



University of Crete
Department of Computer Science



Institute of Computer Science

Expressive Speech Analysis and Classification using adaptive Sinusoidal Modeling

M.Sc. Thesis

Theodora Yakoumaki

Heraklion, April 2015

Expressive Speech Analysis and Classification using adaptive Sinusoidal Modeling

Submitted by

Theodora Yakoumaki

in partial fulfilment of the requirements for the
Master of Science degree in Computer Science

Author:

Theodora Yakoumaki
Department of Computer Science

Examination Committee:

Supervisor

Yannis Stylianou, Professor, University of Crete

Member

Athanasios Mouchtaris, Assistant Professor, University of Crete

Member

Margarita Kotti, Researcher, Imperial College London, UK

Departmental Approval:

Chairman of the
Graduates Studies Committee

Antonis Argyros, Professor, University of Crete

Heraklion, April 2015

Acknowledgements

First of all, I would like to thank my supervisor, Professor Yannis Stylianou, for giving me the opportunity to become a member of his team, for his guidance and trust that he showed during the time we worked together.

I am also specially grateful to have worked with Dr. George Kafentzis. George thank you so much for your continuous support and guidance. It was great pleasure to work with you and learn from you.

Special thanks also to Professor Athanasios Mouchtaris, member of my dissertation committee, for the cooperation we have had for half a year. Also, special thanks to Ms Margarita Kotti for being member of my dissertation committee and for her advices in expressive speech databases in ISCA Summer School.

I would also like to thank all my colleagues at the Multimedia Informatics Laboratory. My warmest thanks to Olina Simantiraki, Sofia Yanikaki and Myros Apostolakis for sharing all these hours inside the lab. A special thank to Maria Koutsogianaki for being so helpful and encouraging. I would also like to thank Tasos for the presentation of my paper in EUSIPCO.

A huge thank to my dearest friend and also colleague at the lab Veronica Morfi for all the moments that we have passed through these 7 years of our studies. Vero, i wish your dreams come true! I would also like to thank my closest friends Amalia, Anna, Valentina, Konstantina, Christina and Fofi for their love and trust all these years. Special thanks also to my beloved friends Mina, Makis and Haris for being there for me providing me encourage, advise and laugh!

Last but not least, the greatest “thank you” to my family: my parents, Yiannis and Eutuxia, for their love, trust, encourage and patient during the years of my studies. My brother Manolis for showing so much patient being in the same house with me all these years. My grandmothers Theodora and Argyro, and my closest relatives Giannis, Eleni, Marianna, Giorgos and May-britt. At last, i would like to thank my famous dog, Fourier, for his unconditional love and obedience.

Thank you all!

Abstract

Emotional (or stressed/expressive) speech can be defined as the speech style produced by an emotionally charged speaker. Speakers that feel *sad*, *angry*, *happy* and *neutral* put a certain stress in their speech that is typically characterized as emotional. Processing of emotional speech is assumed among the most challenging speech styles for modelling, recognition, and classifications. The emotional condition of speakers may be revealed by the analysis of their speech, and such knowledge could be effective in emergency conditions, health care applications, and as pre-processing step in recognition and classification systems, among others.

Acoustic analysis of speech produced under different emotional conditions reveals a great number of speech characteristics that vary according to the emotional state of the speaker. Therefore these characteristics could be used to identify and/or classify different emotional speech styles. There is little research on the parameters of the Sinusoidal Model (SM), namely amplitude, frequency, and phase as features to separate different speaking styles. However, the estimation of these parameters is subjected to an important constraint; they are derived under the assumption of local stationarity, that is, the speech signal is assumed to be stationary inside the analysis window. Nonetheless, speaking styles described as *fast* or *angry* may not hold this assumption. Recently, this problem has been handled by the adaptive Sinusoidal Models (aSMs), by projecting the signal onto a set of amplitude and frequency varying basis functions inside the analysis window. Hence, sinusoidal parameters are more accurately estimated.

In this thesis, we propose the use of an adaptive Sinusoidal Model (aSM), the extended adaptive Quasi-Harmonic Model (eaQHM), for emotional speech analysis and classification. The eaQHM adapts the amplitude and the phase of the basis functions to the local characteristics of the signal. Firstly, the eaQHM is employed to analyze emotional speech in accurate, robust, continuous, time-varying parameters (amplitude and frequency). It is shown that these parameters can adequately and accurately represent emotional speech content. Using a well known database

of pre-labeled narrowband expressive speech (SUSAS) and the emotional database of Berlin, we show that very high Signal to Reconstruction Error Ratio (SRER) values can be obtained, compared to the standard Sinusoidal Model (SM). Specifically, eaQHM outperforms SM in average by 100% in SRER. Additionally, formal listening tests, on a wideband custom emotional speech database of running speech, show that eaQHM outperforms SM from a perceptual resynthesis quality point of view. The parameters obtained from the eaQHM models can represent more accurately an emotional speech signal. We propose the use of these parameters in an application based on emotional speech, the classification of emotional speech. Using the SUSAS and Berlin databases we develop two separate Vector Quantizers (VQs) for the classification, one for amplitude and one for frequency features. Finally, we suggest a combined amplitude-frequency classification scheme. Experiments show that both single and combined classification schemes achieve higher performance when the features are obtained from eaQHM.

Περίληψη

Η εκφραστική (ή αγχωμένη/ συναισθηματική) ομιλία μπορεί να ορισθεί ως το είδος ομιλίας το οποίο παράγεται από έναν ομιλητή ο οποίος είναι συναισθηματικά φορτισμένος. Ομιλητές οι οποίοι αισθάνονται λυπημένοι, θυμωμένοι, χαρούμενοι ή ουδέτεροι προσθέτουν ένα συγκεκριμένο βάρος στην ομιλία τους, το οποίο συνήθως χαρακτηρίζεται ως συναίσθημα. Η επεξεργασία της εκφραστικής ομιλίας θεωρείται μια από τις πιο απαιτητικές διεργασίες για μοντελοποίηση, αναγνώριση και ταξινόμηση συναισθήματος. Η συναισθηματική κατάσταση ενός ομιλητή μπορεί να αποκαλυφθεί από την ανάλυση της ομιλίας του, και μια τέτοιου είδους γνώση θα ήταν χρήσιμη σε καταστάσεις εκτάκτου ανάγκης, σε εφαρμογές υγείας, καθώς και μεταξύ άλλων ως ένα στάδιο επεξεργασίας σε συστήματα αναγνώρισης και ταξινόμησης του συναισθήματος.

Η ακουστική ανάλυση της ομιλίας η οποία παράγεται κάτω από διάφορες συναισθηματικές καταστάσεις αποκαλύπτει έναν εξαιρετικά μεγάλο αριθμό χαρακτηριστικών τα οποία ποικίλουν ανάλογα με τον είδος της συναισθηματικής κατάστασης του ομιλητή. Ως εκ τούτου αυτά τα χαρακτηριστικά θα μπορούσαν να χρησιμοποιηθούν για αναγνώριση και/ή ταξινόμηση διαφόρων συναισθηματικών καταστάσεων. Υπάρχει πολύ μικρή έρευνα πάνω στις παραμέτρους του Ημιτονοειδούς Μοντέλου (SM), (οι οποίες είναι το πλάτος, η συχνότητα και η φάση) ως γνωρίσματα για τον διαχωρισμό των ειδών ομιλίας. Ωστόσο, η εκτίμηση αυτών των παραμέτρων υπόκειται σε έναν πολύ σημαντικό περιορισμό: εξάγονται με την παραδοχή της 'τοπικής στασιμότητας', ότι δηλαδή το σήμα φωνής θεωρείται στάσιμο μέσα σε ένα παράθυρο ανάλυσης. Όμως, είδη ομιλίας τα οποία χαρακτηρίζονται ως γρήγορα ή θυμωμένα ίσως να μην συμφωνούν με αυτή την παραδοχή. Προσφάτως, αυτό το πρόβλημα το χειρίζονται με επιτυχία τα προσαρμόσιμα Ημιτονοειδή Μοντέλα (aSMs), προβάλλοντας το σήμα επάνω σε ένα σύνολο συναρτήσεων βάσης μεταβλητής συχνότητας και πλάτους μέσα σε ένα παράθυρο

ανάλυσης. Ως εκ τούτου, οι ημιτονοειδείς παράμετροι εκτιμούνται με περισσότερη ακρίβεια σε σχέση με τα συνήθη ημιτονοειδή μοντέλα.

Σε αυτή την εργασία, προτείνουμε την χρήση ενός προσαρμόσιμου Ημιτονοειδούς Μοντέλου (aSM), το εκτεταμένο προσαρμόσιμο Σχεδόν - Αρμονικό Μοντέλο (eaQHM), για ανάλυση και ταξινόμηση συναισθηματικής ομιλίας. Το (eaQHM) προσαρμόζει το πλάτος και την φάση των συναρτήσεων βάσης στα τοπικά χαρακτηριστικά του σήματος. Αρχικά, το (eaQHM) καλείται να αναλύσει την εκφραστική ομιλία με πιο ακριβείς, αξιόπιστες, συνεχόμενες, χρονικά - μεταβαλλόμενες παραμέτρους (πλάτη και συχνότητες). Αποδεικνύεται ότι οι παράμετροι αυτοί μπορούν να αναπαραστήσουν το εκφραστικό περιεχόμενο της ομιλίας με επάρκεια και ακρίβεια σε σχέση με τα συνήθη ημιτονοειδή μοντέλα. Χρησιμοποιώντας μια πολύ διαδεδομένη βάση δεδομένων προ-επισημασμένης στενής ζώνης εκφραστικής ομιλίας (SUSAS) και την εκφραστική βάση δεδομένων του Βερολίνου (EmoDB), δείχνουμε ότι μπορούμε να επιτύχουμε πολύ υψηλή αναλογία σφάλματος σήματος ως προς το σφάλμα ανακατασκευής (SRER), σε σύγκριση με το κλασσικό Ημιτονοειδές Μοντέλο (SM). Συγκεκριμένα, το (eaQHM) ξεπερνά το (SM) κατά 100% μέσο όρο (SRER). Επιπλέον, έγιναν επίσημα ακουστικά τέστ, σε μια δεύτερη ευρείας ζώνης βάση δεδομένων με ομιλία, τα οποία δείχνουν ότι το (eaQHM) ξεπερνά το (SM) σε ότι αφορά την ποιότητα ανακατασκευής. Οι παράμετροι οι οποίοι μας παρέχει το (eaQHM) μοντέλο μπορούν να αναπαραστήσουν με ακρίβεια ένα σήμα εκφραστικής ομιλίας. Προτείνουμε την χρήση αυτών των παραμέτρων σε μια εφαρμογή που βασίζεται στην εκφραστική ομιλία, στην ταξινόμηση της εκφραστικής ομιλίας. Χρησιμοποιώντας τις βάσεις δεδομένων της (SUSAS) και (EmoDB) για την κατασκευή δύο χωριστών Διανυσματικών Κβαντιστών (VQ) για ταξινόμηση, ένα για τα πλάτη και ένα για τις συχνότητες ως γνωρίσματα. Τέλος, προτείνουμε ένα συνδυαστικό σχήμα ταξινόμησης με πλάτη και συχνότητες. Τα αποτελέσματα δείχνουν ότι τόσο για τα απλά γνωρίσματα όσο και για τα συνδυαστικά επιτυγχάνεται καλύτερη απόδοση χρησιμοποιώντας το (eaQHM) αντί του (SM)

Contents

Title	1
Acknowledgements	5
Abstract	7
Περίληψη	9
List of Tables	13
List of Figures	15
1 General Introduction	17
1.1 The Mechanism of Human Speech Production	17
1.2 Modeling Speech	19
1.2.1 The Source-Filter Model	19
1.2.2 The Sinusoidal Models	19
1.3 Characteristics of Emotion	19
1.4 Thesis Subject	20
1.5 Thesis Contribution	21
1.6 Thesis Organization	21
2 Related Work	23
2.1 Speech Feature Selection for Emotion Analysis and Recognition	23
2.1.1 Prosodic Features	23
2.1.2 Spectral Features	24
2.1.3 Combination of Spectral and Prosodic Features	24
2.1.4 Sinusoidal Features	25
2.2 Classification Approaches	25
2.2.1 Hidden Markov Models (HMM)	25
2.2.2 Neural Networks (NN)	26
2.2.3 Gaussian Mixture Models (GMM)	26
2.2.4 Support Vector Machines (SVM)	27
2.2.5 Vector Quantizer (VQ)	27
2.3 Conclusions and Discussion	28
2.3.1 The Purpose of This Thesis	28

3	Expressive Speech Analysis using aSMs	31
3.1	Description of the Extended Adaptive Quasi-Harmonic Model	31
3.1.1	Analysis - Initialization	32
3.1.2	Analysis - Adaptation	32
3.1.3	Synthesis	33
3.2	Database Description	34
3.2.1	SUSAS Database	34
3.2.2	Toshiba Database	35
3.2.3	Berlin Database	35
3.3	Analysis and Evaluation	36
3.3.1	Objective Evaluation	36
3.3.2	Subjective Evaluation	38
4	Expressive Speech Classification using Sinusoidal Features	43
4.1	Motivation	43
4.2	VQ based Emotion Classification	45
4.3	Classification	46
4.3.1	Single Feature	46
4.3.2	Combined Features	47
4.4	Classification Results	47
4.4.1	SUSAS	47
4.4.2	Berlin	50
4.4.3	Compared to the state of the art MFCC-based Classification	56
5	Conclusions and Future Work	59
5.1	Overview	59
5.2	Future Research Directions	59
A	Publications	63
	Bibliography	65

List of Tables

1.1	<i>Characteristics of four emotions(Joy, Anger, Sadness and Fear) [78].</i>	20
3.1	<i>The summary of the 35-word vocabulary used in SUSAS. [39]</i>	34
3.2	<i>A sample of Toshiba’s database utterances.</i>	35
3.3	<i>Content of the 10 sentences in Emo DB</i>	36
3.4	<i>Signal to Reconstruction Error Ratio values (dB) for both models on the SUSAS database. Mean and Standard Deviation are given.</i>	37
3.5	<i>Signal to Reconstruction Error Ratio values (dB) for both models on Toshiba database. Mean and Standard Deviation are given.</i>	37
3.6	<i>Signal to Reconstruction Error Ratio values (dB) for both models on Berlin speech database. Mean and Standard Deviation are given.</i>	37
4.1	<i>eaQHM and SM based Confusion Table based on amplitudes for a 64-bit VQ classification between 4 emotions of the SUSAS database. SM classification scores are in parenthesis.</i>	48
4.2	<i>eaQHM and SM based Confusion Table based on amplitudes for a 128-bit VQ classification between 4 emotions of the SUSAS database. SM classification scores are in parenthesis.</i>	48
4.3	<i>eaQHM and SM based Confusion Table based on frequencies for a 64-bit VQ classification between 4 emotions of the SUSAS database. SM classification scores are in parenthesis.</i>	48
4.4	<i>eaQHM and SM based Confusion Table based on frequencies for a 128-bit VQ classification between 4 emotions of the SUSAS database. SM classification scores are in parenthesis.</i>	48
4.5	<i>eaQHM-based Confusion Table based on amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of the normalized SUSAS database.</i>	49
4.6	<i>SM-based Confusion Table based on amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of the normalized SUSAS database.</i>	49
4.7	<i>MFCC based Table for a 128-bit VQ classification between the emotions of the Berlin database.</i>	57
4.8	<i>MFCC based Table for a 128-bit VQ classification between 4 emotions of the Berlin database.</i>	57

List of Figures

1.1	<i>Anatomy of the human speech production system.</i>	18
3.1	<i>Block diagram of the eaQHM system. Dashed line includes the initialization(harmonic) part. Dot-dashed line includes the adaptation part.</i>	33
3.2	<i>Upper part: the reconstruction of word “BREAK” from SUSAS database. Lower part: the reconstruction of the utterance “Albania is an unfortunate country” from Toshiba database. Both reconstructions compare the original to eaQHM and SM reconstruction.</i>	38
3.3	<i>A sample utterance from database of Berlin (angry speech style). In the upper panel is presented the original signal. In the middle and lower panel the reconstruction error using the SM and eaQHM respectively.</i>	39
3.4	<i>Impairment evaluation of the resynthesis quality, for Toshiba acted speech database with the 95% confidence intervals.</i>	40
4.1	<i>An example of analysis of emotional speech: First panel, neutral speech. Second panel, angry speech. Third panel, $f_0(t)$ tracks for each sample. Fourth panel, $A_0(t)$ tracks for each sample.</i>	44
4.2	<i>f_0 histogram from all the SUSAS words in four different emotions.</i>	45
4.3	<i>An example of emotional speaking styles, in time and frequency: First panel, neutral. Second panel, angry. Third panel, soft. Fourth panel, question. The word “Point” is depicted in this example.</i>	46
4.4	<i>The classification scheme based on the combination of features. A_k and f_k denote the instantaneous amplitude and frequency, and ADs denote the average distortion measures.</i>	47
4.5	<i>eaQHM-based graph based on amplitudes for a 128-bit VQ classification between all the emotions of Emo DB.</i>	50
4.6	<i>SM-based graph based on amplitudes for a 128-bit VQ classification between all the emotions of Emo DB.</i>	51
4.7	<i>eaQHM-based graph based on frequencies for a 128-bit VQ classification between all the emotions of Emo DB.</i>	52
4.8	<i>SM-based graph based on frequencies for a 128-bit VQ classification between all the emotions of Emo DB.</i>	52
4.9	<i>eaQHM-based graph based on amplitudes for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.</i>	53
4.10	<i>SM-based graph based on amplitudes for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.</i>	54
4.11	<i>eaQHM-based graph based on frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.</i>	54

4.12	SM-based graph based on frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.	55
4.13	eaQHM-based graph based on combination of amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin. . .	55
4.14	SM-based graph based on combination of amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.	56

Chapter 1

General Introduction

According to the Darwinian theory each emotion contains some physiological and psychological variations which affects the characteristics of speech. Expressive speech is produced by these variations. Emotional speech styles include angry speech, sad speech, etc. In human communication one speaker is able to determine the emotion of the other. However, in human - computer interaction there is still difficulty in accurate emotion recognition through speech. Emotion recognition techniques are trying to automatically recognize the emotional state of the speaker. The knowledge of the emotional state of the speaker could be helpful in emergency conditions [90], health care applications [5] and as a pre-processing step in recognition and classification systems.

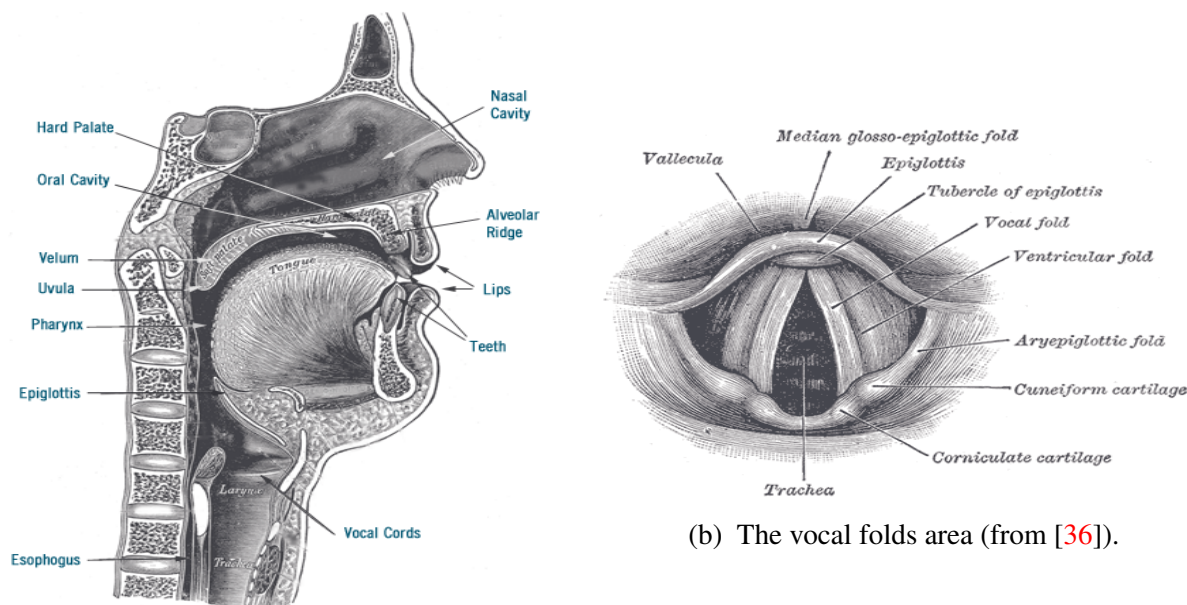
Most of the proposed techniques in emotional speech classification and analysis are based on the use of a large set of spectral and prosodic features, or a combination of them. The Cepstral and Linear Prediction coefficients are used as spectral features and the intonation, power, rhythm are used as prosodic features. Hidden Markov models (HMM) [27, 79, 16, 67, 50, 68], Neural Networks (NN), [22, 65, 6], Gaussian Mixture models (GMM) [57, 4], Support Vector machines (SVM) [34, 55] and Vector Quantization (VQ) [79, 48] are some of the proposed models for emotion classification. The results so far have shown that spectral characteristics of speech are more useful for emotion classification compared to prosodic ones. Sinusoidal models [62] are widely used in analysis [62, 79, 77], coding [63, 1], transformations [76], and synthesis [62, 77, 62] of speech, but their application on emotional speech analysis and classification is rather limited [79, 26, 96]

The goal of these thesis is to evaluate the parameters (amplitude, frequency) estimated by an adaptive sinusoidal model (aSM), called the *extended adaptive Quasi-Harmonic Model - eaQHM* [46] as features for emotional speech analysis and classification, using a well known database (SUSAS), the emotional database of Berlin and a small custom database of Toshiba. The results of amplitude, frequency and a combination of them, applied in a 128-bit Vector Quantizer (VQ), show that eaQHM can outperform the standard Sinusoidal Model in classification scores of expressive speech. In this introductory chapter, the speech production mechanism fundamentals will be discussed and a brief review of the sinusoidal models will be given.

1.1 The Mechanism of Human Speech Production

Speech is produced by a complicated mechanism which we will briefly view. At first in Figures 1.1a and 1.1b we see the basic parts and organs which conduce to speech production and are divided into three groups: the lungs, the larynx, and the vocal tract. Initially, the glottis modulates the airflow which is created by the lungs and crosses the trachea (see Figure 1.1b). The

result of this modulation is either a quasi-periodic or a signal that is called the "voice source". The phonation type is determined as this source crosses the vocal tract, which consists of three cavities: the oral, the nasal, and the pharyngeal. The naturalness of the sound is formed after the source is formed inside the vocal tract. The final waveform is transferred through the lips.



(a) The human vocal tract (from [36]).

(b) The vocal folds area (from [36]).

Figure 1.1 – Anatomy of the human speech production system.

More specifically the *voiced sounds* are formed due to the sound energy of the larynx. The airflow from the lungs is blocked for a short time by the vocal folds, which come very close to each other, thus increasing the subglottal pressure. The vocal folds re-open by the time the resistance provided by the vocal folds becomes less than the subglottal pressure. Then, the vocal folds close as a result of a combination of factors. As long as the vocal cords are supported with pressurized air from the lungs, they will continue to open and close quasi-periodically allowing pulses of air to flow through the glottal opening. The pitch of the produced sound is determined by the fundamental frequency (f_0) produced by the pulses. There is a time variation of this fundamental frequency. Variations over time can carry linguistic information which characterizes the emotional content of the speech. The oral and nasal cavities are embedded in a time-varying acoustic filter, the vocal tract. The sound energy is represented by the vocal tract in certain frequencies whereas it allowed others. The vocal tract's shape, length and volume specifies the *formants*, the frequencies with the local maximum energy. *Anti-formants* are named the frequencies where the local energy is repressed.

The larynx is also the source of energy for *unvoiced sounds*. In unvoiced consonants such as /s/ and /f/, the vocal folds can be totally open, while in phonemes like /h/ the vocal folds are in an intermediate position. The vocal cords, in *stop consonants*, such as /p/, /t/, or /k/, are moving all of a sudden from an entirely closed position where the air flow is totally cut, to a completely open position where the concentrated airflow is released and then a glottal stop is produced.

1.2 Modeling Speech

The speech production mechanism has various mathematical representations. Two views dominate for this representation, which are different but not separate: (a) in the first approach, which follows the *source-filter model*, the mathematical model represents the actual speech production mechanism as a linear, time-varying system, excited by an input signal that differs according to the type of voicing (voiced or unvoiced speech), (b) the second approach, the speech signal is presented as a time-series summation of amplitude and frequency modulated sinusoids. This approach is said to follow a *sinusoidal model*.

1.2.1 The Source-Filter Model

The source-filter theory of speech is presented in the work of Fant [30], stating that speech usually results from a combination of the larynx, which is a *source* of sound energy, modulated by a filter whose characteristics are based on the shape of the vocal tract. The outcome of this combination is a shaped spectrum with broadband energy peaks which is called the *source-filter model of speech production*.

The source-filter theory is implemented as a series of glottal pulses (*the source*) for voiced sound representing the velocity of the glottal volume. The fundamental frequency is defined by the distance of the consecutive pulses of the signal, whereas in unvoiced speech, the vocal folds do not quasi-periodically open and close, so there are no pulses and its characteristics are modelled by a zero-mean white Gaussian noise model. The lip radiation is embodied into the source in some models, since there is a high-pass like structure in the characteristics of the lip radiation spectrum. In this case, the derivative of the glottal flow is the source [31, 75] which is driven into the vocal tract filter. In digital speech processing an apparently challenging problem is to estimate the source and the vocal tract accurately, specially in voiced speech [60, 2, 28, 19, 97, 75, 18, 11, 33, 25, 91, 92, 23, 3].

1.2.2 The Sinusoidal Models

According to the source-filter theory, there is a binary glottal excitation model. Voiced speech is generated by glottal pulses and unvoiced speech seems to fit well to a random noise model. The excitation waveform of sinusoidal-based models is synthesized of arbitrary amplitudes, frequencies and phases. The major difference among sinusoidal-based representation is how the sinusoidal parameters are estimated along with the assumptions on the nature of the model [62, 37, 82, 85, 86, 35, 44, 42, 43]. The time and frequency modulation of these sinusoids as they pass through the vocal tract and are radiated by the lips produces the speech waveform, with the assumption of *local stationarity* of speech into small intervals. This means that in a short time analysis window (20 – 30 ms), speech can be modelled as sinusoids with constant amplitude and frequency values. Although this assumption is useful in practical implementations, it is not totally reliable.

1.3 Characteristics of Emotion

An emotional state is correlated with a particular physiological state that affects the speech. Emotion can be characterized in two dimensions: activation and valence [5, 32]. Activation is the needed energy to express a specific emotion. The pitch is changing depending of the emotion.

As an example, anger, fear or joy are characterized by an increment of the pitch (fundamental frequency) as it is shown in Table 1.1, along with an increment of the heart rate, blood pressure and sub-glottal pressure. Also, the mouth becomes dry and there is muscle tremor. The speech becomes loud, fast, enunciated and it is expressed with strong high-frequency energy. Articulation rate is also increased, which is calculated by measuring the length of voiced segments. On the other hand, when someone is bored or sad, the heart rate and blood pressure decrease and salivation increases. The produced speech is slow. The pitch is low and so is the high-frequency energy. The rate of articulation is decreased.

<i>Characteristics of Emotion</i>				
	Joy	Anger	Sadness	Fear
Pitch mean	High	very High	very Low	very High
Intensity mean	High	very High	Low	Medium/ High
Speaking rate	High	Low - male High - female	High - male Low - female	High

Table 1.1 – *Characteristics of four emotions (Joy, Anger, Sadness and Fear)* [78].

As follows, activation features like pitch, articulation, high-frequency energy and the quality of voice are strongly correlated with the emotion. Nevertheless, it is still difficult to identify the correct emotional state of the speaker, because some emotions are correlated. For example, both angry and happy emotions correspond to high activation like pitch and high-frequency energy but they different affect is passed on. This difference is described as the valence dimension, but there is no agreement within researchers on how, or even if, acoustic features are correlated with this valence dimension [54]. Hence, classification between high and low activation emotion is possible, but there is still challenge in classification between different emotions.

1.4 Thesis Subject

In this thesis, the main focus is on **Expressive Speech Analysis and Classification using adaptive Sinusoidal Modeling**.

Adaptive Sinusoidal Models (aSMs) have been recently developed for analysis. Their main attribute is that they deal with the assumption of *local stationarity* that the Sinusoidal Models (SMs) hold inside the analysis window for the parameters estimation. This problem is handled by projecting the signal onto a set of amplitude and frequency varying basis functions inside the analysis window [46, 72]. Initially, as suggested by Pantazis [73], the deterministic part of speech can be modeled using an adaptive quasi-harmonic model, whereas the stochastic part can be modeled as time-modulated and frequency-modulated noise. The adaptive model that is used in this thesis is called the *extended adaptive Quasi-Harmonic Model - eaQHM* and proposes both amplitude and phase adaptation [46] providing a high quality, quasi-harmonic representation of speech signals. The parameters obtained from the eaQHM model can represent more accurately an emotional speech signal. We wanted to study if these accurate parameters could benefit an application based on emotional speech. The application that was chosen is the classification of emotional speech based on the parameters estimated by the eaQHM.

In this thesis, we take advantage of the parameters estimated for the entire speech waveform by the eaQHM to perform analysis, re-synthesis and classification of emotional speech. First of all, to see if eaQHM can efficiently represent emotional speech, an emotional speech analysis

and synthesis was carried out, using three emotional speech databases, the well known *SUSAS* database, the emotional database of Berlin and a small acted database of Toshiba. The Signal to Reconstruction Error Ratio (SRER) was calculated for both SM and eaQHM showing that eaQHM outperforms standard SM by more than 100%, on average, in terms of SRER. Also, informal listening tests in the high quality database of Toshiba, showed that eaQHM-based resynthesized speech samples were indistinguishable from the original ones. After emotion speech analysis and re-synthesis, a classification task was implemented. For the classification purposes four emotions of *SUSAS* database were selected the (*angry, neutral, soft* and *question*) and six and four emotions of Berlin database. The parameters obtained from the eaQHM (amplitude and frequency) were used separately and combined, as features for a simple 128-bit VQ classifier.

1.5 Thesis Contribution

This thesis contributes the following achievements:

- The eaQHM is applied on expressive speech analysis. It is shown that eaQHM can analyze and resynthesize various styles of expressive speech with very high accuracy and quality. **This work has been accepted in EUSIPCO 2014 [47].**
- The eaQHM is applied on the problem of emotion classification and it is shown that it achieves classification scores that are higher than standard sinusoidal approaches. **This work has been accepted in Interspeech 2014 [95].**

1.6 Thesis Organization

The rest of this thesis is organized in three parts, as follows:

1. Chapter 2 : **Related Work**

This Chapter presents a literature review on speech feature selection for emotion analysis and classification problems.

2. Chapter 3 : **Expressive Speech Analysis using aSMs**

This Chapter demonstrates the use of the extended adaptive Quasi-Harmonic Model (eaQHM) in order to analyze and resynthesize emotional speech. Objective and subjective evaluation results are presented.

3. Chapter 4 : **Expressive Speech Classification using Sinusoidal Features**

This Chapter presents the results of classification tasks based on the features (amplitude and frequency) obtained from eaQHM compared to the features obtained from the standard SM. Additionally, a combined amplitude-frequency classification scheme is suggested.

4. Chapter 5 : **Conclusions and Future Work**

This Chapter concludes the thesis and proposes many interesting research directions for further investigation.

5. Appendix I presents the publications made during this thesis.

Subjective evaluations are supported with on-line demonstration pages that verify the conclusions derived.

Chapter 2

Related Work

In this chapter, we will make a literature review on speech feature selection for emotion analysis and classification problems. The chapter is divided in three sections: a) feature selection, b) classifiers, c) discussion.

2.1 Speech Feature Selection for Emotion Analysis and Recognition

We will first start with the problem of feature selection in emotion recognition. The goal is to develop robust features which carry the information related to glottal parameters. As it is suggested in [88] glottal parameters can yield beneficial information for emotion classification. If we do so, we will be able not only to correctly classify our expressive speech signal, but also to properly transform these features in order to synthesize emotional speech from neutral speech. These features should be language and speaker independent. In speech community, researchers are trying to deal with the problem of the appropriate feature selection, but still there is no clear agreement on which feature set is the best.

The most commonly used features for emotion analysis and classification are separated in two categories: prosodic and spectral features.

2.1.1 Prosodic Features

The word prosody comes from the Greek word *Προσωδία* and is used to describe the rhythm, stress and intonation of speech. Numerous features that may not be encoded by grammar or by choice of vocabulary, can be reflected by prosody, because prosody is absent in writing. Prosodic features are based on fundamental frequency and properties such as the mean, variance, energy, shimmer, jitter, F_0 contour. They are often estimated over a whole utterance and are used as long term statistics. It has shown that prosodic features are able to discriminate the high and low activation emotions, but there is confusion in discrimination of some articulation states [58].

Initially, in [45] the authors described a perception experiment with prosody features that was designed to make soft classification of emotional speech. The results were helpful for achieving more suitable acoustic patterns, when synthesizing emotional speech. The relationship between emotional states and prosody is presented in [52], showing that prosodic characteristics could be useful in emotion classification. In [89] the authors analyse the acoustic parameters in different emotional speech styles and create a classification tree, showing promising results for future work in classification and synthesis. Conversion on phoneme durations and intensity is evaluated

in [29] in order to create expressive speech, but the results of expressiveness of the converted sentences were far beyond natural emotional speech. The relationship between emotional states and prosody is investigated in [87], where the characteristics of prosody patterns were analyzed in order to show that probability of occurrence for accents and boundaries were different between 4 emotions. In [64] it was presented, for a Spanish database, the contribution of prosody to the recognisability of the uttered emotion greatly varies from one to another.

2.1.2 Spectral Features

The spectral features involve formant frequencies, along with their corresponding bandwidths. Ordinary spectral features are the Mel frequency cepstral coefficients (MFCC), the Linear prediction coefficients (LPC) [51], one-sided autocorrelation linear prediction coefficients (OS-ALPC) [9] and they are extracted in each frame with typical length 20-30 ms as a short-time representation of the speech signal.

In [93], pitch, log energy, formant, mel-band energies, and MFCCs were selected as base features for emotion recognition in SUSAS database and good results were achieved in a 4-class emotion recognition, using various classifiers. The MFCC method was used to extract features for emotion recognition in [61], with average recognition accuracy 80% between 3 different databases. A performance comparison between the short time log frequency power coefficients (LFPC) and linear prediction (LPCC) and mel-frequency cepstral coefficients (MFCC) is presented in [69]. The results of the comparison reveal that LFPC is a better choice for emotion classification than the traditional features. Additionally, in [9] the authors formulated two new feature extraction methods, a modified mel-frequency scale (M-MFCC), and an exponential-logarithmic scale (ExpoLog) in order to improve the recognition under the presence of stress. The authors in [12] proposed novel spectrally weighted mel-frequency cepstral coefficient (WMFCC) features for emotion recognition, based on the location of the formants carry information related to the emotion. The results of this study showed that classifiers with the WMFCC features perform significantly better compared to standard spectral classifiers. Acoustic models were trained with spectral features in [14], using the TIMIT corpus and the performance was tested by two emotional speech databases with accuracy up to 78% in the binary classification task.

2.1.3 Combination of Spectral and Prosodic Features

The emotional state is difficult to automatically identify and analyse using only one type of the aforementioned features. As a result, it is reasonable to use different types of features in order to achieve better classification results. Luengo, Navas and Hernandez in [55] suggested that as both kinds of features have different characteristics, they can be treated separately, creating two sub-classifiers and combining their results. The same authors also suggested that if the whole feature set is considered, spectral envelope parametrizations are more informative than the prosodic ones, using a combination of prosody features, spectral envelope and voice quality in [56] for emotion discrimination. Also, in [34] results show that a single type of feature is unable to achieve sufficient recognition accuracy, so they rely on a combination of them using the strength of excitation from zero frequency filtering method and the spectral band energy ratio related to the excitation source. In [94] the authors used modulation spectral features in combination with prosodic features for automatic emotion recognition achieving an overall recognition rate of 91.6% between seven emotion categories. A large feature set and a psychologically-inspired

binary cascade classification schema is proposed in [49] in order to separate commonly confused pairs of emotions with 87.7% best emotion recognition accuracy.

2.1.4 Sinusoidal Features

The parameters of the Sinusoidal Model(SM) have recently been engaged in analysis and/or classification of expressive speech. A new set of harmony features, correlated to the valence and activation dimension of emotion, for automatic emotion recognition from speech signals is proposed in [96]. These harmony features are based on the psychoacoustic harmony perception and improve by 4% the recognition performance, if they are used in addition to the state of art features. In [79] authors have shown that parameters obtained from SM (amplitude, frequency and phase) can be successfully used in characterization and classification of expressive speech, showing better results compared to linear prediction and cepstral features. Moreover, the authors in [26] explore the sinusoidal modeling framework for voice transformations finalized to the analysis and synthesis of emotional speech.

2.2 Classification Approaches

Emotion classification from speech signals can be considered as a supervised machine learning problem that requires training data. Training data are usually recordings or samples that are labelled with the emotion that they express. These datasets are represented as vectors of features(MFCCs, pitch values etc) which are extracted frame by frame from the sample. Thereafter the training set is given to the learning algorithm that produces a classifier. A good classifier will be able to predict the correct emotion labels for the testing samples.

Many different classification schemes have been proposed for the recognition and/or classification of emotional speech. The most common ones are the Hidden Markov Models (HMM), Neural Networks(NN), Gaussian Mixture Models, Support Vector Machines (SVM) and Vector Quantization.

2.2.1 Hidden Markov Models (HMM)

A Hidden Markov Model (HMM) is a statistical model which is assumed to be a Markov process with hidden states and can be represented as the simplest dynamic Bayesian network. In this model, the state is not precisely visible, but the output is visible depended on the state. Each state has a probability distribution over the possible output tokens. HMM gives information about the sequence of states through the sequence of tokens. The model is referred with the adjective 'hidden' owing to the passing states sequence. Due to its physical relation with the production mechanism of speech signal, the HMM has been broadly used in speech applications, such as isolated word recognition and speech segmentation. Generally, the HMM provides classification accuracy in emotion classification that is comparable to other well-known classifiers.

Mathematically, for a sequence of observable data vectors x_1, x_2, \dots, x_T a hidden Markov chain is assumed, with K the number of states, $\pi_i, i = 1, \dots, K$, the initial state probabilities and $a_{ij}, i = 1, \dots, K, j = 1, \dots, K$ the transition probability from state i to j . For the true sequence

s_1, \dots, s_T the likelihood for the observations is given by

$$\begin{aligned} p(x_1, s_1, \dots, x_T, s_T) &= \pi_{s_1} b_{s_1}(x_1) \alpha_{s_1, s_2} b_{s_2}(x_2) \dots \alpha_{s_{T-1}, s_T} b_{s_T}(x_T) \\ &= \pi_{s_1} b_{s_1}(x_1) \prod_{t=2}^T \alpha_{s_{t-1}, s_t} b_{s_t}(x_t) \end{aligned}$$

where $b_i(x_t) \equiv P(x|s_t = i)$ the observation density of the i th state.

In [68] the authors used the HMM with different number of states in order to classify emotions into six categories (anger, disgust, fear, sadness and surprise) with average accuracy of 78%. In [83] the authors chose a back-end HMM owing to its capability of modelling utterances by a sequence of vectors to capture longer-term temporal information. Ramamohan and Dandapat in [79] used a four state HMM VQ-based classifier in order to evaluate the performance of sinusoidal model features for recognition of different stress classes resulting 87% average classification rates. A continuous one-state HMM with GMM was used in [8] based on prosodic features. In [38] is described a speech emotion recognition system by use of HMM aiming at improving speech emotion recognition rate in a Chinese corpus, the results show that the method is effective, high speed and accurate.

2.2.2 Neural Networks (NN)

In computer science, artificial neural networks (ANN) are forms of computing architecture inspired by biological neural networks. The NN are used to estimate functions that depend on a large unknown number of inputs. An ANN classifier generally has plenty of design parameters as presented in [5], e.g. the form of the neuron activation function, the hidden layers number and the number of neuron in each layer. They are known to be more effective in modeling nonlinear mappings. The ANN usually have better classification performance compared to HMM and GMM, when the training sample number is low.

In [22] the authors built a NN classifier in order to recognize different emotional states and they achieved an accuracy over 90% in *hot anger* and *neutral states*. Eight sub-neural networks, one for eight emotions, compose the neural network in [65] achieving a recognition rate of approximately 50%. Both artificial neural network-based speaker identifier and the ground truth where proposed in [84] to evaluate the importance of speaker-specific information in emotion recognition. Four Feed-Forward Neural Networks are used in [66] for the classification of speech samples into four emotions, where each network has twelve input neurons and one output in the range $[0, 1]$. A Neural Network based classifier is also formulated in [40] with SUSAS database achieving classification rates from 60% to 61% for a five word vocabulary size. At last, in [74] the authors achieved the highest classification scores which was 85% for fear emotion when a NN classifier was used.

2.2.3 Gaussian Mixture Models (GMM)

The Gaussian Mixture Model (GMM) is a model of the probability of density estimation using a convex combination of multi-variate normal densities. It can be considered as an only one state HMM. The GMMs are efficient in modeling multi-modal distributions and they have less requirements in testing and training compared to HMM. Determining the best number of Gaussian components is a challenging problem.

A Gaussian Mixture Model is a weighted sum of M component Gaussian densities as given

by the equation:

$$p(x|\lambda) = \sum_{i=1}^M w_i g(x|\mu_i, \Sigma_i) \quad (2.1)$$

where x is a D -dimensional continuous-valued data vector, $w_i, i = 1, \dots, M$ are the mixture weights, and $g(x|\mu_i, \Sigma_i), i = 1, \dots, M$ are the component densities.

In [55] the authors used spectral features to train a GMM for each different emotion combining the results with a Support Vector Machine (SVM). Also, in [21] the same selected features were fed to two classifiers, a GMM and a SVM. A GMM used in [57] for the classification of spectral and prosodic features of Basque emotional speech database, gave the best result with a 98.4% accuracy when using 512 mixtures.

2.2.4 Support Vector Machines (SVM)

The Support Vector Machine (SVM) are supervised learning models that analyze data and recognize patterns. They have been used in classification and regression analysis. SVMs are based on the use of kernel functions for mapping in a non-linear way the initial features to a high-dimensional space where a linear classifier could be used. The SVM have also been used in pattern recognition applications as it is extensively presented in [15]. The SVM are used in many emotional speech recognition studies as follows:

In [93] the authors used a Gaussian SVM for the classification of SUSAS database. They achieved a 96% accuracy for stressed/neutral style classification and 70% for a 4-class speaking style classification. A SVM for 86 prosodic features classification was used in [57] achieving an overall accuracy of 93.5%, and when only six prosodic features were used, the accuracy reduced only by 1.18%. Two sub-classifiers were created in [55], one using prosodic features and the other one using spectral ones. The results were combined with a fusion system based on SVM, achieving a result of 77% for a 2-class discrimination. Instant amplitude- and frequency- derived features were fed in a binary SVM with linear kernel in [21] with 77% accuracy. In [20] the authors used the SVM classification system based on cross-correlation features, achieving an overall accuracy as high as 84.55%. A recognition accuracy of 87.7% is achieved by SVMs with linear kernels in EmoDB in [49].

2.2.5 Vector Quantizer (VQ)

Vector Quantization (VQ) is a conventional technique from signal processing, based on the principle of block coding, which allows the modeling of probability density function by the distribution of prototype vectors. VQ divides a large set of points, called vectors, into groups which have about the same number of points closest to them. Each group is represented by its centroid point, as in k-means. VQ is used for lossy data correction, pattern recognition, density estimation and clustering.

Linde Buzo Gray (LBG) Algorithm as presented in [53], is an efficient and intuitive algorithm for vector quantizers designing with quite general distortion measures. It is based either on a known probabilistic model or on a long training sequence of data. The LBG Algorithm is presented below:

1. Determine the N size of codebook.
2. Randomly select N codewords to be the initial codebook.
3. Using the Euclidean distance measure, clusterize the vectors around each codeword.
4. Compute new set of codewords.

$$y_i = \frac{1}{m} \sum_{j=1}^m x_{ij}$$

where i is the component of each vector, m is the number of vectors in the cluster.

5. Repeat (2) and (3) until the codewords don't change.

Rammamohan and Dandapat in [79] evaluated the performance of sinusoidal model features for recognition of different stress classes with a vector classifier, with average success rate of 91% for frequency features. In [48] the authors used spectral features and a 16-bit Vector Quantizer to handle input data and to identify six emotional states from the input signals. The results show that *anger* and *boredom* are clearly distinguished from other emotional categories.

2.3 Conclusions and Discussion

In this Chapter, we presented the work that has been done so far in emotional speech analysis and classification, which comes up with two important issues: the feature selection to characterize various emotions and the classification techniques that are used.

Most of the studies are still trying to find speech features and their relation to the emotional content of speech. The ability of a model to identify emotions is strongly associated with the relevant feature selection. Many classifiers have been proposed for emotional speech identifications such as HMM, GMM, ANN etc. However, the decision of the most appropriate classifier is complicated because different emotional corpora with different experimental setups were applied.

The lack of uniformity in the way methods and features are evaluated, does not allow us to make appropriate comparisons and to declare which classifier and which set of features is the best.

2.3.1 The Purpose of This Thesis

In spite of its wide range of applications [59], the Sinusoidal Model (SM) [62] has not been thoroughly engaged in analysis and/or classification of stressed speech until recently [79, 26]. In these approaches, the parameters of sinusoids (amplitude, frequency and phase) are suggested as features for classification tasks. Although, these parameters obtained from sinusoidal analysis, come up with a significant constraint; they are extracted under the assumption of *local stationarity*, that is, the speech signal is considered as *stationary* inside the analysis window. However, this is not the case for speech styles characterized as “*fast*” or “*angry*”.

Recently, the adaptive Sinusoidal Models (aSMs) [46, 70, 24] have managed to cope with this problem by projecting the signal onto a set of amplitude-and frequency- varying basis functions inside the analysis window. This way, the parameters represent the underlying signal more closely as an AM-FM decomposition. In this thesis, we propose an adaptive Sinusoidal Model (aSM), the *extended adaptive Quasi-Harmonic Model*, for emotional speech analysis and classification. The aim of this thesis is:

1. Firstly, to show that the aSMs can properly represent a speech signal with emotional content. For these purpose, emotional utterances are analyzed and resynthesized using the *eaQHM*. The performance of this task is compared to the standard SM using the Signal-to-Reconstruction-Error Ratio (SRER) values, and subjective listening tests, where applicable.
2. We study if the parameters obtained from the *eaQHM* (amplitude, frequency) have important information, which could be used in an emotion classification task.

(1) and (2) will be presented in Chapters 3 and 4 respectively.

Chapter 3

Expressive Speech Analysis using aSMs

In this chapter the extended adaptive Quasi-Harmonic Model (eaQHM) [46] is utilized to demonstrate its ability to analyze and resynthesize emotional speech. The speech corpus for the analysis and resynthesis are the SUSAS database [39], a small high quality, wideband database of acted speech of Toshiba and the emotional database of Berlin [13]. Subjective listening tests have been conducted to prove the transparency of the resynthesized speech. It is also shown that eaQHM can efficiently model all styles of emotional speech in these databases with high precision, and this is demonstrated via Signal-to-Reconstruction-Error-Ratio (SRER) values, compared to the standard SM.

3.1 Description of the Extended Adaptive Quasi-Harmonic Model

The eaQHM is a high-resolution tool to accurately estimate the parameters of an AM-FM decomposed signal. The speech signal is described as an AM-FM decomposition in the full-band (e.g. from 0 Hz to the Nyquist frequency)

$$d(t) = \sum_{k=-K}^K A_k(t) e^{j\phi_k(t)} \quad (3.1)$$

where $A_k(t)$ is the instantaneous amplitude and $\phi_k(t)$ is the instantaneous phase of k^{th} component, respectively. The instantaneous phase term is given by

$$\phi_k(t) = \phi_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^t f_k(u) du \quad (3.2)$$

where $\phi_k(t_i)$ is the instantaneous phase value at the analysis time instant t_i , f_s is the sampling frequency, and $f_k(t)$ is the instantaneous frequency of the k^{th} component. The analysis part explains how to obtain the aforementioned parameters accurately. The analysis part is divided into two steps: an initialization step, where a first approximation of the speech signal is obtained under a harmonic assumption, and an adaptation step, where the parameters of the initialization step are iteratively refined.

3.1.1 Analysis - Initialization

A *continuous* f_0 estimation for all frames, noted by \hat{f}_0 , is obtained at first, using the SWIPE pitch estimator [17] (although any pitch estimator can be used). The next step is to assume a full-band harmonicity to obtain a first estimate of the instantaneous amplitudes of all the harmonics. Using standard harmonic analysis [86], the parameters $|a_k(t_i)|$, $\phi_k(t_i)$ are obtained, where t_i is the i^{th} analysis time instant. As a final step, the overall signal can be synthesized by interpolating the $|a_k|$ and \hat{f}_0 values over successive analysis time instants t_i , resulting in an approximation of Equation 3.3.

$$\hat{d}(t) = \sum_{k=-L}^L \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \quad (3.3)$$

where

$$\hat{A}_k(t) = |a_k(t)| \quad (3.4)$$

$$\hat{\phi}_k(t_i) = \angle a_k(t_i) \quad (3.5)$$

and

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^t (k\hat{f}_0(u) + \gamma(u)) du \quad (3.6)$$

The Equations 3.4, 3.5 are estimates of $A_k(t)$, $\phi_k(t)$ and $\gamma(t)$ is a phase correction term to ensure phase coherence, as described in [72]

3.1.2 Analysis - Adaptation

In order to converge to quasi-harmonicity, the projection of the signal onto a set of amplitude and frequency varying basis functions is suggested in [46], by using the parameters a_k and b_k of the Quasi-Harmonic Model (QHM) [71]. This yields the eaQHM model, which can be formulated in a single frame as:

$$d(t) = \left(\sum_{k=-L}^L (a_k + tb_k) \left(\hat{A}_k(t) e^{j\hat{\phi}_k(t)} \right) \right) w(t) \quad (3.7)$$

with $\hat{A}_k(t)$, $\hat{\phi}_k(t)$ as in Eqs. (3.4, 3.5). In this model, a_k , b_k are the complex amplitude and the complex slope of the k^{th} component, and $\hat{A}_k(t)$, $\hat{f}_k(t)$, $\hat{\phi}_k(t)$ are estimates of the instantaneous amplitude, frequency, and phase of the k^{th} component, respectively, from the previous analysis step. The a_k , b_k parameters are obtained via Least Squares [46]. The adaptation is completed by using the frequency correction mechanism first introduced in [71]. This mechanism provides a frequency correction $\hat{\eta}_k$, for each component. Hence, at the first adaptation, for the analysis time instant t_i , the instantaneous phases become

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^t (\hat{f}_k(u) + \gamma(u)) du \quad (3.8)$$

where $\hat{f}_k(t) = kf_0(t) + \hat{\eta}_k(t)$. Then, a Least Squares solution for the a_k , b_k using these refined frequencies (and phases) leads to a better estimate of the instantaneous amplitudes $\hat{A}_k(t) =$

$|a_k(t)|$ and the $\hat{\eta}_k$ terms. By iteratively adding the $\hat{\eta}_k$ term of the current adaptation on the k^{th} -frequency track of the previous adaptation, the frequency tracks deviate from strict harmonicity and represent the underlying actual frequencies better. Finally, this adaptation scheme continues until a convergence criterion is met, which is related to the overall Signal-to-Reconstruction-Error Ratio (SRER) [73]. The SRER is defined as

$$SREER = 20 \log_{10} \frac{std(d(t))}{std(d(t) - \hat{d}(t))} \quad (3.9)$$

where $d(t)$ is the original waveform, $\hat{d}(t)$ is the model representation, and $std(\cdot)$ is the standard deviation.

3.1.3 Synthesis

During synthesis, the k^{th} instantaneous amplitude track, $\hat{A}_k(t)$, is computed via either linear or spline interpolation of the successive estimates from the last adaptation step. The k^{th} instantaneous frequency track, $\hat{f}_k(t)$, is also computed via spline interpolation. As for the k^{th} instantaneous phase track, $\hat{\phi}_k(t)$, the non-parametric approach based on the integration of instantaneous frequency is followed, as is shown in the adaptation steps of the analysis. Then, the speech signal can be approximated as:

$$\hat{d}(t) = \sum_{k=-L}^L \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \quad (3.10)$$

A block diagram of the algorithm is depicted in Figure 3.1.

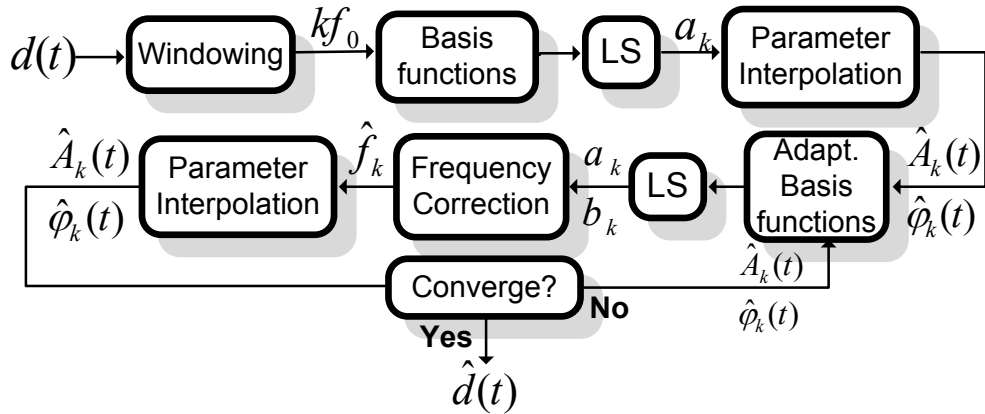


Figure 3.1 – Block diagram of the eaQHM system. Dashed line includes the initialization(harmonic) part. Dot-dashed line includes the adaptation part.

3.2 Database Description

In order to evaluate the analysis, re-synthesis and later on in classification of expressive speech signals based on the use of the eaQHM model, we selected 3 emotional speech databases. The SUSAS database, a small custom database of Toshiba and the Berlin database. Firstly, we used SUSAS database in order to evaluate the performance of eaQHM in large set of limited sentence length data. SUSAS is a pre-labelled well-known database of short communication words. Unfortunately, the SUSAS database is not proper for listening tests because of its low quality. In order to use a database of higher quality for listening tests, we selected a small wideband custom database of Toshiba, consisting of male-female acted utterances. At last, we used the Berlin database, which is a high quality database consisting of male-female utterances in German in order to extend our research results in a dataset consisting of a different spoken language than English.

3.2.1 SUSAS Database

The SUSAS (Speech Under Simulated and Actual Stress) database was used in analysis, resynthesis and classification tasks, as presented in [39]. SUSAS database was developed in the 1990s in order to study the variations of speech production and classification under stressed conditions. It consists of five stressed domains i) talking styles ii) single tracking task or speech produced in noise (Lombard effect), iii) dual tracking computer response task, iv) actual subject motion-fear tasks (G-force, Lombard effect, noise, fear), v) psychiatric analysis data (speech under depression, fear, anxiety). A total of 16,000 isolated-word utterances were produced by 32 speakers. The using vocabulary in the database consists of 35 aircraft, greatly confusable words (e.g., /go-oh-no/, /wide-white/, /six-fix/). The simulated speech under stress data consists of ten speaking styles, while actual speech stress data consists of speech during the performance of i) dual-tracking workload computer tasks, ii) subject motion-fear tasks.

35-Word SUSAS Vocabulary Set					
brake	eighty	go	nav	six	thirty
change	enter	hello	no	south	three
degree	fifty	help	oh	stand	white
destination	fix	histogram	on	steer	wide
east	freeze	hot	out	strafe	zero
eight	gain	mark	point	ten	

Table 3.1 – *The summary of the 35-word vocabulary used in SUSAS. [39]*

In our study the simulated part was used in order to evaluate our model. In this part 9 U.S. English male speakers, of three main dialects (general USA, New England/Boston, and New York City accent), under different simulated stress conditions (angry, clear, fast, Lombard, loud, neutral, question, slow, soft, and two conditions where the speaker was recorded while performing medium and light physical activities) have been recorded. The production of speech was done while the speakers were seated in a quiet environment. Each speaking style corpus has 70 speech files per speaker, which consists of isolated, short communication, two token words such as “hello”, “break”, “go”, “point” and “destination”. The sampling rate in all files was 8kHz and were represented using 16-bit integers. Given the token count for each stress condition, sums

to a about 1190 token per speaker. Due to the fact that the spoken words are acoustically similar, the database is difficult enough for several applications, speech recognition and classification.

3.2.2 Toshiba Database

The Toshiba database is a custom, small database of acted speech. This database consists of one male and one female subject, acting in four different speaking styles (*angry, sad, happy* and *neutral*), in a recording studio, with a total number of 20 waveforms sampled at 16000 Hz. The Toshiba database was only used in listening test. A sample of Toshiba's database spoken utterances is presented in Table 3.2.

Female Speaker
<i>You can change these destinations at any time</i>
<i>As long as we live we will never see another achievement like it</i>
<i>The above mentioned celebrities also wrote to the Times</i>
<i>Mark asked Tom to remember his phone</i>
Male Speaker
<i>Albania is an unfortunate country</i>
<i>You find players give them opportunities and watch them capitalize on it</i>
<i>My souffle had turned into an amorphous lump</i>
<i>Though my thumb wasn't nearly as delicious</i>

Table 3.2 – A sample of Toshiba's database utterances.

3.2.3 Berlin Database

In order to extend our analysis and classification research in another high quality database with different spoken language, we selected Emotional Database of Berlin (Emo DB) [13]. EmoDB is a labelled acted database comprising 6 basic emotions *anger, joy, sadness, fear, disgust* and *boredom* as well as *neutral* speech. Ten professional native German actors (5 female and 5 male) simulated these emotions, producing 10 utterances (5 short and 5 longer sentences), which could be used in every-day communication and are interpretable in all applied emotions. The recordings were made using a Sennheiser MKH 40 P 48 microphone and a Tascam DA P1 portable DAT-recorder in an anechoic chamber. Recordings were taken with 16 kHz sampling frequency. The recorded speech material consists of about 800 sentences (7 emotions * 10 actors * 10 sentences + some second versions). The content of the 10 sentences and their translation is depicted in Table 3.3.

<i>Emotional Database of Berlin</i>	
German Text	English Translation
Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the fridge.
Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
In sieben Stunden wird es soweit sein.	In seven hours it will be.
Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

Table 3.3 – *Content of the 10 sentences in Emo DB*

3.3 Analysis and Evaluation

In this section, the evaluation procedure is described, along with the dataset selection and the parameter estimation. For this task, the three databases mentioned above are used, (i) for testing the performance of SRER of eaQHM compared to the SRER of SM model (ii) for listening tests where only the database of Toshiba was used.

3.3.1 Objective Evaluation

At first, it is important to show that eaQHM can decompose the speech signals into AM-FM components that represent the signal closer than SM. For this, all speech files in the databases have been analyzed and resynthesized from their AM-FM components, and the corresponding SRER has been computed for each speech utterance. For this analysis, the window size was 30 ms for the SM and 3 local pitch periods for the eaQHM, both of Hamming type. A step size of 2.5 ms was selected for both models. Table 3.4 shows the mean and the standard deviation of SRER for all speakers, for most common speaking styles in SUSAS database.

SRER Performance (SUSAS)				
Model	Speaking Styles			
	Angry	Loud	Clear	Fast
SM	16.6 (3.06)	16.8 (3.01)	16.8 (3.06)	16.7 (3.03)
eaQHM	32.3 (5.61)	32.8 (5.59)	32.6 (5.62)	32.9 (5.58)
Model	Question	Soft	Neutral	Slow
	SM	16.8 (3.00)	16.7 (3.05)	16.8 (3.01)
eaQHM	32.8 (5.57)	32.9 (5.61)	32.9 (5.58)	32.9 (5.60)

Table 3.4 – *Signal to Reconstruction Error Ratio values (dB) for both models on the SUSAS database. Mean and Standard Deviation are given.*

This clearly demonstrates the quality and the performance stability of the adaptive model compared to the SM on a large database of isolated words of different expressive speaking styles. It is interesting to note that both models appear to be very stable around a mean of about 16.6 and 32.5 dB, for the SM and the eaQHM respectively. Although the distribution of SRERs is wider in eaQHM-analysis, the mean is high enough to show that in almost all cases the eaQHM manages to capture most of the information present in the speech signal, for *all* speaking styles. Since the SUSAS database contains short, low-sampled, and isolated words only, it would be interesting to observe the behaviour of the models in high-quality running expressive speech. For this, the database of Toshiba and Berlin were used. The waveforms were analyzed and the results are depicted in Table 3.5 for Toshiba database and in Table 3.6 for Berlin. The results show apparently the ability of eaQHM to properly reconstruct an emotional utterance.

SRER Performance (Toshiba Database)					
Model	Speaking Styles				
	Female Speaker				
	Angry	Happy	Neutral	Sad	
SM	14.8 (1.36)	17.5 (3.0)	16.5 (1.36)	21.2 (1.64)	
eaQHM	28.8 (1.24)	33.1 (1.81)	34.9 (2.23)	34.8 (3.60)	
Model	Male Speaker				
	SM	17.0 (1.45)	14.3 (0.76)	16.0 (1.67)	16.5 (1.63)
	eaQHM	35.7 (2.04)	31.6 (3.49)	33.3 (2.56)	33.1 (2.74)

Table 3.5 – *Signal to Reconstruction Error Ratio values (dB) for both models on Toshiba database. Mean and Standard Deviation are given.*

SRER Performance (Berlin Database)								
Model	Speaking Styles							
	Female Speaker							
	Angry	Boredom	Disgust	Fear	Happy	Sad	Neutral	
SM	11.2 (1.5)	17.2 (1.5)	14.1 (1.7)	13.5 (2.0)	12.7 (1.6)	13.1 (3.1)	15.3 (2.8)	
eaQHM	29.4 (3.9)	33.2 (3.2)	30.0 (3.6)	27.4 (3.0)	29.0 (3.0)	26.8 (4.2)	31.5 (2.5)	
Model	Male Speaker							
	SM	11.9 (1.8)	17.1 (0.89)	14.3 (1.1)	13.4 (1.3)	12.3 (1.8)	14.9 (1.4)	15.1 (1.8)
	eaQHM	32.5 (3.8)	36.6 (2.3)	30.7 (3.6)	29.7 (3.2)	30.3 (3.1)	32.4 (1.9)	35.0 (1.8)

Table 3.6 – *Signal to Reconstruction Error Ratio values (dB) for both models on Berlin speech database. Mean and Standard Deviation are given.*

It is evident that the adaptive model can handle word-isolated (i.e. SUSAS) and running expressive speech see Figure 3.2 and Figure 3.3, equally well.

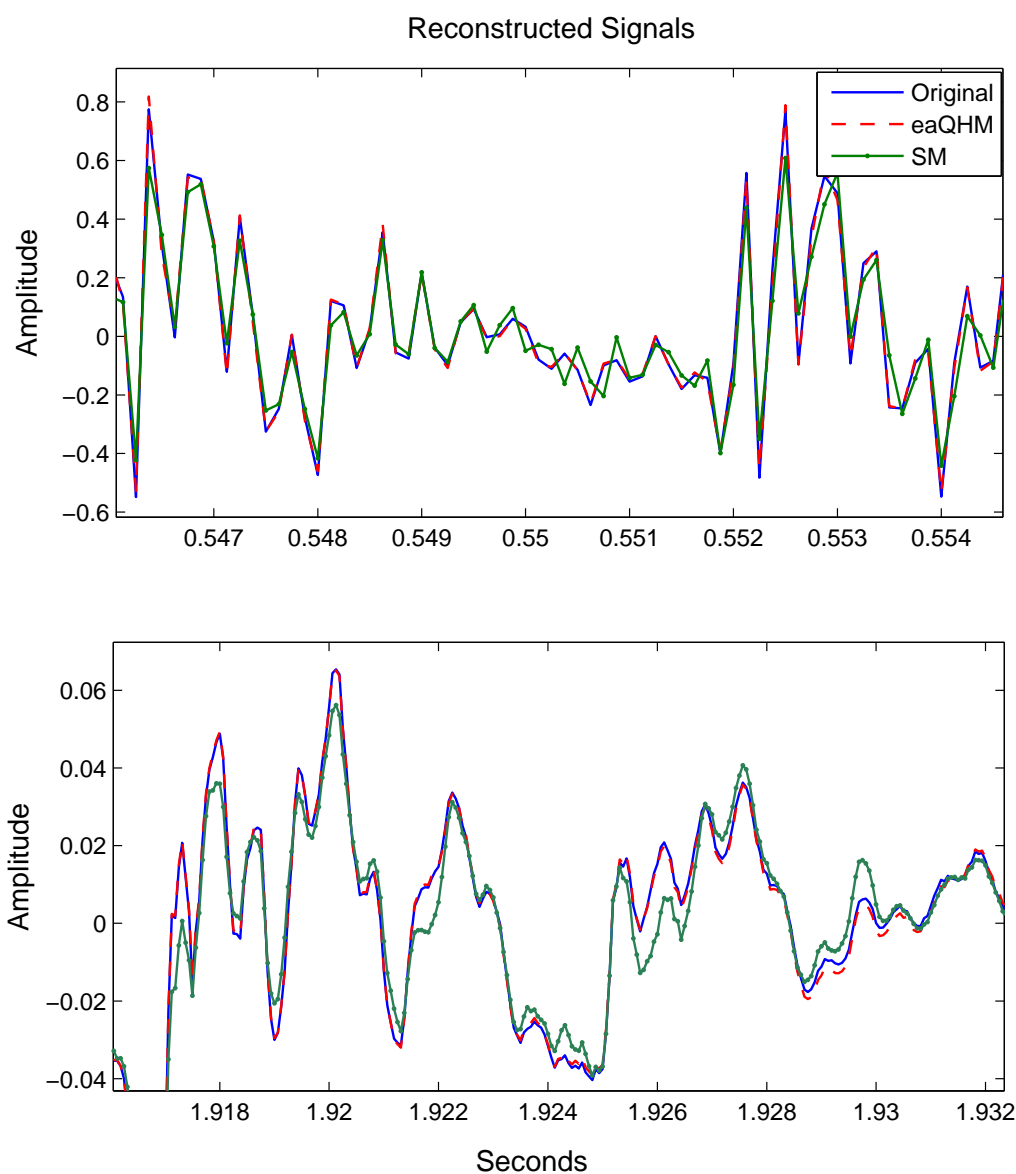


Figure 3.2 – *Upper part: the reconstruction of word “BREAK” from SUSAS database. Lower part: the reconstruction of the utterance “Albania is an unfortunate country” from Toshiba database. Both reconstructions compare the original to eaQHM and SM reconstruction.*

3.3.2 Subjective Evaluation

Towards our intention for demonstrating the perceptual differences in the resynthesis part, the SUSAS database was judged to perform poorly from a perceptual point of view due to the recording noise and the low sampling frequency. Informal listening tests showed that the eaQHM-based resynthesized speech samples were indistinguishable from the original ones, but this was the case for the standard sinusoidal model as well, in most of the cases. After careful listening, only a

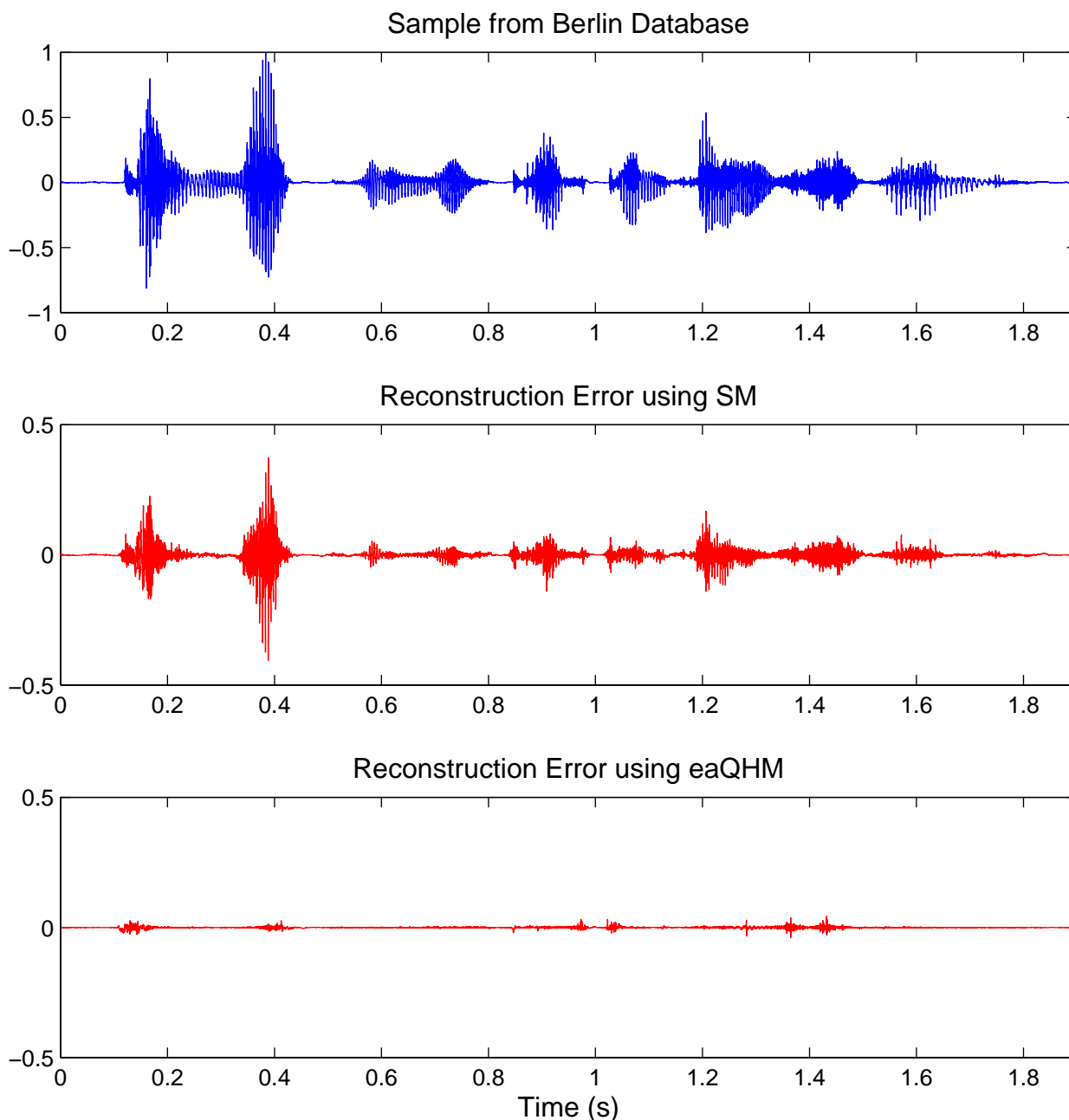


Figure 3.3 – A sample utterance from database of Berlin (angry speech style). In the upper panel is presented the original signal. In the middle and lower panel the reconstruction error using the SM and eaQHM respectively.

minority of waveforms demonstrated perceptual differences between the models but they were not enough in quantity to justify a listening test with this database. However, since the subjective evaluation is critical in synthesizing speech, especially in the case of expressiveness, a formal, on-line listening test was designed¹ using the small, high-quality database of Toshiba. The listeners were asked to evaluate the overall quality of the resynthesized speech based on the two models. A total of 32 listeners participated in this test, and the results are depicted in Figure 3.4

¹<http://www2.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.Exprtest>

along with the 95% confidence intervals. It should be noted that only 5 of them are familiar with signal processing. According to the preference test, almost all listeners noted eaQHM as being almost indistinguishable to the original one.

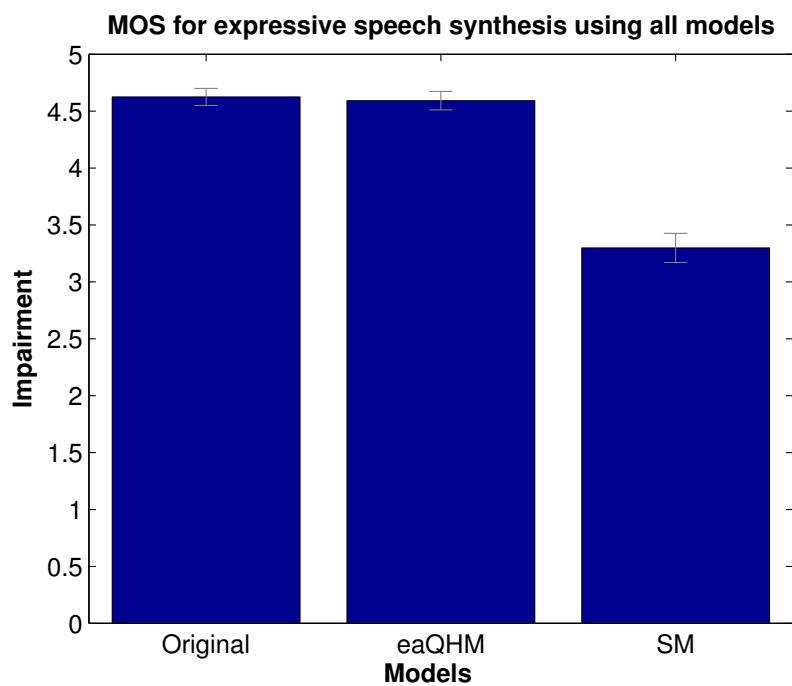


Figure 3.4 – Impairment evaluation of the resynthesis quality, for Toshiba acted speech database with the 95% confidence intervals.

Chapter 4

Expressive Speech Classification using Sinusoidal Features

In this chapter the *extended adaptive Quasi-Harmonic Model - eaQHM* is applied on emotional speech analysis for classification purposes. The parameters of the model (amplitude and frequency) are used as features for the classification. Using a well known database of narrowband expressive speech (SUSAS) and the Emotional Database of Berlin, we develop two separate Vector Quantizers (VQ) for classification, one for the amplitude and one for the frequency features. However, single feature classification is inappropriate for higher-rate classification. Thus, we suggest a combined amplitude-frequency classification scheme, where the classification scores of each VQ are weighted and ranked, and the decision is made based on the minimum value of this ranking. These classifiers are compared to the standard Sinusoidal Model (SM) classifiers which we also constructed.

4.1 Motivation

In order to use the sinusoidal features (amplitude, frequency) for emotion classification, we wanted to study how they behave. In Figure 4.1 we present the amplitude and frequency parameters of a speech sample (word “No”) from SUSAS database, pronounced with different emotional content (*angry, neutral*). Clearly, the amplitudes and frequencies of the fundamental are different, and this is the case for higher components as well. The latter is verified by the work in [79].

Additionally, the fundamental frequency obtained from eaQHM, for all speakers, from SUSAS database was studied deeper. Also, it has been described in [81] that the pitch frequency varies when the emotions are different. A study of the frequency contours could provide informations on their dependence between different emotions. For example, for the same speech uttered when different emotion is expressed, the duration of the speech utterance varies. For comparing the features between different emotions, a frequency histogram was composed for the f_0 features behaviour in each *Angry, Neutral, Question* and *Soft* speaking styles.

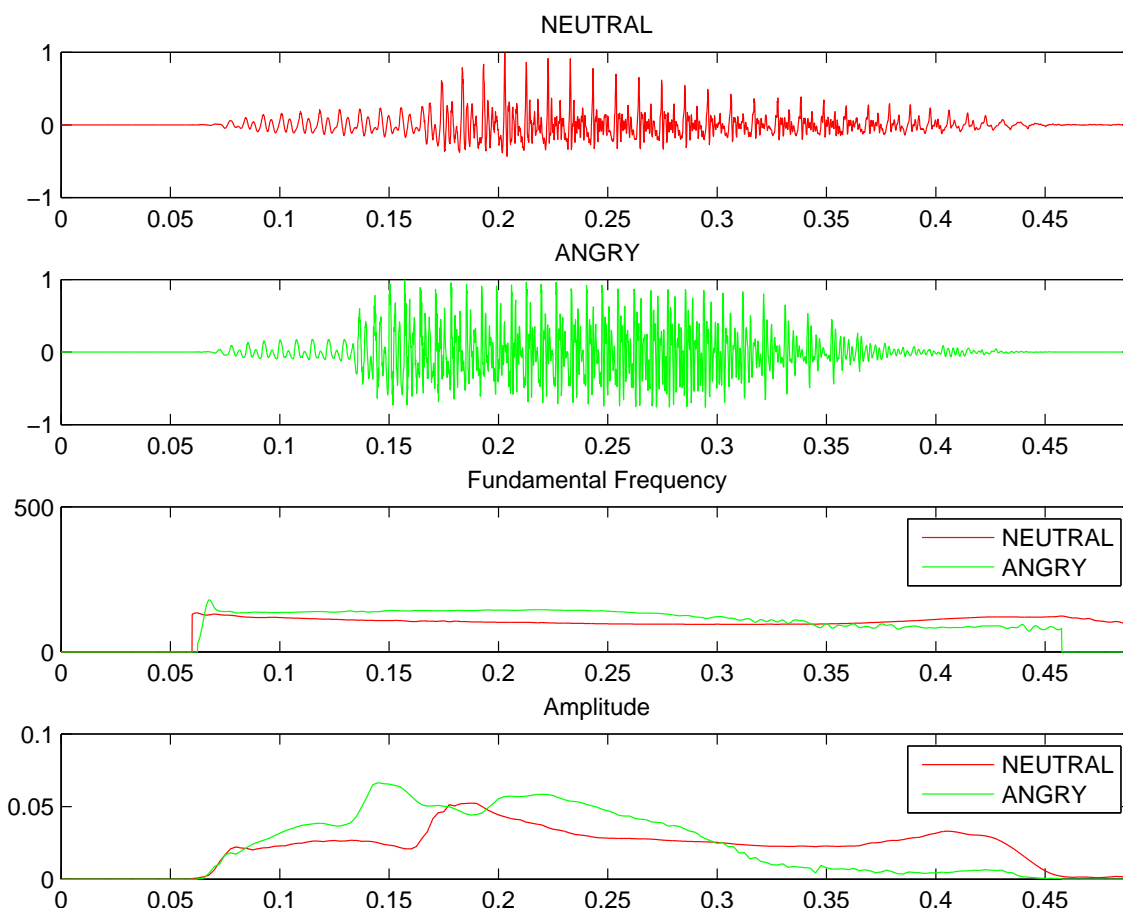


Figure 4.1 – An example of analysis of emotional speech: First panel, neutral speech. Second panel, angry speech. Third panel, $f_0(t)$ tracks for each sample. Fourth panel, $A_0(t)$ tracks for each sample.

For the histogram of f_0 , each word was separated into 12 equal intervals. For each of the four emotions, all words were analyzed using the *eaQHM*. The hamming analysis window size was set at 3 local pitch periods and the step size at 2.5 ms. The results are depicted in Figure 4.2.

The results show that the *soft* emotion takes most of its frequency values around 100 Hz, *question* and *angry* around 130 Hz and *neutral* around 150 Hz. We observe that the distributions for *angry*, *question* and *neutral* emotions are wider, because they have many f_0 values in frequencies above 150 Hz compared to the *soft* emotion.

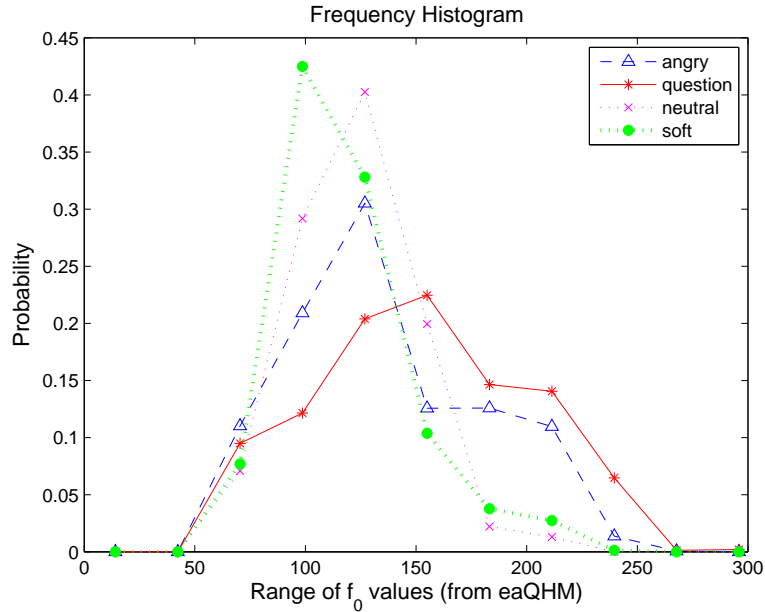


Figure 4.2 – f_0 histogram from all the SUSAS words in four different emotions.

As already discussed, a discrimination between different emotional speaking styles is of great interest. Considering a sinusoidal analysis, it has been reported that amplitude and frequency values of the sinusoidal components can be used successfully to characterize the different expressive classes (emotions) in a speech signal [79]. Since the eaQHM can compute these parameters more accurately, it is expected that their discrimination properties among different speaking styles are similar or better than those reported in the literature for the standard SM. An example of a single word (“point”) is presented in Figure 4.3 in four different speaking styles, along with the corresponding spectrograms that partly reveal their differences. It can be seen that these differences appear in amplitude strength, frequency variations, energy distributions, formant positioning, timings, duration of vowels and consonants, etc. Sinusoidal modeling can capture these differences in the form of AM-FM components [79]. Due to its adaptive processing, we propose that eaQHM can provide parameters that are highly accurate, which makes them more suitable for an emotion classification task than the same parameters obtained from a standard SM.

4.2 VQ based Emotion Classification

To evaluate our suggestion, classification tasks based on a 64-bit and a 128-bit Vector Quantizer (VQ) were designed. The speech signals from databases were separated, the 70% of them were used for training and the rest 30% were kept for testing. All discrete time waveforms were normalized to unit energy, as in

$$x[n] = \frac{x[n]}{\sqrt{\sum_{n=0}^{L-1} x^2[n]}} \quad (4.1)$$

where L is the signal length in samples. Both models used an analysis frame rate of 2.5 ms. The 10 strongest components of the magnitude spectrum of the FFT and the 10 highest sinusoidal amplitudes provided by the LS, along with their corresponding frequencies, were extracted from each analysis frame. The analysis window was set at 30 ms for the SM, and at 3 local pitch periods for the eaQHM. There was no distinction between voiced and unvoiced parts of speech.

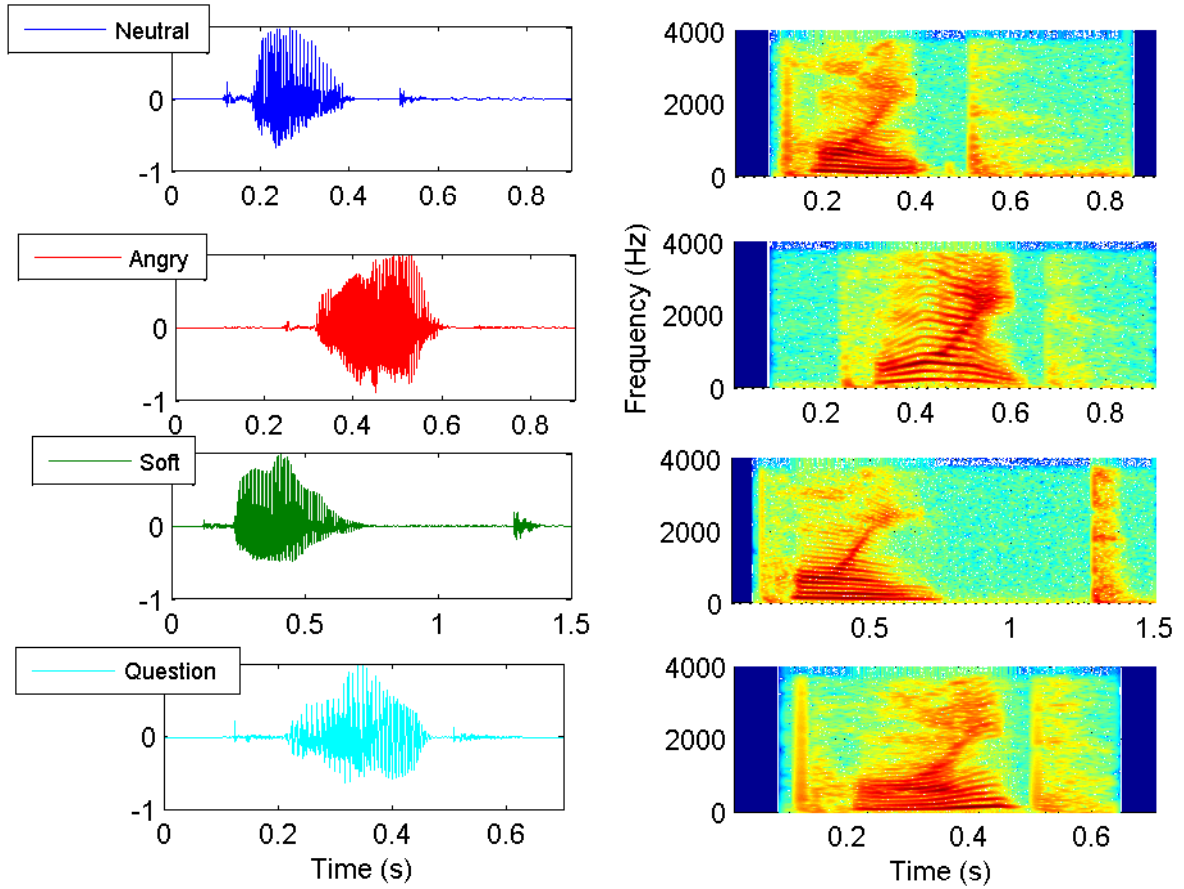


Figure 4.3 – An example of emotional speaking styles, in time and frequency: First panel, neutral. Second panel, angry. Third panel, soft. Fourth panel, question. The word “Point” is depicted in this example.

4.3 Classification

4.3.1 Single Feature

At first, two classification tasks were set, each one using different features (amplitudes and frequencies). Having M spectral vectors x_i containing the selected features (amplitudes or frequencies), the data matrix X is created as

$$X = [x_1 \ x_2 \ \dots \ x_M] \quad (4.2)$$

The codebooks are then designed based on the minimization of the Average Distortion (AD) between the training vectors and the codebook vectors in matrix Y , where

$$Y = [y_1 \ y_2 \ \dots \ y_C] \quad (4.3)$$

and C is the codebook size. The AD is defined as

$$AD = \frac{1}{C} \sum_{k=1}^C \min_{y_i \in Y} d^2(x_k, y_i) \quad (4.4)$$

where $d(x, y)$ is the Euclidean Distance (ED) between vectors x and y . For each of the four emotions mentioned above, a codebook was designed using the LBG algorithm [53]. The emotion are recognized by the minimum average distortion.

4.3.2 Combined Features

Since single-feature based classification leads to low classification scores, a combined classification scheme is suggested. The ADs obtained from amplitude and frequency based VQs are normalized by the highest corresponding AD. Then, the ADs of the corresponding emotions are added. Finally, the emotion with the minimum sum of ADs is selected as the recognized emotion. This way, when the VQs have decided differently, the VQ which is more “confident” in its decision (the minimum AD is far less than other ADs) can influence the final outcome. Figure 4.4 illustrates the proposed scheme.

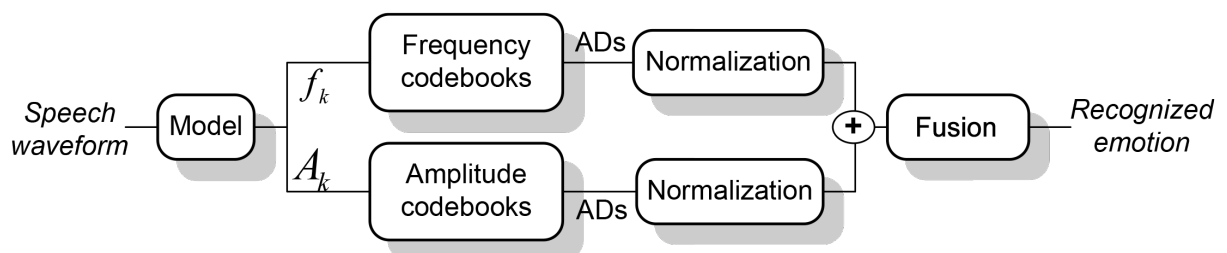


Figure 4.4 – The classification scheme based on the combination of features. A_k and f_k denote the instantaneous amplitude and frequency, and ADs denote the average distortion measures.

4.4 Classification Results

Classification tasks in two of the three databases were performed using the parameters obtained from eaQHM into 64-bit and 128-bit VQ classifiers. The Toshiba database was not used in this task because it contained only 20 waveforms. The classification tasks from single feature and combined feature classification for SUSAS and Berlin using the features obtained from eaQHM are compared with the features obtained from the SM.

4.4.1 SUSAS

Using a subset corpus of the SUSAS, labelled as *Angry*, *Neutral*, *Soft*, and *Question*. A total number of 2520 waveforms (630 per emotion) were used. A number of 756 waveforms were kept for testing (189 per emotion), while the rest were used for training. The confusion Tables for the amplitude-based classification for the 64-bit and 128-bit VQ are given in 4.1 and 4.2, whereas for the corresponding frequency-based ones are given in 4.3 and 4.4.

In general, the parameters obtained from the eaQHM lead to better classification scores in almost all cases(except one). It can be seen that in both cases and both quantizers the *angry* speaking style stands out of the rest of speaking styles. This is expected since this speaking style is very different than the others in terms of amplitude and frequency distributions [79]. Furthermore, the *angry* speaking style has the highest correct classification percentage for both models and both sets of features. The *question* speaking style is the most difficult one to correctly classify when the frequencies are used as features, and we can see that it is mostly confused with

64-bit VQ Classification in % - Amplitudes					
Predicted Class					
C		Angry	Neutral	Soft	Question
l	Angry	73(73)	16(17)	4(2)	7(8)
a	Neutral	4(6)	62(70)	22(16)	12(8)
s	Soft	4(5)	30(30)	53(51)	13(14)
s	Question	5(4)	25(27)	13(16)	57(53)

Table 4.1 – *eaQHM* and *SM* based Confusion Table based on amplitudes for a 64-bit VQ classification between 4 emotions of the SUSAS database. *SM* classification scores are in parenthesis.

128-bit VQ Classification in % - Amplitudes					
Predicted Class					
C		Angry	Neutral	Soft	Question
l	Angry	77(72)	14(14)	2(3)	7(11)
a	Neutral	4(4)	64(63)	18(18)	14(15)
s	Soft	3(5)	31(30)	56(50)	10(15)
s	Question	6(4)	21(22)	13(20)	60(55)

Table 4.2 – *eaQHM* and *SM* based Confusion Table based on amplitudes for a 128-bit VQ classification between 4 emotions of the SUSAS database. *SM* classification scores are in parenthesis.

64-bit VQ Classification in % - Frequencies					
Predicted Class					
C		Angry	Neutral	Soft	Question
l	Angry	67(67)	9(10)	6(2)	17(21)
a	Neutral	4(4)	56(42)	17(26)	22(28)
s	Soft	1(2)	26(23)	57(53)	15(21)
s	Question	16(15)	22(33)	14(21)	48(31)

Table 4.3 – *eaQHM* and *SM* based Confusion Table based on frequencies for a 64-bit VQ classification between 4 emotions of the SUSAS database. *SM* classification scores are in parenthesis.

128-bit VQ Classification in % - Frequencies					
Predicted Class					
C		Angry	Neutral	Soft	Question
l	Angry	71(70)	6(6)	7(5)	21(18)
a	Neutral	6(6)	55(38)	24(28)	15(27)
s	Soft	3(3)	13(25)	65(59)	14(13)
s	Question	17(18)	18(24)	14(25)	55(33)

Table 4.4 – *eaQHM* and *SM* based Confusion Table based on frequencies for a 128-bit VQ classification between 4 emotions of the SUSAS database. *SM* classification scores are in parenthesis.

the *neutral* speaking style. On the other hand, the *soft* speaking style has the lowest classification

score when the amplitudes are used as features.

The results from the combined classification scheme proposed in Section 4.3.2 for eaQHM and SM are shown in Tables 4.5 and 4.6. Using this scheme, on average, the eaQHM correctly classifies 65% of the utterances in the database, whereas the SM reaches 54%. Apparently, not all speaking styles were favoured by this combined scheme. Mostly the *angry* and the *question* speaking style achieved significant increase of their classification rates in both models. While *angry* speaking style already had a relatively high percentage, the *question* speaking style has interestingly increased its correct classification score. However, the *soft* and *neutral* speaking style did not significantly change their percentages. This suggests that a weighted sum of the ADs before ranking may be more appropriate.

VQ Combined Classification in %					
Predicted Class					
C		Angry	Neutral	Soft	Question
l	Angry	83	5	1	11
a	Neutral	15	58	12	15
s	Soft	10	18	56	16
s	Question	20	6	11	63

Table 4.5 – eaQHM-based Confusion Table based on amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of the normalized SUSAS database.

VQ Combined Classification in %					
Predicted Class					
C		Angry	Neutral	Soft	Question
l	Angry	77	5	5	13
a	Neutral	4	48	24	24
s	Soft	2	29	54	15
s	Question	17	24	21	38

Table 4.6 – SM-based Confusion Table based on amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of the normalized SUSAS database.

4.4.2 Berlin

A classification task, by the parameters obtained from eaQHM (amplitude and frequency), was also performed in this Emotional database of Berlin (EmoDB), using the 70% of the total sentences for training and the rest 30% for testing. The classification was performed using 128-bit VQ classifiers, where the frame rate during the analysis was 2.5 ms as it is described in Chapter 4.3.1. The feature extraction was performed as it is described in Chapter 4.2 but no normalization to unit energy was performed.

Firstly, we evaluated the single feature classification rate with the eaQHM and SM models in all 6 speaking styles of Emo DB. The classification scores for amplitudes for eaQHM and SM are presented in Figures 4.5 and 4.6 respectively, whereas for the frequencies as features are shown in Figures 4.7 and 4.8. In this task, the classifier for amplitudes, when the parameters were obtained from eaQHM, could correctly classify the most distinguish emotion, the *angry* emotion, with 76% accuracy between the rest five emotion, while the next emotion with the highest classification rate was the emotion of *fear*. The *neutral* and *happy* emotions achieved the highest classification scores based on the frequency classifier, where the *angry*, *boredom* and *fear* emotions had almost the same accuracy. The *sad* emotion was mostly confused with the *boredom* and *neutral* emotion.

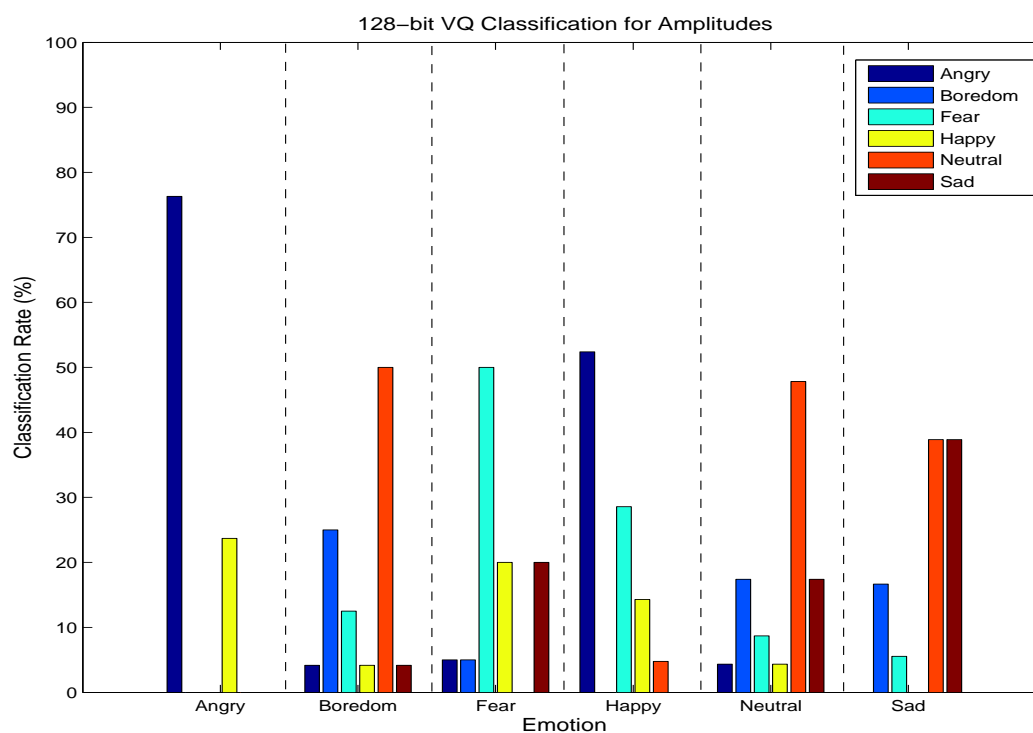


Figure 4.5 – eaQHM-based graph based on amplitudes for a 128-bit VQ classification between all the emotions of Emo DB.

The classification results from SM features show that the *angry* emotion had the best classification scores, especially when the amplitudes were set as features with score 82%. The next emotion with the highest score, for amplitudes as features, is the *neutral* with 57%. The *fear*, *happy* and *sad* emotions are easily confusable with the rest emotions. The *boredom* emotion is the only that it is not correctly classified, it is classified as *neutral* with 67% score. When

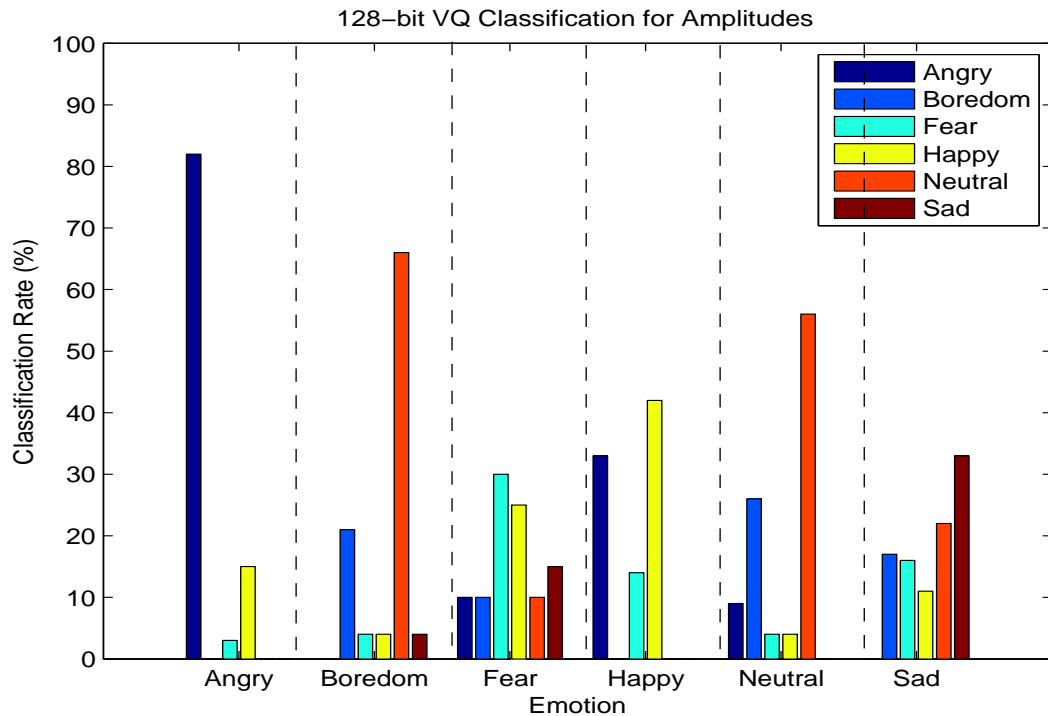


Figure 4.6 – SM-based graph based on amplitudes for a 128-bit VQ classification between all the emotions of Emo DB.

frequencies are set as features, the classification rate of *angry* emotion is decreased and the classification of *happy* reaches the 66%. The *boredom* is also and in this case classified as *neutral*. The rest emotions have low classification rates.

The results from this classification task, as we expected were not as satisfactory because of the plurality of the emotions. Thus, we chose four of the six emotions, the *angry*, *fear*, *neutral* and *sad* to evaluate again the two models. The results of the single features classification between these 4 emotions are depicted in Figures 4.9 and 4.10, for amplitudes for eaQHM and SM and for frequencies at Figures 4.11 and 4.12 respectively. When the parameters of eaQHM are set as features to the classifiers, the best classification score 94% is achieved for the *angry* when amplitudes are set as features. Both *fear* and *neutral* emotions achieve 65% classification score. The *sad* emotion is the mostly confused with the *neutral*, when the classification is based on amplitudes. Although, *neutral* is the most distinguishable emotion, when we have classification based on frequencies, with classification score above 90%. The *angry* and *fear* emotions are following with classification scores 82% and 75% respectively. The *sad* emotion is also confused with the *neutral* one when frequencies are considered and achieves about 60% classification score.

When the SM parameters are obtained for the classification task with 4 emotions of the Berlin database the best classification rates are for the *angry* emotion with 89% when frequencies and amplitudes are used and 91% for the *neutral* when frequencies are used. The *sad* emotion is following with 72% when the frequencies are used for features, but when the amplitudes are used is easily confused with the rest emotions. The *fear* emotion has almost the same results, it is classified only when the amplitudes are set as features with 60% classification rate.

A combined feature classification was also evaluated for EmoDB. The ADs from the separate VQs for amplitude and frequency were normalized by the highest corresponding AD and then

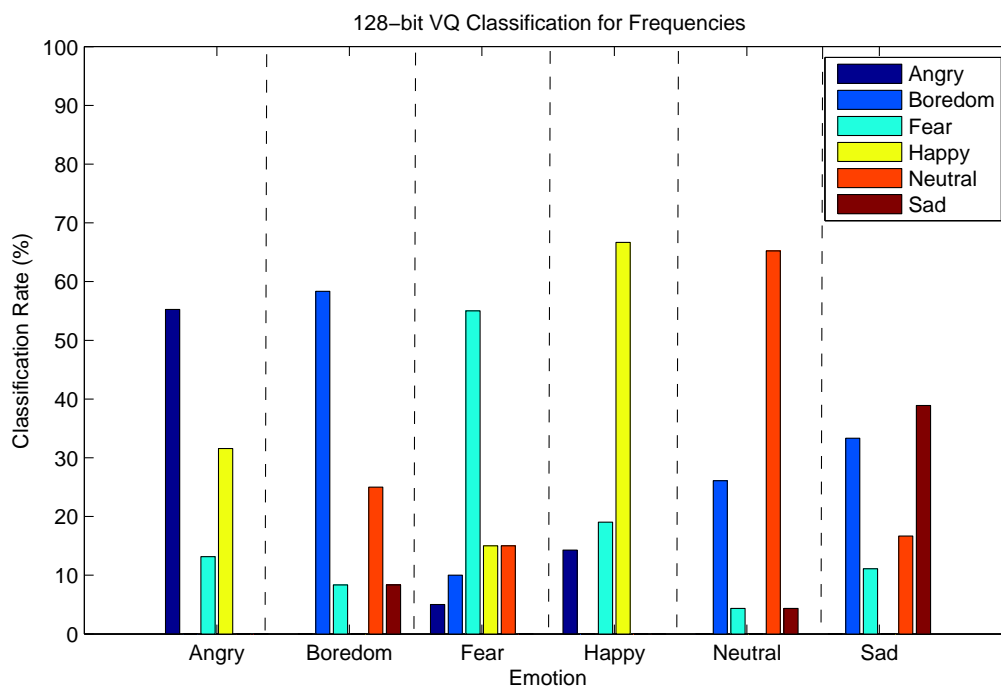


Figure 4.7 – *eaQHM*-based graph based on frequencies for a 128-bit VQ classification between all the emotions of *Emo DB*.

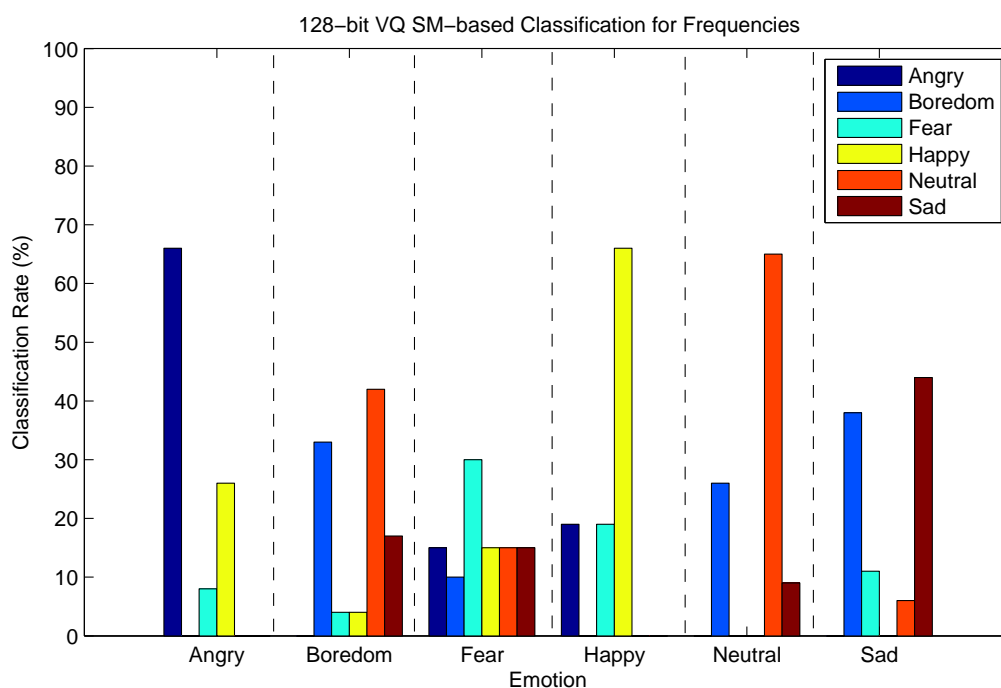


Figure 4.8 – *SM*-based graph based on frequencies for a 128-bit VQ classification between all the emotions of *Emo DB*.

the ADs of the corresponding emotions are added as it is described in Chapter 4.3.2. The results obtained from the combined feature classification task are shown in Figure 4.13 for the *eaQHM* and in Figure 4.14 for the *SM*.

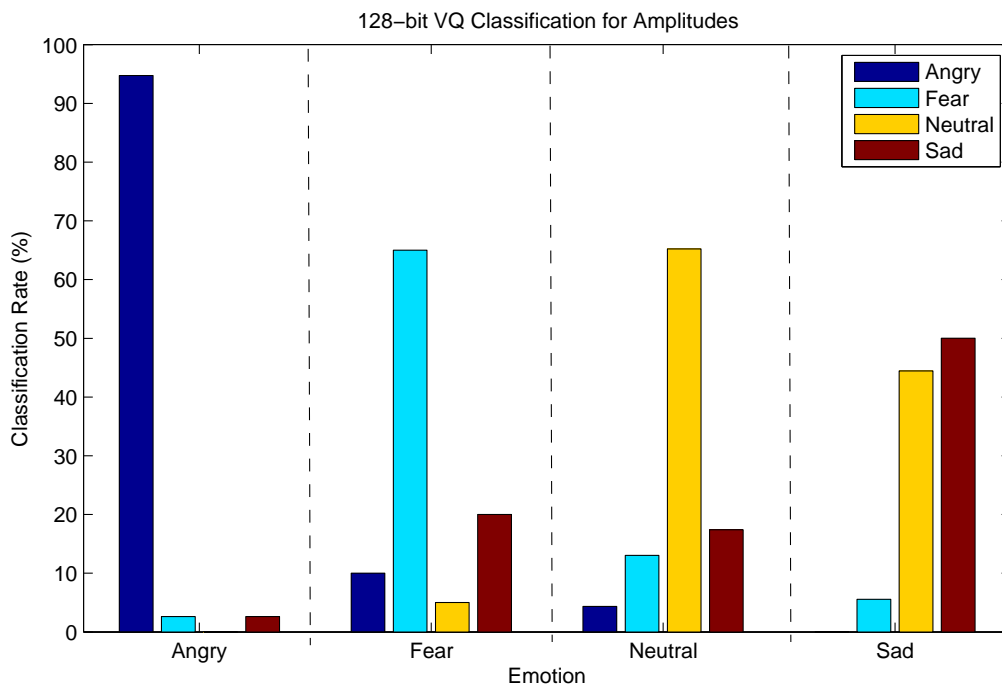


Figure 4.9 – eaQHM-based graph based on amplitudes for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.

The topmost classification score by the eaQHM is 95% for the *neutral* emotion and the *angry* emotion is following with classification rate 89%. The emotion of *fear* was the same rate 75% as with the frequency classifier, whereas the *sad* emotion was reached classification score at 67% but it is still the mostly confusable with the *neutral* emotion. On average, the features obtained from eaQHM correctly classify about 82% of the utterances in the database. Whereas only the *angry* emotion was a lower combined classification score in contrast with the single feature classification.

The results for the SM show that could only correctly classify the 3 out of the 4 emotions. The best classification score was achieved for the *neutral* emotion with 91%, when the *angry* emotion is following with 89%, which is the same classification score with the single feature classification. The best improvement is for the *sad* with achieves 83% score, when *fear* emotion is still not classified correctly. On average, the features obtained from SM classify correctly about 75% of the total utterances in the database.

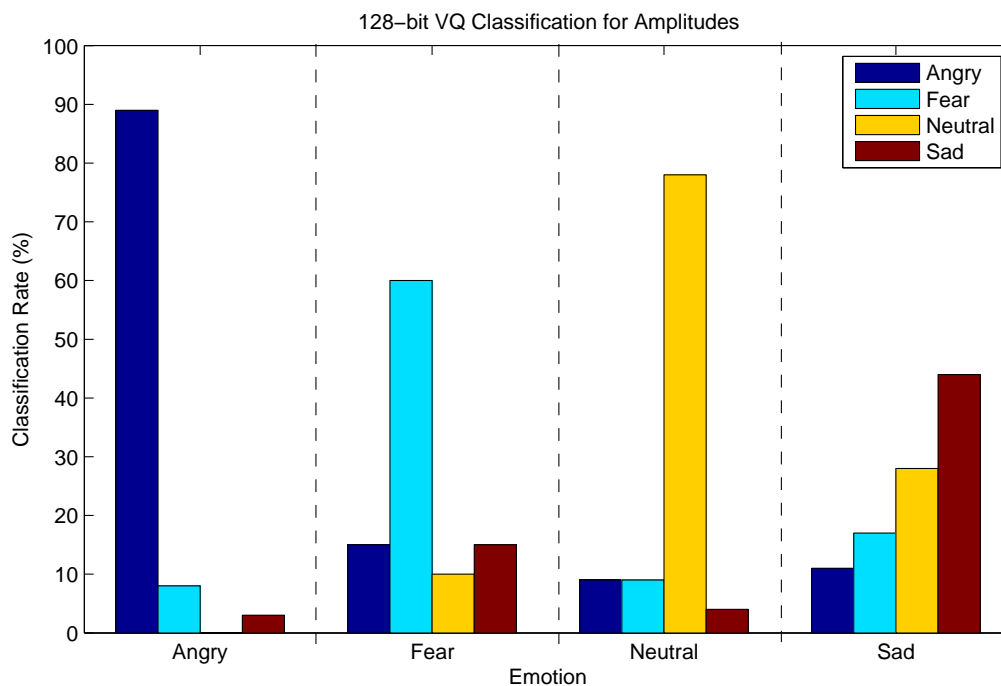


Figure 4.10 – SM-based graph based on amplitudes for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.

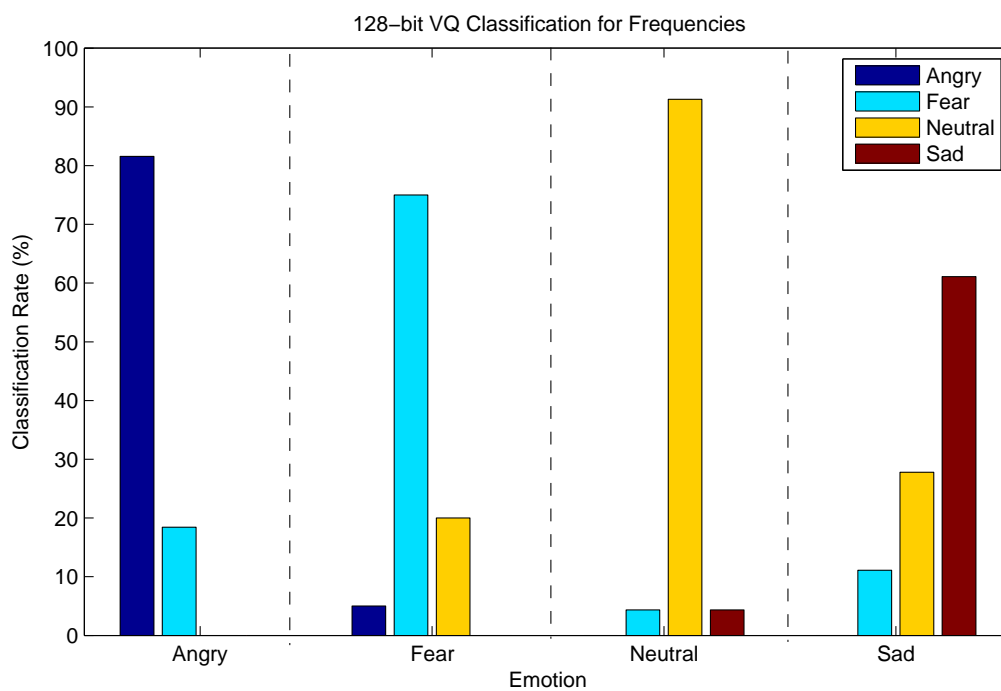


Figure 4.11 – eaQHM-based graph based on frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.

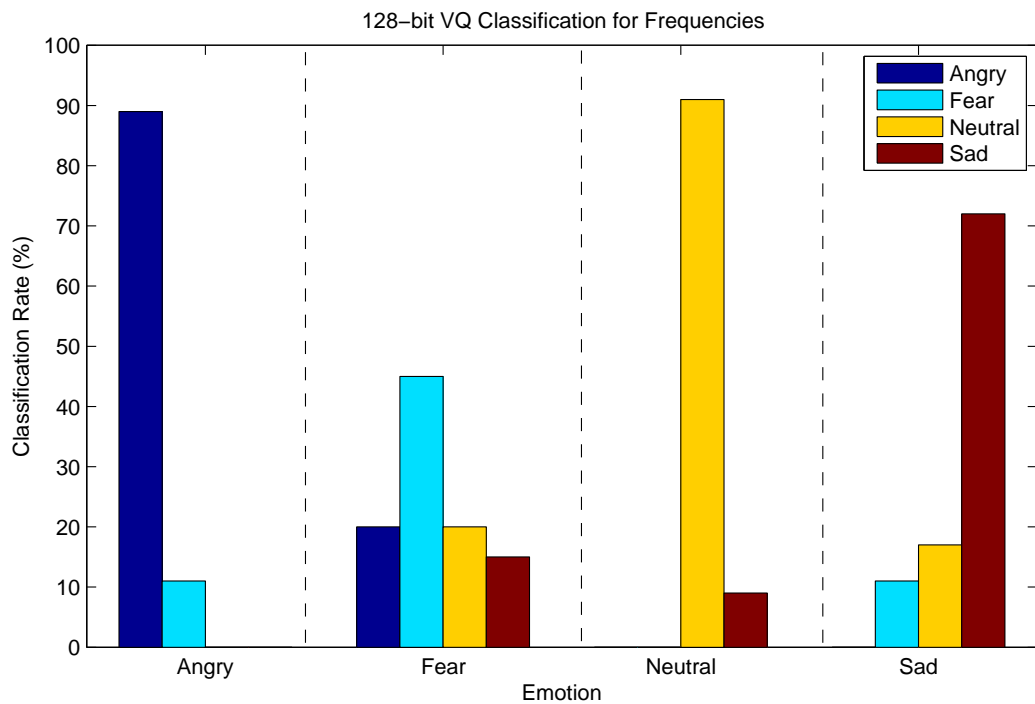


Figure 4.12 – SM-based graph based on frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.

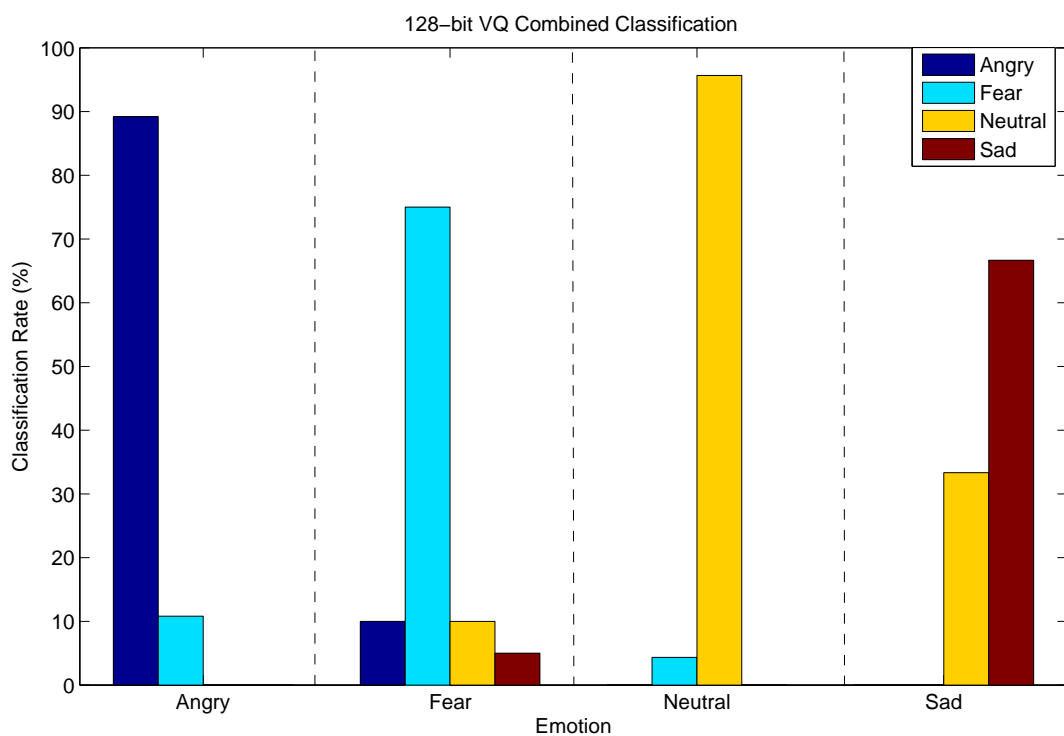


Figure 4.13 – eaQHM-based graph based on combination of amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.

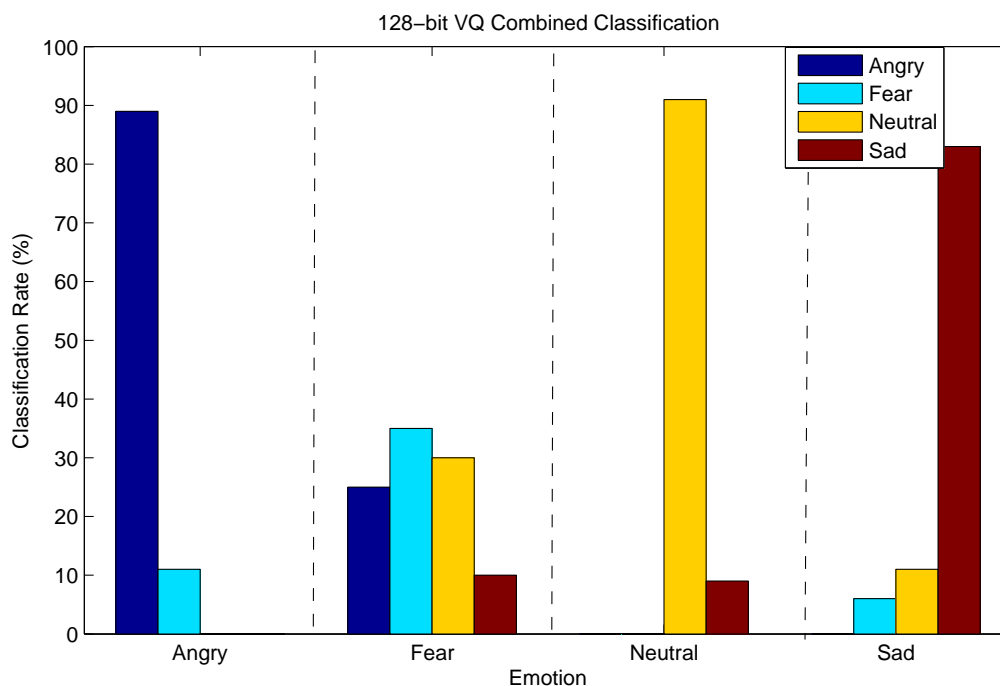


Figure 4.14 – SM-based graph based on combination of amplitudes and frequencies for a 128-bit VQ classification between 4 emotions of Emotional database of Berlin.

4.4.3 Compared to the state of the art MFCC-based Classification

The sounds generated by a human are filtered by the shape of vocal tract, which determines what sound comes out. The accurate determination of vocal tract's shape could give an accurate representation of the *phoneme* being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and MFCCs are used to precisely represent this envelope. The MFCCs are widely used in automatic speech and speaker recognition [21, 90, 10, 78]. As a result, we extended our study with the classification of the MFCCs as features using the emotional database of Berlin in order to compare the results with the classification task from the eaQHM. The classification task was also performed using 128-bit VQ classifiers, where the frame rate during the analysis was 2.5 ms and 30 ms the analysis window, the melcepst function from the voicebox was used. Similarly, the same 70% of the total sentences were used for training and the rest 30% for testing. Initially, a classification task using 12 MFCCs, for each analysis frame, was performed in all styles of emotional database of Berlin and in 4 selected emotion, the results are presented in Table 4.7 and Table 4.8 respectively.

128-bit VQ Classification in MFCC					
Angry		Boredom		Fear	
Angry	27/38	Angry	0/24	Angry	0/20
Boredom	1/38	Boredom	1/24	Boredom	1/20
Fear	1/38	Fear	1/24	Fear	4/20
Happy	9/38	Happy	2/24	Happy	4/20
Neutral	0/38	Neutral	3/24	Neutral	2/20
Sad	0/38	Sad	17/24	Sad	9/20
Total	71%		4%		20%
Happy		Neutral		Sad	
Angry	1/21	Angry	0/23	Angry	0/18
Boredom	0/21	Boredom	0/23	Boredom	0/18
Fear	1/21	Fear	0/23	Fear	0/18
Happy	13/21	Happy	0/23	Happy	0/18
Neutral	4/21	Neutral	4/23	Neutral	0/18
Sad	2/21	Sad	19/23	Sad	18/18
Total	62%		20%		100%

Table 4.7 – MFCC based Table for a 128-bit VQ classification between the emotions of the Berlin database.

128-bit VQ Classification in MFCCs							
Angry		Fear		Neutral		Sad	
Angry	28/38	Angry	2/20	Angry	0/23	Angry	0/18
Fear	5/38	Fear	7/20	Fear	0/23	Fear	0/18
Neutral	5/38	Neutral	3/20	Neutral	4/23	Neutral	0/18
Sad	0/38	Sad	8/20	Sad	19/23	Sad	18/18
Total	74%		35%		17%		100%

Table 4.8 – MFCC based Table for a 128-bit VQ classification between 4 emotions of the Berlin database.

Chapter 5

Conclusions and Future Work

5.1 Overview

In this work, we presented an application of an adaptive sinusoidal model, named eaQHM, on the problem of emotional speech analysis and classification. It was shown that different emotional speech styles can be effectively represented by the adaptivity mechanism of eaQHM, yielding very accurate AM-FM decomposition. This was demonstrated through resynthesis of the original speech signals from its AM-FM components and by evaluating the Singal-to-Reconstruction-Error-Ratio (SRER). A formal listening test was designed to evaluate the perceptual quality of the resynthesized speech and showed that eaQHM-resynthesized emotional speech is indistinguishable from the original. Instantaneous parameters of sinusoidal model were used to perform emotion classification from speech signal. Results showed that a Vector Quantization classification based on eaQHM achieves higher classification scores for a subset of SUSAS database and in emotional database of Berlin, both on single-feature classification based on the sinusoidal parameters and on their combination.

5.2 Future Research Directions

Many interesting research directions can be further investigated such as the use of phase information in combination with amplitudes and frequencies. In [79], the number of *phase reversals* is suggested as a feature. Although it has shown to be useful, a more intuitive measure could be suggested. In [80], the notion of relative phase shift (RPS) is revisited and phase structure is shown to be revealed through RPS. It would be interesting to examine if there are different patterns in RPS structures that can help discriminate emotional content in speech, combined with the standard amplitude and frequency features.

Furthermore, it is has shown in [27, 40, 41] that an implicit information provided by the sinusoidal amplitudes is important in emotion recognition. However, when considering only a part of the full-band, such as the 10 highest spectral peaks, a significant part of the spectrum is not taken into account. Better classification scores may contribute by the inclusion of that part. Moreover, higher frequency components were suggested to be disregarded in sinusoidal model-based emotion classification as inappropriate for the task [79]. Nevertheless, the aSMs are able to follow the dynamics of speech in the upper bands, and thus to reveal the spectral details that are blurred due to the time-frequency trade-off of the FFT-based estimation.

Additionally, vowels have received increasing attention when it comes to emotion recognition. However, consonants are shown to be important as well (see for example [7]). Since our

model is full-band and models both voiced and unvoiced parts of speech using AM-FM components, it would be interesting to examine whether there is any useful information embedded in the sinusoidal representation that is able to distinguish emotions. However, a robust voiced/unvoiced detector (VUD) should be employed for this task. Finally, different classifiers can be used, such as HMMs, SVMs, or GMMs, for a more efficient classification.

Appendix A

Publications

During this work, the following conference papers were published (in chronological order):

1. Kafentzis, G. P., **Yakoumaki T.**, Mouchtaris A., Stylianou Y.,
Analysis of Emotional Speech using an Adaptive Sinusoidal Model,
In European Signal Processing Conference (EUSIPCO), 2014.
2. **Yakoumaki T.**, Kafentzis G. P., Stylianou Y.,
Emotion Classification using adaptive Sinusoidal Modeling,
In Conference of International Speech Communication Association (INTERSPEECH),
2014.

Bibliography

- [1] S. Ahmadi and A. S. Spanias. Low bit-rate speech coding based on an improved sinusoidal model. *Speech Communication*, 34(4):369 – 390, 2001.
- [2] P. Alku. Glottal Wave Analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication*, 11:109–118, 1992.
- [3] P. Alku. Glottal inverse filtering analysis of human voice production : a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana - Academy Proceedings in Engineering Sciences*, 36:623–650, 2011.
- [4] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray. Speech emotion recognition using Gaussian Mixture Vector autoregressive models. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 957–960, 2007.
- [5] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, pages 572–587, 2011.
- [6] M. Bhatti, Y. Wang, and L. Guan. A neural network approach for human emotion recognition in speech. In *Proc. of the International Symposium on Circuits and Systems*, volume 2, pages II–181–4 Vol.2, 2004.
- [7] D. Bitouk, R. Verma, and A. Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, 52(7-8):613–625, 2010.
- [8] M. L. Björn Schuller and G. Rigoll. *Automatic Emotion Recognition by the Speech Signal*, 2002.
- [9] S. Bou-Ghazale and J. Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4):429–442, July 2000.
- [10] S. E. Bou-Ghazale and J. H. L. Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4):429–442, 2000.
- [11] B. Bozkurt, B. Doval, C. D’Alessandro, and T. Dutoit. Zeros of Z-Transform Representation with Application to Source-Filter Separation in Speech. *IEEE Signal Processing Letters*, 12:344–347, 2005.
- [12] E. Bozkurt, E. Erzin, c. E. Erdem, and A. T. Erdem. Formant position based weighted spectral features for emotion recognition. *Speech Commun.*, 53(9-10):1186–1197, Nov. 2011.
- [13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *INTERSPEECH’05*, pages 1517–1520, 2005.
- [14] C. Busso, S. Lee, and S. S. Narayanan. Using neutral speech models for emotional speech analysis. In *Proceedings of InterSpeech*, page 2225?2228, Antwerp, Belgium, Aug. 2007.

- [15] H. Byun and S.-W. Lee. Applications of support vector machines for pattern recognition: A survey. In A. V. Seong-Whan Lee, editor, *Pattern Recognition with Support Vector Machines*, pages 213–236. Springer, 2002.
- [16] D. A. Cairns and J. Hansen. Nonlinear analysis and classification of speech under stressed condition. *Journal of Acoustical Society of America (JASA)*, pages 3392–3400, 1994.
- [17] A. Camacho and J. G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *Journal of Acoustical Society of America (JASA)*, 124:1628–1652, 2008.
- [18] M. Campedel-Oudot, O. Cappe, and E. Moulines. Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach. *IEEE Transactions on Speech and Audio Processing*, 9(5):469–481, Jul 2001.
- [19] O. Cappe, J. Laroche, and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1995.
- [20] S. Chandaka, A. Chatterjee, and S. Munshi. Support vector machines employing cross-correlation for emotional speech recognition. *Measurement*, 42(4):611 – 618, 2009.
- [21] T. Chaspari, D. Dimitriadis, and P. Maragos. Emotion classification of speech using modulation features. In *European Signal Processing Conference (EUSIPCO)*, pages 1552–1556, 2014.
- [22] K. Dai, H. J. Fell, and J. MacAuslan. Recognizing emotion in speech using neural networks. In *Proceedings of the IASTED International Conference on Telehealth/Assistive Technologies*, pages 31–36, Anaheim, CA, USA, 2008. ACTA Press.
- [23] G. Degottex. *Glottal source and vocal-tract separation*. PhD thesis, UPMC-Ircam, France, 2010.
- [24] G. Degottex and Y. Stylianou. Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2085–2095, 2013.
- [25] H. Deng, R. Ward, M. Beddoes, and M. Hodgson. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):445–455, March 2006.
- [26] C. Drioli, G. Tisato, P. Cosi, and F. Tesser. Emotions and voice quality: Experiments with sinusoidal modeling. In *In Proceedings of VOQUAL’A03*, pages 127–132, 2003.
- [27] B. D. Womack and J. H. L. Hansen. N-channel hidden markov models for combined stressed speech classification and recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 7:668–676, 1999.
- [28] A. El-Jaroudi and J. Makhoul. Discrete All-Pole Modeling. *IEEE Transactions on Signal Processing*, 39:411–423, 1991.
- [29] D. Erro, E. Navas, I. Hernáez, and I. Saratxaga. Emotion conversion based on prosodic unit selection. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(05):974–983, 2010.
- [30] G. Fant. *Acoustic Theory of Speech Production*. Mouton De Gruyter, 1970.
- [31] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4:1–13, 1985.
- [32] R. Fernandez. *A computational model for the automatic recognition of affect in speech*. PhD thesis, Massachusetts Institute of Technology, 2004.

- [33] Q. Fu and P. Murphy. Robust Glottal Source Estimation based on Joint Source-Filter Model Optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:492–501, 2006.
- [34] P. Gangamohan, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana. Excitation source features for discrimination of anger and happy emotions. In *interspeech*, pages 1253–1257, 2014.
- [35] E. B. George and M. J. T. Smith. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, 5(5):389–406, 1997.
- [36] H. Gray and W. H. Lewis. *Anatomy of the Human Body*. Lea & Febiger, 1918.
- [37] D. W. Griffin. *Multiband Excitation Vocoder*. PhD thesis, M.I.T, 1987.
- [38] L. S. Han Zhiryan and W. Jian. *Speech emotion recognition system based on integrating feature and improved HMM*.
- [39] J. Hansen and S. Bou-Ghazale. Getting started with SUSAS: A speech under simulated and actual stress database. *EUROSPEECH*, 4:1743 – 1746, 1997.
- [40] J. H. L. Hansen and B. Womack. Feature analysis and neural network based classification of speech under stress. *IEEE Transactions on Audio, Speech, and Language Processing*, 4:307–313, 1996.
- [41] J. H. L. Hansen, B. D. Womack, and L. M. Arsian. A source generator based production model for environmental robustness in speech recognition. In *Proc. ICSLP*, pages 1003 – 1006, 1994.
- [42] R. Heusdens, R. Vafin, and W. Kleijn. Sinusoidal modeling using psychoacoustic-adaptive matching pursuits. *Signal Processing Letters, IEEE*, 9(8):262–265, Aug 2002.
- [43] J. Jensen and R. Heusdens. A comparison of sinusoidal model variants for speech and audio representation. In *in Proc. EUSIPCO'02*, 2002.
- [44] J. Jensen, S. Jensen, and E. Hansen. Exponential sinusoidal modeling of transitional speech segments. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, volume 1, pages 473–476 vol.1, 1999.
- [45] Y. K. Jianhua Tao and A. Li. Prosody conversion from neutral speech to emotional speech. *IEEE Audio*, 14(04):1145–1154, 2006.
- [46] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou. An Extension of the Adaptive Quasi-Harmonic Model. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Kyoto, March 2012.
- [47] G. P. Kafentzis, T. Yakoumaki, A. Mouchtaris, and Y. Stylianou. Analysis of emotional speech using an adaptive sinusoidal model. In *European Signal Processing Conference (EUSIPCO)*, pages 1492 – 1496, 2014.
- [48] P. Khanna and M. S. Kumar. Application of vector quantization in emotion recognition from human speech. In *Information Intelligence, Systems, Technology and Management*, volume 141, pages 118–125. Springer Berlin Heidelberg, 2011.
- [49] M. Kotti and F. Patern? Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *International Journal of Speech Technology*, 15(2):131–150, 2012.
- [50] O. W. Kwon, K. Chan, J.Hao, and T. W. Lee. Emotion recognition by speech signals. In *EUROSPEECH*, pages 125–128, 2003.

- [51] R. L. Rabiner. *Digital Processing of Speech Signals*. Person Education, 1978.
- [52] S.-M. Lee and J.-Y. Choi. *Analysis of emotion in speech using perceived and automatically extracted prosodic features*.
- [53] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Communications*, 28(1):84–95, Jan. 1980.
- [54] J. Liscombe. *Prosody and speaker state: paralinguistics, pragmatics, and proficiency*. PhD thesis, Columbia University, 2007.
- [55] I. Luengo, E. Navas, and I. Hernandez. Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge. In *Interspeech*, pages 332–335, 2009.
- [56] I. Luengo, E. Navas, and I. Hernandez. Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 12(6):490–501, 2010.
- [57] I. Luengo, E. Navas, I. Hernandez, and J. Sanchez. Automatic emotion recognition using prosodic parameters. In *Interspeech*, pages 493–496, 2005.
- [58] M. Lugger and B. Yang. The relevance of voice quality features in speaker independent emotion recognition. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, volume 4, pages 17–20, 2007.
- [59] M. W. Macon, D. D. J. Blumenthal, D. M. A. Clements, and D. R. M. Mersereau. Applications of sinusoidal modeling to speech and audio signal processing. In *report in Georgia Institute of Technology*, 1993.
- [60] J. Makhoul. Linear Prediction: A Tutorial Review. *Proceedings of the IEEE*, 63:561–580, 1975.
- [61] R. K. Mandar Gilke, Pramod Kachare and V. P. Rodrigues. Mfcc-based vocal emotion recognition using ann. In *International Conference on Electronics Engineering and Informatics*, volume 49, pages 150 – 154, 2012.
- [62] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:744–754, 1986.
- [63] R. J. Mcaulay and T. F. Quatieri. Low-rate speech coding based on the sinusoidal model. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*. Marcel Dekker Inc., New York, 1992.
- [64] J. M. Montero, J. Gutiirrez-arriola, J. Colas, E. Enrnquez, and J. M. Pardo. Analysis and modelling of emotional speech in spanish. In *In Proc. of ICPhS*, 1999.
- [65] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. *Neural Computing & Applications*, 9:290–296, 2000.
- [66] D. K. Noam Amir, Ori kerret. *Classifying emotions in speech: a comparison of methods*.
- [67] N. Nogueiras, A. Moreno, A. Bonafonte, and J. Marino. Speech emotion recognition using Hidden Markov Models. In *EUROSPEECH*, pages 2679–2682, 2001.
- [68] T. Nwe, S. Foo, and L. D. Silva. Speech emotion recognition using Hidden Markov Models. *Speech Communication*, 41:603–623, 2003.
- [69] T. L. Nwe, S. W. Foo, and L. C. D. Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623, 2003.

- [70] Y. Pantazis. *Adaptive AM-FM Signal Decomposition With Application to Speech Analysis*. PhD thesis, Computer Science Department, University of Crete, 2010.
- [71] Y. Pantazis, O. Rosec, and Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In *Interspeech*, Brisbane, September 2008.
- [72] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AM-FM signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:290–300, 2011.
- [73] Y. Pantazis, G. Tzedakis, O. Rosec, and Y. Stylianou. Analysis/Synthesis of Speech based on an Adaptive Quasi-Harmonic plus Noise Model. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, March 2010.
- [74] V. A. Petrushin. Emotion recognition in speech signal: experimental study, development, and application. In *INTERSPEECH*, pages 222–225. ISCA, 2000.
- [75] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7:569–587, 1999.
- [76] T. Quatieri and R. McAulay. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6):1449–1464, Dec 1986.
- [77] T. Quatieri and R. McAuley. Audio signal processing based on sinusoidal analysis/synthesis. In M. Kahrs and K. Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*, chapter 9, pages 343–416. Kluwer Academic Publishers, 2002.
- [78] S. Ramakrishnan. Recognition of emotion from speech: A review. *SPEECH ENHANCEMENT, MODELING AND RECOGNITION—ALGORITHMS AND APPLICATIONS*, page 121, 2012.
- [79] S. Ramamohan and S. Dandapat. Sinusoidal model-based analysis and classification of stressed speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):737–746, 2006.
- [80] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez. Simple representation of signal phase for harmonic speech models. *Electronics Letters*, 7, 2009.
- [81] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256, 2003.
- [82] X. Serra. *A System for Sound Analysis, Transformation, Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.
- [83] V. Sethu, E. Ambikairajah, and J. Epps. On the use of speech parameter contours from emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 19:14, 2013.
- [84] M. Sidorov, S. Ultes, and A. Schmitt. Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4836–4840, May 2014.
- [85] P. Stoica, R. Moses, B. Friedlander, and T. Soderstrom. Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements. *IEEE Transactions on Audio, Speech, and Language Processing*, 37(3):378–392, Mar 1989.

- [86] Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, E.N.S.T - Paris, 1996.
- [87] J.-Y. C. Suk-Myung Lee. *Analysis of emotion in speech using perceived and automatically extracted prosodic features*.
- [88] R. Sun, E. M. II, and J. F. Torres. Investigating glottal parameters for differentiating emotional categories with similar prosodics. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.
- [89] J. Tao and Y. Kang. Features importance analysis for emotional speech classification. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, pages 449–457, Berlin, Heidelberg, 2005. Springer-Verlag.
- [90] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162 – 1181, 2006.
- [91] F. Villavicencio, A. Röbel, and X. Rodet. Improving lpc spectral envelope extraction of voiced speech by true envelope estimation. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, 2006.
- [92] D. Vincent, O. Rosec, and T. Chonavel. A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 525–528, 2007.
- [93] O. wook Kwon, K. Chan, J. Hao, and T. won Lee. Emotion recognition by speech signals. In *In Proceedings of International Conference EUROSPEECH*, pages 125–128, 2003.
- [94] S. Wu, T. H. Falk, and W.-Y. Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768 – 785, 2011. Perceptual and Statistical Audition.
- [95] T. Yakoumaki, G. P. Kafentzis, and Y. Stylianou. Emotional speech classification using adaptive sinusoidal modelling. In *Interspeech*, 2014.
- [96] B. Yang and M.Lugger. Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90:1415–1423, 2003.
- [97] B. Yegnanarayana and R. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *Speech and Audio Processing, IEEE Transactions on*, 6(4):313–327, Jul 1998.

This work has been partially funded by Toshiba Europe Limited.

The Toshiba logo is centered on a solid red square background. It consists of the word "TOSHIBA" in a bold, white, sans-serif font. Below it, the phrase "Leading Innovation" is written in a smaller, white, sans-serif font, followed by three white chevron symbols (»») pointing to the right.

TOSHIBA
Leading Innovation >>>

