustic, Speech, and Signal Processing (ICASSP), pages=980–984, year=2014,*.

[70] L. Vidrascu and L. Devillers. Detection of real-life emotions in call centers. In *Proceedings of the Anual conferenece of the international speech and communication association (INTERSPEECH)*, number 10, pages 1841–1844, 2005.

[71] I. Siegert. *Emotional and user - specific cues for improved analysis of naturalistic interactions*. PhD thesis, Magdeburg, Universität, 2015.

[72] G. Fant. *The acoustic theory of the speech production*. Hague, The Netherlands: Mounton & Co, 1960.

[73] P. Becker. Structural and relational analyses of emotions and personality traits. *Zeitschrift fur Differentielle und Diagnostische Psychologie*, 22(3):155–172, 2001.

[74] R. R. Cornelius. *The science of emotion: Research and tradition in the psychology of emotions*. Prentice-Hall, Inc, 1996.

[75] T. C. Schneirla. An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. pages 1–42. 1959.

[76] H. Schlosberg. Three dimensions of emotion. *Psychological review*, 61(2):81–88, 1954.

[77] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 19–24, 2000.

[78] R. W. Picard. *Affective computing*. MIT press, 2000.

[79] S. Bedoya-Jaramillo, J. R. Orozco-Arroyave, J. D. Arias-Londoño, and J. F. Vargas-Bonilla. Emotion recognition from telephony speech using acoustic and nonlinear features. In *Proceedings of the International Carnahan Conference on Security Technology*, pages 1–5, 2013.

[80] J. C. Vasquez-Correa, N. Garcia, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth. Emotion recognition from speech under environmental noise conditions using wavelet decomposition. In *Proceedings of the International Carnahan Conference on Security Technology (ICCST)*, pages 1–5, 2015.

[81] H. Bořil, P. Boyraz, and J. H. Hansen. Towards multimodal driver's stress detection. In *Digital Signal Processing for in-Vehicle Systems and Safety*, pages 3–19. 2012.

[82] C. E. Williams and K. N. Stevens. Vocal correlates of emotional states. *Speech Evaluation in Psychiatry*, pages 221–240, 1981.

[83] I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[84] G. Fairbanks and W. Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Communications Monographs*, 6(1):87–104, 1939.

[85] I. Fónagy and K. Magdics. Emotional patterns in intonation and music. *STUF-Language Typology and Universals*, 16(1-4):293–326, 1963.

[86] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.

[87] P. Boersma and D. Weenik. PRAAT: a system for doing phonetics by computer. Report of the Institute of Phonetic Sciences of the University of Amsterdam., 1996.

[88] V. Abolhasanizadeh, E. Ayazi, A. Sharifi M., and H. Karimabadi. The effect of age and sex on the acoustic characteristics of speech. In *Language Design: Journal of Theoretical and Experimental Linguistics*, volume 16, pages 105–116, 2014.

[89] M. Babel and D. Bulatov. The role of fundamental frequency in phonetic accommodation. *Language and Speech*, 55(2):231–248, 2012.

[90] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman. Stress and emotion classification using jitter and shimmer features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, 2007.

[91] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[92] M. Westphal. The use of cepstral means in conversational speech recognition. In *Proceedings of Eurospeech*, 1997.

[93] E. Zwicker. Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33:248, 1961.

[94] T. Jehan. *Creating music by listening*. PhD thesis, Massachusetts Institute of Technology, 2005.

[95] C. Gobl and A. N. Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1):189–212, 2003.

[96] E. Yumoto, W. J. Gould, and T. Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71(6):1544–1550, 1982.

[97] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara. Normalized noise energy as an acoustic measure to evaluate pathologic voice. *The Journal of the Acoustical Society of America*, 80(5):1329–1334, 1986.

[98] D. Michaelis, T. Gramss, and H. W. Strube. Glottal-to-noise excitation ratio–a new measure for describing pathological voices. *Acta Acustica united with Acustica*, 83(4):700–706, 1997.

[99] F. Takens. On the numerical determination of the dimension of an attractor. *Lecture Notes in Mathematics*, 1125:99–106, 1985.

[100] M. B. Kennel, R. Brown, and H. D. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3404–3411, 1992.

[101] P. Grassberger and I. Procaccia. Characterization of strange attractors. *Physical Review Letters*, 50(5):346, 1983.

[102] H. E. Hurst, R. P. Black, and Y. Simaika. *Long-term storage: an experimental study*. Constable, 1965.

[103] A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22(1):75–81, 1976.

[104] C. T. Ishi, H. Ishiguro, and N. Hagita. Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. *EURASIP J. Audio, Speech and Music Processing*, 2010, 2010.

[105] K. Funaki. A time-varying complex AR speech analysis based on GLS and ELS method. In *Eurospeech*, pages 1–4, 2001.

[106] A.G. Poulimenos and S.D. Fassois. Parametric time-domain methods for non-stationary random vibration modelling and analysis — a critical survey and comparison. *Mechanical Systems and Signal Processing*, 20(4):763 – 816, 2006.

[107] G. N. Fouskitakis and S. D. Fassois. Functional series TARMA modelling and simulation of earthquake ground motion. *Earthquake Engineering & Structural Dynamics*, 31(2):399–420, 2002.

[108] L. D. Avendaño Valencia and S. D. Fassois. Generalized stochastic Constraint TARMA models for in-operation identification of wind turbine non-stationary dynamics. *Key Engineering Materials*, 569:587– 594, 2013.

[109] D. Rudoy, T. F. Quatieri, and P. J. Wolfe. Time-varying autoregressive tests for multi-scale speech analysis. In *Proceedings of the Anual conferenece of the international speech and communication association (INTERSPEECH)*, pages 2839–2842, 2009.

[110] S. Mallat. *A wavelet tour of signal processing*. Academic press, 2nd edition, 1999.

[111] S. H. Chen and J. F. Wang. Speech enhancement using perceptual wavelet packet decomposition and teager energy operator. In *Real World Speech Processing*, pages 51–65. 2004.

[112] J. Yao and Y. T. Zhang. Bionic wavelet transform: a new time-frequency method based on an auditory model. *IEEE Transactions on Biomedical Engineering* , 48(8):856–863, 2001.

[113] S. M. Govindan, P. Duraisamy, and X. Yuan. Adaptive wavelet shrinkage for noise robust speaker recognition. *Digital Signal Processing*, 33:180–190, 2014.

[114] I. Daubechies and S. Maes. A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models. *Wavelets in medicine and biology*, pages 527–546, 1996.

[115] I. Daubechies, J. Lu, and H. T. Wu. Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool. *Applied and Computational Harmonic Analysis*, 30(2):243–261, 2011.

[116] G. Thakur, E. Brevdo, N. S. Fučkar, and H. T. Wu. The synchrosqueezing algorithm for time-varying spectral analysis: robustness properties and new paleoclimate applications. *Signal Processing*, 93(5):1079–1094, 2013.

[117] AMR speech Codec; General description (3GPP TS 26.071 version 11.0.0 Release 11), October 2012.

[118] Libav, open source audio and video processing tools.

[119] Adaptive Multi-Rate Wideband (AMR-WB) speech Codec; Transcoding functions (3GPP TS 26.190 version 12.0.0 Release 12), October 2014.

[120] Chris Bagwell. Sox - sound exchange, the swiss army knife of audio manipulation, 2013.

[121] International Telecommunication Union (ITU). 7 kHz audio-coding within 64 kbit/s. Recommendation ITU-T G.722, 2012.

[122] International Telecommunication Union (ITU). 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM). Recommendation ITU-T G.726, 1990.

[123] Internet Engineering Task Force (IETF). SILK Speech Codec, 2010.

[124] Internet Engineering Task Force (IETF). Definition of the Opus Audio Codec. RFC 6716, 2012.

[125] Google Inc. "WebRTC." Internet: http://www.webrtc.org.

[126] C. H. Lin, W. K. Liao, W. C. Hsieh, J. Liao, W, and J. C. Wang. Emotion identification using extremely low frequency components of speech feature contours. *The Scientific World Journal*, 2014, 2014.

[127] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):443–445, 1985.

[128] Y. Hu and P. C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Transactions on Speech and Audio Processing*, 11(4):334–341, 2003.

[129] B2030A Studio Monitor: Specifications. http://www.music-group.com/Categories/Behringer/Loudspeaker-Systems/Studio-Monitors/B2030A/p/P0135/Features.

[130] SM63 Shure Microphone. http://es.shure.com/americas/products/microphones/sm/sm63-handheld-microphone,.

[131] Fast Track C400, http://avid.force.com/pkb/articles/download/Fast-Track-C400-Drivers.

[132] Internet Engineering Task Force (IETF). SILK Speech Codec, 2010.

[133] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, L. D. Avendaño, and Elmar Nöth. Time dependent ARMA for automatic recognition of fear-type emotions in speech. *Lecture Notes in Artificial Intelligence*, 9302:110–118, 2015.

[134] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, and S. Narayanan. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 2015.

[135] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 3(4):251–266, 1995.

[136] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[137] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.