New Spectral Features and Classifier Architectures for Emotion Recognition from Spontaneous Speech

by

Elif Bozkurt

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Electrical and Computer Engineering

Koç University

July 29, 2010

Koç University
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Elif Bozkurt

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assoc. Prof. Dr. Engin Erzin

Prof. Dr. A. Tanju Erdem

Assoc. Prof. Dr. Çiğdem Eroğlu Erdem

Assoc. Prof. Dr. Yücel Yemez

Asst. Prof. Dr. S. Serdar Kozat

Date:

ABSTRACT

In this thesis, we propose formant position based weighted Mel Frequency Cepstral Coefficient (WMFCC) features for spontaneous emotion recognition from speech problem and compare performance results with commonly used feature sets such as Mel Frequency Cepstral Coefficients (MFCC), Line Spectral Frequency (LSF) features, formants and prosody. Since, the LSF features are positioned close to each other around formant frequencies, we propose normalized inverse harmonic mean function to weight critical band energies for the extraction of MFCC features. We evaluate both the standard and weighted MFCC feature sets with left-to-right Hidden Markov Model (HMM) structures for the five class emotion recognition task. Experimental results on the spontaneous FAU Aibo emotional corpus indicate that WMFCC features perform significantly better than standard spectral features. The HMM classifier with the standard MFCC features attain 39.43 % unweighted recall rate, whereas proposed WMFCC features based HMM classification brings 1.92 % improvement. Another contribution of this thesis is the fusion of classifiers using WMFCC, MFCC and LSF features.

ÖZETÇE

Bu tezde doğal konuşmadan duygu tanıma problemi için biçimlendirici konumu ağırlıklı Mel frekans kepstral katsayısı (AMFKK) özniteliklerini sunuyoruz ve başarım sonuçlarını sıkça kullanılan Mel frekans kepstral katsayıları (MFKK), Doğru Spektral frekans (DSF) katsayıları, biçimlendiriciler ve bürün öznitelikleri başarımları ile karşılaştırıyoruz. DSF öznitelikleri biçimlendirici frekansları çevresinde birbirine yakın konumlandığından, MFKK özniteliklerinin çıkarımında kritik bant enerji değerlerini normalleştirilmiş ters harmonik ortalama fonksiyonu ile ağırlandırıyoruz.Beş sınıflı duygu tanıma problemi için hem standart hem de ağırlıklı MFKK öznitelik vektörlerini sol-sağ yapılı saklı Markov modeller (SMM) ile eğitiyoruz. FAU Aibo duygu yüklü konuşma veritabanı üzerindeki deney sonucları AMFKK özniteliklerinin standart spektral özniteliklerden daha iyi başarım sağladığını ortaya koyuyor. Standart MFKK öznitelikleri % 39.43 başarım sağlarken, AMFKK özniteliklerinin SMM ile sınıflandırılması başarımda 1.92 % değerinde bir artış sağlıyor. Bu tezde ayrıca AMFKK, MFKK ve DSF öznitelikleri kullanılarak eğitilen farklı SMM sınıflandırıcılarının karar kaynaşımı da inceleniyor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Emerging technological advances are inspiring researchers for enriching the meaning of human-computer interaction. The wide use of telecommunication services and multimedia devices will need to have human-centered designs instead of computer centered ones [1]. Consequently, an accurate perception of a user's affective state by computer systems will become a basic requirement for more natural human-computer interaction process [2] [3]. In this sense, the orientation of emotion research is headed towards real, life-like speech-driven advanced applications which has motivated us to investigate spontaneous affect recognition from speech signals.

Initial efforts for affect-sensitive human-computer interaction systems include call-center applications, where problems due to unsatisfactory interaction can be detected and the frustrated customer can be offered assistance of human operators [4][5][6]. Recognition of emotion largely helps to design more natural communication for intelligent automobile systems [7], interactive game and movie systems [8], as well. Similarly, emotion-aware tutoring systems can be included into pedagogical strategies to improve a student's performance and learning [9].

Although extensively investigated, computer recognition of emotions from speech signals is still an open problem. The emotion categories depend on the database used. The most popular description of the basic categorization include the six emotions in addition to neutral state, namely, *happiness*, *sadness*, *fear*, *anger*, *disgust* and *surprise* [10]. This description is also regarded to be cross-cultural indicating that humans perceive these basic emotions in the same way regardless of their cultural background [10]. Some researchers also study cognitive states such as *interest*, *puzzlement*, *frustration* and *boredom* in addition to these basic emotions [11].

Early research in the affective-computing field is based on acted datasets, where actors are asked to speak with a predefined emotion, as a simplification of how emotions happen in real world [12] [13]. This simplification makes it easier to search the acoustic correlation between features and emotion classes. Such approaches aim to recognitize a small number of basic emotions. However, there are objections against the use of acted emotions. It was shown that acted and spontaneous samples differ in the view of features and accuracies [14]. Emotion classifiers have not been successful in realistic contexts when they are trained from acted emotions. Some experiments supported the opinion that acted emotional speech is not felt when spoken and is perceived more strongly than real emotional speech [15].

On the other hand, spontaneous emotional speech datasets introduce difficulties such as highly imbalanced emotion categories as the distributions of recorded emotion samples depend on the content of the dataset. Moreover, utterances may match more than one emotional category given that humans are able to express mixtures of emotions. Thus, it is hard to detect everyday emotions for both humans and computers. Usually, human labelers annotate the spontaneous speech data, since it is not feasible to ask speakers what kind of emotion they have felt during the recordings. One side effect is that mislabeled samples may introduce ambiguity for the training process that majority vote rule is used in the labeling process [16]. Hence, for assessment of the spontaneous emotion classification performance, one should consider human performance on the same task as well.

1.1. System Overview and Contribution

The main contribution of this thesis is to investigate new speech parameter representations that carry emotional cues and try to efficiently model these features for spontaneous emotion recognition task. We evaluate the statistical significance of recognition performances of introduced features with those of well-known, commonly used ones under the same test conditions. Our research has three main contributions:

(i) We propose the use of *Line Spectral Frequency* (LSF) features for emotion recognition which have not been previously employed for this task to the best of our knowledge.

(ii) We introduce *Formant Position-based Weighted Mel Frequency Cepstral Coefficients* (WMFCC), which weights the critical band energies in the computation of MFCC features based on LSF features.

(iii) We investigate decision fusion of different classifiers modeling spectral and prosody features for improved recognition performances.

The remainder of this thesis is structured as follows. In Chapter 2, the necessary background and literature review on emotion recognition from spontaneous speech are provided. The commonly used speech features and classifiers are reviewed.

In Chapter 3, our initial research results obtained in INTERSPEECH 2009 Emotion Challenge is overviewed, spontaneous speech dataset and used feature sets and classifiers are explained.

In Chapter 4, proposed emotion recognition system overview, the employed spectral and prosody features together with the proposed WMFCC features are presented. HMM based classifier architecture for emotion recognition and the decision fusion method are explained. Experiments to assess the performance of the proposed system are also discussed.

Finally, the concluding remarks and future work are presented in Chapter 5.

Chapter 2

LITERATURE REVIEW

Speech is an important communicative modality in human-to-human interaction that conveys affective information through *linguistic* and *acoustic* content. Similarly, for affect-sensitive human-computer interaction systems, the aim of speech emotion recognizer is to estimate emotional state of the speaker given a speech fragment as an input. Although, some researchers report improvement in recognition performance by using linguistic content in addition to acoustic content such as information on language, discourse or context, extraction of this information is a challenging task [17] [18].

Linguistic content based features are extracted manually or directly from transcripts. However, for real life applications, spoken content needs to be recognized by automatic speech recognition (ASR) systems where existing systems cannot reliably recognize the verbal content of emotional speech. Affective word dictionaries are prerequisite for the ASR systems that is a difficult task to build these dictionaries as well, since it is hard to anticipate a speaker’s word choice associated with his/her affective state. Moreover, linguistic content conveying emotion, is language dependent, which is a drawback for generalizing from one language to another [19].

Moreover, in phonetics, acoustic parameters are traditionally categorized as prosodic, spectral and voice quality features. Prosody characteristics are mostly defined as pitch (fundamental frequency, F0), intensity, loudness, speaking rate, duration, pause and rhythm

[8][20][21]. It is well known that for different emotional states, the speech signal carries different prosodic patterns [22]. Hence, prosodic features such as pitch and speech intensity can be used to model different emotions. For example, high values of pitch appear to be correlated with happiness, anger, and fear, whereas sadness and boredom seem to be associated with low pitch values [22]. Intensity based features describe the energy change of the signal over time. Duration based features model the effect of the speaking style on the duration of the spoken utterance. Most popular voice quality features are jitter, shimmer and harmonics to noise ratio (HNR) where jitter is a measure for the cycle-to-cycle variation of the period length, shimmer is peak or average amplitude and HNR is the measure of periodicity of a sound.

Spectral features describe the characteristics of a speech signal in the frequency domain in addition to features like harmonics and formants. Harmonics are multiples of the fundamental frequency and are specified by their frequency and amplitude. Formants are a representation of the vocal tract resonances. Formants have been used to describe the shape of the vocal tract during emotional speech production [8][21]. Each formant is characterized by its center frequency and bandwidth. Experimental analysis has shown that the first and second formants are affected by the emotional states of speech more [23] [24] [25].

Mel frequency cepstral coefficients (MFCC) are also spectral features, widely used for speech recognition and have been designed to extract what is being spoken. They have been successfully used for emotion recognition too [12][26]. Other spectral features useful for speech recognition such as linear predictive cepstral coefficients (LPCC) and mel filter bank (MFB) features are also used for emotion recognition [27].

Acoustic content feature vectors can be represented as long or short-time ones. Long-time features are statistical information estimated over the entire utterance length, most commonly mean, minimum, median, maximum and standard deviation values, and less frequently skewness, a measure of asymmetry and kurtosis, a measure of the peakedness of probability density are also used. Short-time features are determined in a smaller time window, usually 20 to 100 msec. For both of the representation types, the feature vector is often extended to include the first (delta) and second order (delta delta, acceleration) derivatives [14].

In addition to the search for applicable features, an essential aspect of affective-computing is the classification of emotion patterns. A lot of work has been done to develop new and to improve existing automatic classification techniques [27]. Esentially, the percentage of correctly recognized samples is the standard criterium used when evaluating an automatic classifier. Emotion classification methods can be mainly grouped in two, as those of that estimate the probability density function of the features and those of that discriminate emotional states without any estimation of the feature distributions for each emotional state where former models short-time features and the latter long-time features. Popular classifiers include linear discriminant analysis (LDA), artificial neural networks (ANN), support vector machines as well as Gaussian mixture models (GMMs) and hidden Markov models [28][29].

Nevertheless, some researchers reject the idea of categorizing emotions into discrete classes. They claim that in real life people may express combination of emotions with different levels. So, they adopt the continuous representation of emotions in three dimensions, namely *valence*, *activation* and *dominance* as shown in Fig. 2.1 [30] [31]. Valence describes how negative or positive is a specific emotion. For example anger is negative and happiness is positive. Activation, ranging from being passive to being active, describes the internal excitation of a speaker, such as anger is active but sadness is passive, although they have close valence values. Dominance, ranging from weak to strong, represents the apparent strength of the speaker. In contrast to categorical labeling system, raters need a special training to use dimensional labeling system [32] that continuous representation of emotions is rarely used in publicly available datasets.

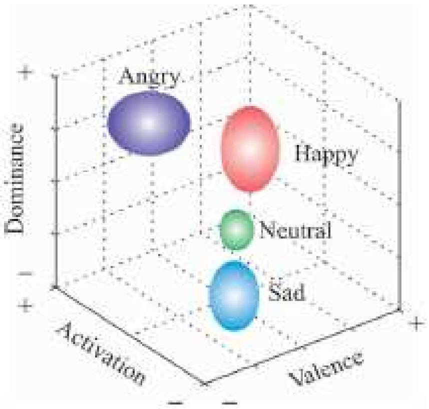

Fig. 2.1. Continuous representation of emotions in three dimensions

Chapter 3

THE INTERSPEECH 2009 EMOTION CHALLENGE

3.1. Introduction

In comparison to speech processing tasks such as automatic speech and speaker recognition, there is a lack of common databases and test-conditions for the evaluation of emotion recognition specific features and classifiers. Existing emotional speech data sources are scarce, mostly monolingual, and small in terms of number of recordings or number of emotions. Among these sources the Berlin emotional speech dataset (EMO-DB) is composed of acted emotional speech recordings in German those are perceived stronger than real emotional speech [33]. Other acted databases include SUSAS (Speech Under Simulated and Actual Stress databases), and DES (Danish Emotional Speech) databases which were seldom made public and the spoken content was mostly predefined [34][35]. The VAM (Vera am Mittag) database consists of audio-visual recordings of German TV talk show with spontaneous and emotionally rich content where speech content is labeled in terms of emotion primitives *valence*, *activation* and *dominance*. The portion of this dataset conveying emotional content well, has unbalanced distribution of emotions and is small in number of recordings [36].

Moreover, partitioning of the existing datasets for evaluation with cross-validation or percentage splits   prevents exact reproducibility. Only Leave-One-Subject-Out cross validation or cross-corpora tests would ensure true speaker indepence. The spontaneous FAU Aibo Emotion Corpus as distributed in the INTERSPEECH 2009 Emotion Challenge, has clearly defined training and test partitions, with guaranteed speaker independence and different room acoustics as needed in most real life settings [37]. Further information on existing databases is detailed in [1] and [2].

3.2. The INTERSPEECH 2009 Emotion Challenge Overview

The FAU Aibo corpus was collected during interaction of 51 children (ages 10-13, 21 male, 30 female, totally 9.2 hours of speech without pauses) with the pet robot Aibo. The robot was actually controlled by a human operator, whereas children were made to believe that the robot was responding to their instructions, where obedient or disobedient responses evoked children’s emotional reactions. The corpus was recorded at two different schools: data of one school was used for training and the other for testing purposes in the challenge [28].

The INTERSPEECH 2009 Emotion Challenge introduced five and two class emotion classification tasks. The five-class classification problem covers emotions *anger*, *empathy*, *neutral*, *positive* and *rest*, with distributions as summarized in Table 3.1. The two-class classification task consists of classes *negative* (anger and emphatic) and *idle* (all non-negative states). Table 3.2 shows the two class grouping of the same dataset. As the classes were unbalanced, the evaluation of the challenge was primarily based on unweighted average recall value (UA) that is the average recall of all classes, and secondly, on the weighted average (WA) recall value (accuracy). The baseline results were produced by the challenge organizers with dynamic and static modeling of low level descriptors of pitch, energy, etc. UA rates for the five and two class tasks were determined as 38.2% and 67.7%, respectively [37].

	Anger	**Emphatic**	**Neutral**	**Positive**	**Rest**	**Total**
Train	881	2093	5590	674	721	9959
Test	611	1508	5377	215	546	8257
Total	1492	3601	10967	889	1267	18216

Table 3.1. FAU AIBO dataset instances for the five-class emotion recognition task in the INTERSPEECH 2009 Emotion challenge

	Negative	**Idle**	**Total**
Train	3358	6601	9959
Test	2465	5792	8257
Total	5823	12393	18216

Table 3.2. FAU AIBO dataset instances for the two-class emotion recognition task in the INTERSPEECH 2009 Emotion challenge

3.3. Feature Extraction and Classification Methods for the Challenge

We investigated various spectral and prosody features, early fusion of different features and late fusion of different classifiers for the INTERSPEECH 2009 Emotion Challenge. In this investigation, we used GMM based emotion classifiers to model the color of spectral and prosody features, and HMM (Hidden Markov Model) based emotion classifiers to model temporal emotional prosody patterns. Spectral features we used for the challenge consist of mel-frequency cepstral coefficients (MFCC), line spectral frequency (LSF) features and their derivatives, whereas prosody-related features consist of mean normalized values of pitch, first derivative of pitch and speech intensity.

Although some of these features have been recently employed for emotion recognition, our investigation included the following novelties: (i) we used LSF features, which are good candidates to model prosodic information since they are closely related to formant frequencies [39], (ii) we employed a novel multibranch HMM structure to model temporal prosody patterns of emotion classes, and (iii) we investigated data fusion of different features and decision fusion of different classifiers.

3.3.1. Prosody Features

The speech signal carries different prosodic patterns for different emotional states [22]. Hence, prosodic features such as pitch and speech intensity can be used to model different emotions. For example, high values of pitch appear to be correlated with happiness, anger, and fear, whereas sadness and boredom seem to be associated with low pitch values [22].

The pitch features of the emotional speech are estimated using the autocorrelation method [38]. Since pitch values differ for each person and the system ideally should be speaker-independent, speaker normalization is applied. For each window of speech with non-zero pitch values, the mean pitch value of the window is removed to achieve speaker normalization. The regions between utterances without a valid pitch (zero-value pitch segments) are filled with zero-mean and unit-variance Gaussian noise to avail proper training of the HMM classifiers. Then, *pitch*, *1st derivative of pitch*, and *intensity* values are used as normalized prosody features, which will be denoted as $\boldsymbol{f}_P$.

3.3.2. Mel Frequency Cepstral Coefficient Features

The *mel-frequency cepstral coefficient* (MFCC) parametric representation is among the most widely used spectral features for emotion recognition [12][26]. MFCCs are expected to model the varying nature of speech spectra across different emotions. We extract MFCC features using a 25 msec Hamming window at intervals of 10 msec and cover frequency range from 300 Hz to the Nyquist frequency. 12 cepstral coefficients with the log-energy is represented as $\boldsymbol{f}_C$.

3.3.3. Line Spectral Frequency Features

Another spectral feature is the *line spectral frequency* (LSF) representation of the linear prediction filter, that was introduced by Itakura, is closely related to formant frequencies [39]. Linear prediction analysis of speech assumes that a short stationary segment of speech can be represented by a linear time invariant all pole filter of the form $H(z) = \frac{1}{A(z)}$, which is a p^{th} order model for the vocal tract.

The LSF decomposition refers to expressing the p-th order inverse filter $A(z)$ in terms of two polynomials $P(z) = A(z) - z^{p+1}A(z^{-1})$ and $Q(z) = A(z) + z^{p+1}A(z^{-1})$, which are used to represent the LP filter as,

$$H(z) = \frac{1}{A(z)} = \frac{2}{P(z) + Q(z)} \tag{3.1}$$

The polynomials $P(z)$ and $Q(z)$ each have $p/2$ zeros on the unit circle, which are interleaved in the interval $[0, \pi]$. These p zeros form the LSF feature representation for the LP model. Note that the formant frequencies correspond to the zeros of $A(z)$. Hence, $P(z)$ and $Q(z)$ will be close to zero at each formant frequency, which implies that the neighboring LSF features will be close to each other around formant frequencies. This property relates the LSF features to the formant frequencies [40], and makes them good candidates to model emotion related information in the speech spectra. We represent the LSF feature vector, estimated over 20 msec frames centered on each 30 msec analysis window of speech as a p = 16 dimensional vector $\boldsymbol{f}_L$.

3.3.4. Dynamic Features

Temporal changes in the spectra play an important role in human perception of speech. One way to capture this information is to use dynamic features, which measure the change in short term spectra over time. The dynamic feature of the i^{th} analysis window is calculated using the following regression formula,

$$\Delta f(i) = \frac{\sum_{k=1}^{K}\left[f(i+k) - f(i-k)\right]k}{2\sum_{k=1}^{K} k^2} \qquad (3.2)$$

where the number of analysis windows in the regression computation is set to $2K + 1 = 5$. The MFCC feature vector, $\boldsymbol{f}_C$, is extended to include the first and second order derivative features, and the resulting dynamic feature vector is represented as $f_{C\Delta} = \left[f_C^{'} \quad \Delta f_C^{'} \quad \Delta\Delta f_C^{'}\right]$ where prime represents vector transpose. Likewise, the LSF feature vector with dynamic features is denoted as $\boldsymbol{f}_{L\Delta}$.

We also combine the pitch-intensity and the MFCC features to form the feature vector $\boldsymbol{f}_{PC}$, and when the first and second order derivatives of this combined feature are also included, we have the feature vector $\boldsymbol{f}_{PC\Delta}$ for non-zero pitch segments.

3.3.5. HMM-based Features

We employed a novel multi-branch HMM structure to model temporal prosody patterns for emotion recognition. Under different emotions, people utter with different intonations, which create different temporal prosody patterns. We employed unsupervised training of parallel multi-branch HMM structures through spectral and prosody features. The HMM structure Λ with B parallel branches is shown in Fig. 3.1, where each branch has N left-to-right states. One can expect that each branch models certain emotion dependent prosody pattern after an unsupervised training process, which includes utterances from different emotional states. After the unsupervised training process we can split the multi-branch HMM Λ into single branch HMM structures, λ_1, λ_2, … , λ_B. Let us define the likelihood of a speech utterance U for the i^{th} branch HMM as,

$$p_i = P(U \mid \lambda_i) \tag{3.3}$$

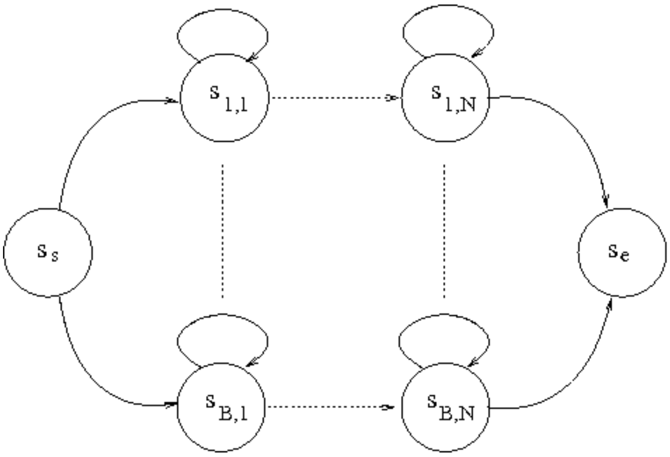

Fig. 3.1. Paralel-branch HMM structure

Then the sigmoid normalization is used to map likelihood values to the [0, 1] range for all utterances [9]. This new set of likelihoods for the utterance *U* define an HMM-based emotion feature set $\boldsymbol{f}_H$,

$$f_H(i) = \left[1 + e^{-\left(\frac{p_i - \bar{p}}{2\sigma} + 1\right)}\right]^{-1} \tag{3.4}$$

where $\bar{p}$ and σ are the mean and the standard deviation of the likelihood p_i over all the training data, respectively. The HMM based emotion feature set $\boldsymbol{f}_H$ is a *B* dimensional vector. We refer to two possible set of features $\boldsymbol{f}_{HP}$ and $\boldsymbol{f}_{HPC}$ when the multi-branch HMM is trained over $\boldsymbol{f}_P$ and $\boldsymbol{f}_{PC\Delta}$ features, respectively.

We experimented the HMM structure with different parameters: we set the number of branches to five and evaluated performance of the model for number of states per branch from 3 to 10 and number of Gaussian components per state up to 12. Since prosody features are extracted every 10 msec, we considered minimum event size from 30 msec to 100 msec for number of states from 3 to 10, respectively. Then, for the 2 and 5-class recognition problems we trained GMM classifiers using the HMM-recognition scores as features.

3.3.6. Gaussian Mixture Model based Emotion Recognition

In the GMM based classifier, probability density function of the feature space is modeled with a diagonal covariance GMM for each emotion. Probability density function, which is defined by a GMM, is a weighted combination of *K* component densities given by

$$p(f) = \sum_{k=1}^{K} w_k \, p(f \mid k) \tag{3.5}$$

where *f* is the observation feature vector and w_k is the mixture weight associated with the k^{th} Gaussian component. The weights satisfy the constraints,

$$0 \le w_k \le 1 \text{ and } \sum_{k=1}^{K} w_k = 1 \tag{3.6}$$

The conditional probability *p(f |k)* is modeled by Gaussian distribution with the component mean vector μ_k, and the diagonal covariance matrix Σ_k.

The GMM for a given emotion is extracted through the expectation-maximization based iterative training process using a set of training feature vectors representing the emotion. In the emotion recognition phase, posterior probability of the features of a given speech utterance is maximized over all emotion GMM densities. Given a sequence of feature vectors for a speech utterance, F = {f_1, f_2, . . . , f_T}, let's define the *log* likelihood of this utterance for emotion class *e* with a GMM density model γ_e as,

$$\rho_{\gamma_e} = \log p(F \mid \gamma_e) = \sum_{t=1}^{T} \log p(f_t \mid \gamma_e) \tag{3.7}$$

where $p(F|\gamma_e)$ is the GMM probability density for the emotion class *e* as defined in (3.5). Then, the emotion GMM density that maximizes posterior probability of the utterance is set as the recognized emotion class,

$$\in = \arg\max_{e \in E} \rho_{\gamma_e} \tag{3.8}$$

where *E* is the set of emotions and $\in$ is the recognized emotion.

3.3.7. Decision Fusion

Decision fusion is used to compensate for possible misclassification errors resulting from a given feature set decision with other available feature set decisions hence resulting in a more reliable overall decision. In decision fusion, scores resulting from each unimodal classification are combined to arrive at a conclusion. Decision fusion is especially effective when contributing modalities are not correlated and resulting partial decisions are statistically independent.

We considered a weighted summation based decision fusion technique to combine different classifiers [41]. The GMM based classifiers output likelihood scores for each emotion and utterance. Likelihood streams need to be normalized prior to the decision fusion process. First, for each utterance, likelihood scores of both classifiers are mean-removed over emotions. Then, sigmoid normalization is used to map likelihood values to the [0, 1] range for all utterances [41]. After normalization, we have two score sets for each GMM based classifier composed of likelihood values for each emotion and utterance. Let us denote normalized log-likelihoods of GMM based classifiers as $\overline{\rho}_{\gamma_e}$ and $\overline{\rho}_{\lambda_e}$ respectively, for the emotion class *e*. The decision fusion then reduces to computing a single set of joint log-likelihood ratios, ρ_e, for each emotion class *e*. Assuming the two classifiers are statistically independent, we fuse the two classifiers, $\gamma_e \oplus \lambda_e$, by computing the weighted average of the normalized likelihood scores

$$\rho_e = \beta\overline{\rho}_{\gamma_e} + (1-\beta)\overline{\rho}_{\lambda_e} \tag{3.9}$$

where the value β weighs the likelihood of the first GMM classifier, and it is selected in the interval [0, 1] to maximize the recognition rate.

3.4. THE INTERSPEECH 2009 Emotion Challenge Results

INTERSPEECH 2009 Emotion challenge participants did not have access to the labels of the test data, and all model selection and training was based only on the training data. Each participant could upload instance predictions to receive the confusion matrix and results from the test data set up to 25 times.

In this section, we present the experimental results using all of the features described in section 3.3. The GMM mixture components and the decision fusion parameter β are optimally selected to maximize emotion recall rate on a part of the training corpus where GMMs have mixtures up to 50 and β is in the range [0, 1]. We used one third of the training set as validation set for model selection purposes. Then, based on the selected parameters, we retrained models using all the available training data. In fact, leave-one-speaker-out cross validation strategy would be more preferable for model selection, but concerning the time limit in challenge, we chose a simpler and faster approach. Recognition rates for the uni-modal GMM classifiers are given in Table 3.3. $\boldsymbol{f}_{PC\Delta}$ GMM and $\boldsymbol{f}_{C\Delta}$GMM classifiers have the highest UA rate as 66.39 % and 39.94 % for the two and five-class recognition problems, respectively.

Features	**Recall (%)**			
	2-class		**5-class**	
	UA	**WA**	**UA**	**WA**
$\boldsymbol{f}_{C\Delta}$	66.36	62.09	39.94	41.29
$\boldsymbol{f}_{L\Delta}$	66.05	60.24	39.10	41.78
$\boldsymbol{f}_{L}$	63.36	65.25	33.68	40.39
$\boldsymbol{f}_{PC\Delta}$	66.39	60.70	39.10	46.66

$\boldsymbol{f}_{HP}$	-	-	24.56	21.30
$\boldsymbol{f}_{HPC}$	59.82	57.43	29.53	27.48

Table 3.3. Emotion recognition rates with unimodal GMM based classifiers

Decision fusion of different classifiers has been realized as defined in Section 3.3.7. The highest recognition rates for each decision fusion are listed in Table 3.4. Decision fusion of classifiers provides statistically significant improvement over unimodal classifiers. Among the decision fusion of GMM based classifiers, $\boldsymbol{f}_{PC\Delta}$ and $\boldsymbol{f}_{L\Delta}$ fusion yields the highest 5-class recognition rate, 40.90 %, with β = 0.57, where β is the weight of the first classifier in the fusion. In addition, fusion of $\boldsymbol{f}_{C\Delta}$ and $\boldsymbol{f}_{L\Delta}$ has 67.52 % UA rate for the 2-class recognition problem when β = 0.64. We observe that the $\boldsymbol{f}_{HPC}$ feature set with 3 states per HMM branch and 12 Gaussian components per state yields the best results with a classification accuracy of 59.82 % and 29.53 % for 2 and 5-class classification tasks respectively. When we apply a second stage decision fusion to $\boldsymbol{f}_{C\Delta}$ and $\boldsymbol{f}_{L}$ fusion results with HMM-based feature $\boldsymbol{f}_{HPC}$, we obtain 67.90 % and 41.59 % recognition rates, respectively.

Classifier Fusion	**Recall (%)**			
	2-class		**5-class**	
	UA	**WA**	**UA**	**WA**
$\gamma(f_{C\Delta}) \oplus \gamma(f_{L})$	67.49	64.44	40.47	42.07
$\gamma(f_{C\Delta}) \oplus \gamma(f_{L\Delta})$	67.52	62.58	40.76	43.71
$\gamma(f_{PC\Delta}) \oplus \gamma(f_{L\Delta})$	67.44	61.64	40.90	47.83
$\gamma(f_{C\Delta}) \oplus \gamma(f_{HP})$	-	-	40.22	41.37
$\gamma(f_{C\Delta}) \oplus \gamma(f_{HPC})$	-	-	40.10	41.50

$(\gamma(f_{C\Delta}) \oplus \gamma(f_{L\Delta})) \oplus \gamma(f_{HP})$	-	-	40.69	43.33
$(\gamma(f_{C\Delta}) \oplus \gamma(f_{L\Delta})) \oplus \gamma(f_{HPC})$	67.90	63.03	41.59	44.17

Table 3.4. Emotion recognition rates after the decision fusion

We summarize all paper submissions accepted to the challenge in Table 3.5 for comparison of selected methods. Most popular feature sets include mel frequeny cepstral coefficients, harmonics to noise ratio, pitch, energy and zero crossing rate, whereas most popular classifiers are Gaussian Mixture Models and Support Vector Machines. Almost every participant applied decision fusion method subsequent to classification. Our emotion recognition system ranked second and fourth for the five and two-class classification tasks, respectively.

	Paper Title & Authors	**Feature Set & Classifier**	**Best UA (%) (rank)**	
			2-class	**5-class**
1	*Brno University of TechnologySystem for Interspeech 2009 Emotion Challenge* Kockmann, Burget and Cernocky	RASTA applied MFCC ΔΔ parameters are modeled with JFA (Joint factor analysis) to cope with speaker session variability for the 2-class problem. For the 5-class problem this feature set-model pair is fused with SDC (shifted delta cepstra) features modeled with JFA based on multiclass linear regression fusion. It is also reported that voice activity recognition with a Hungarian phone recognizer is applied prior to feature extraction process.	68.3 (3)	41.65 (1)
2	*Improving Automatic Emotion Recognition from Speech Signals* Bozkurt, Erzin, E. Erdem and Erdem	MFKK ΔΔ and Line spectral frequency (LSF) feature vectors are modeled with Gaussian mixture models (GMM) and then first stage decision fusion is applied. Hidden Markov model (HMM) based emotion features are extracted and modeled with GMMs. Recognition results with the score score from the first stage decision fusion are fused with the best HMM-based feature set-GMM structure.	67.9 (4)	41.59 (2)

3	*GTM-URL Combination to theInterspeech 2009 Emotion Challenge* Planet, Iriondo, Socoro, Monzo and Adell	Functionals of features zero crossing rate (ZCR), RMS energy, F0, HNR (harmonics to noise ratio) and MFCCs are modeled with Naïve Bayes classifier.	-	41.16 (5)
4	*Acoustic Emotion Recognition Using Dynamic Bayesian Networks and Multi-space Distributions* Chicote, Fernandez, Lutfi, Cuesta, Guarasa, Montero, Segundo and Pardo	Multispace distribution (MSD) approach is applied for F0 voiced/unvoiced segments prior to feature extraction. MFCC ΔΔ, Log F0 ΔΔ, Log Energy ΔΔ features are modeled with Dynamic Bayesian network.	67.06 (7)	38.24 (8)
5	*Emotion Recognition Using A Hierarchical Binary Decision Tree Approach* Lee, Mower, Busso, Lee and Narayanan	Z-normalization and binary logistic regression with forward selection as feature selection are applied to functionals of features zero crossing rate (ZCR), RMS energy, F0, HNR (harmonics to noise ratio) and MFCCs. Then, selected features are modeled with multi-stage Bayesian logistic regression.	-	41.57 (3)
6	*Combining Spectral and Prosodic Information for Emotion recognition in the Interspeech 2009 Emotion Challenge* Luengo, Navas and Hernaez	Firstly, Mel scale short time log frequency power coefficients with ΔΔ parameters are modeled with GMMs wit 32 mixtures. Secondly, for non-pause segments statistics for intonation, power, rhythm, regression, voice quality, sentence end features are ranked with LDA (linear discriminant analysis) and modeled with RBF-based (radial basis funcion) SVMs (support vector machines). Finally, these results are fused using SVMs.	67.19 (6)	41.38 (4)
	Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features Polzehl, Sundaram, Ketabdar, Wagner and Metze	An automatic speech recognizer is used to transcribe words based on degree of emotional salience. Next, information gain filter is applied for feature selection of functionals of features intensity, F0, MFCC, Formants, ZCR, duration, HNR. Both acoustic and linguistic features are modeled with RBF-SVMs and then fused.	67.55 (5)	-
8	*Cepstral and Long-term Features for Emotion Recognition* Dumouchel, Dehak, Attabi, Dehak and Boufaden	Only voiced parts of speech is used to extract MFCC ΔΔ features. For the 5-class task these features are modeled with MAP adapted UBM-GMMs*. For the 2-class task in addition to adaptation, training with SVMs is applied. Then, best scores of adaptation and SVM-based recognition results are fused with linear logistic regression fusion**.	*69.72 (2) **70.29 (1)	39.4 (6)
9	*Exploring the Benefits of Discretization of Acoustic*	Feature selection with CFS (correlation based subset selection) is applied to functionals of F0,	66.4	39.4

	Features for Speech Emotion Recognition Vogt and Andre	energy, spectral, cepstral voice segments, voice quality, jitter, shimmer and modeled with Naïve Bayes classifier.	(8)	(7)

Table 3.5. The INTERSPEECH 2009 Emotion Challenge results

Chapter 4

FORMANT POSITION BASED WEIGHTED SPECTRAL FEATURES FOR SPONTANEOUS EMOTION RECOGNITION

4.1. Introduction

In this section, we introduce formant position based weighted Mel frequency cepstral coefficients (WMFCC) to attack the emotion recognition problem. Emotion has a considerable influence on formant positioning [22] and the LSF features are known to concentrate around formant positions [40]. However, formant features are hard to track accurately and LSF's are easy to compute. Based on these facts, we propose a spectral weighting function based on LSF features to weight the critical band energies in the computation of MFCC features. We derive WMFCC features from the weighted critical band energies, and employ them for improved emotion recognition task.

4.2. System Overview

A block diagram of our automatic speech driven emotion recognition system is given in Fig. 4.1. Speech is the only input modality that drives the emotion recognition system. The overall system is trained and tested on the FAU Aibo Emotional dataset which contains speech utterances reflecting the five emotions, namely, *anger, emphatic, neutral, positive*, and *rest*. In the training part of our system, which is shown in the upper half of Fig. 4.1, first the emotional speech data is parameterized into the short-term acoustic features. Features extracted include: spectral features, such as *Mel-frequency Cepstral Coefficients* (MFCC), *Line Spectral Frequency* (LSF) features, *formants*, *formant position based weighted MFCC features* and their dynamic parameters (i.e., the first and second derivatives), as well as the prosody-related features consisting of *mean normalized pitch*, *first derivative of pitch*, and *intensity*. Then, we use Hidden Markov Model (HMM) -based emotion classifiers for modeling the temporal emotional speech patterns of each feature set [45].

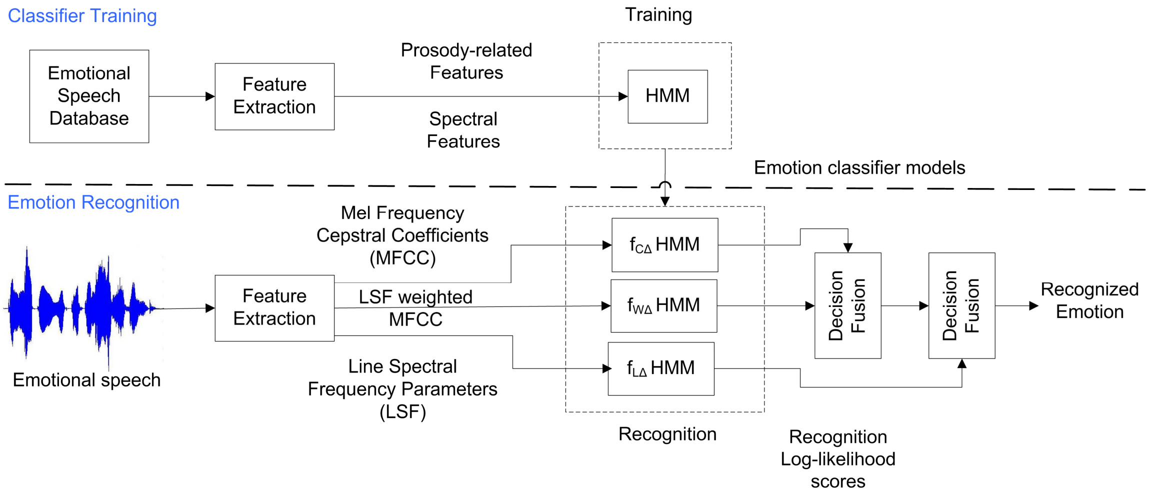

Fig. 4.1. Proposed emotion recognition from speech system overview.

The emotion recognition part of the system is illustrated in the bottom half of Fig. 4.1. Similar to the classifier training part of our system, first input speech is parameterized as spectral and prosody-related features. Then, speech-driven emotion recognition is carried out using the previously trained HMM classifiers. Next, decision fusion of classifiers is applied and at the end of the fusion step, the emotion class with the highest recognition score is accepted to be the recognized emotion.

4.3. Feature Extraction

Two types of information sources are available to determine the emotional status of a speaker from his/her speech, the acoustic content and the linguistic content of the speech. In this study, we only consider the acoustic content by using both prosody-related features and spectral features. The features that are utilized and the proposed formant position based weighted MFCC feature set are defined in the following sub-sections.

$\boldsymbol{f}_{P}$	Mean normalized pitch, derivative of pitch and intensity
$\boldsymbol{f}_{F\Delta}$	First two formants F1 and F2 with dynamic features
$\boldsymbol{f}_{C}$	MFCC features
$\boldsymbol{f}_{C\Delta}$	MFCC and dynamic features
$\boldsymbol{f}_{L}$	LSF features
$\boldsymbol{f}_{L\Delta}$	LSF and dynamic features
$\boldsymbol{f}_{W\Delta}$	WMFCC and dynamic features

Table 4.1. Feature set representations.

The notation that we use to represent the features are summarized in Table 4.1. We denote prosody-related features with $\boldsymbol{f}_{P}$ whereas, represent spectral features like MFCCs and LSFs with $\boldsymbol{f}_{C}$ and $\boldsymbol{f}_{L}$, respectively. Moreover, Δ symbol in the representations, stands for the dynamic features such as first and second derivatives. First two formant frequency features with dynamic parameters are expressed as $\boldsymbol{f}_{F\Delta}$. Spectral features, extended to include the dynamic features are symbolized with $\boldsymbol{f}_{C\Delta}$ and $\boldsymbol{f}_{L\Delta}$, respectively. Finally, the proposed formant position-based weighted MFCCs with dynamic features are symbolized as $\boldsymbol{f}_{W\Delta.}$

4.3.1. Prosody-related Features

Prosody-related features are extracted as detailed in Section 3.3.1.

4.3.2. Formants

Formants are the resonant frequencies of the vocal tract filter. Emotion has a considerable influence on formant positioning, especially on the placement of first two formants F1 and F2 [23]. We employ the first two formant frequencies with delta and acceleration parameters (first and second derivatives), and denote them as $\boldsymbol{f}_{F\Delta}$.

Formants are extracted using the PRAAT speech analysis software with standard settings [42]. The maximum number of formants are tracked (five) and the maximum frequency of the highest formant is set to 8000 Hz for all speakers. The time step between two consecutive analysis frames is selected as 10 msec within an analysis window of size 25 msec. The default value for amount of pre-emphasis (50 Hz) is used.

4.3.3. Mel Frequency Cepstral Coefficients

Mel frequency cepstral coefficient features are extracted as detailed in Section 3.3.2.

4.3.4 Line Spectral Frequency Coefficients

Line spectral frequency coeffcient features are extracted as detailed in Section 3.3.3.

4.3.5 LSF weighted MFCC Features

In Section 3.3.3, we noted that neighboring LSF features will be close to each other around formant frequencies. Using this fact, inverse harmonic mean (IHM) weighting function was introduced for weighted quantization of LSF parameters [43]. The IHM weighting function is defined as,

$$w_i = \begin{cases} \dfrac{1}{f_L^{I+1} - f_L^i} & i = 1 \\ \dfrac{1}{f_L^i - f_L^{i-1}} + \dfrac{1}{f_L^{i+1} - f_L^i} & i = 2,3,\ldots,p-1 \\ \dfrac{1}{f_L^i - f_L^{i-1}} & i = p \end{cases} \tag{4.1}$$

Where f_L^i is the *i*-th line spectral frequency for *p*-th order filter and w_i is the corresponding IHM weight. In order to normalize and further control the weight of the high frequency spectral contributions, we define a normalized IHM weighting function as,

$$\overline{w_i} = \left(\frac{w_i}{\sum_j w_j} \right)^{\alpha} \quad i = 1,2,\ldots,p \tag{4.2}$$

where α is the control parameter.

In the extraction of MFCC features, each analysis frame is first multiplied with a Hamming window and transformed to frequency domain using Fast Fourier Transform (FFT). Mel-scaled triangular filter-bank energies, e_i, which are located at critical band frequencies m_i, are calculated over the square magnitude of the spectrum and represented in logarithmic scale [44]. Since the first two formant positionings are reported to be influential on emotion recognition [23], we propose an IHM based weighting of the critical band energies. Let's consider the critical band frequency m_i falling between two neighboring line spectrum frequencies f_L^{n-1} and f_L^n. Then the critical band weighting function is formed with a linear interpolation of normalized IHM weightings,

$$v_i = \frac{\overline{w}_n\left(m_i - f_L^{n-1}\right) + \overline{w}_{n-1}\left(f_L^n - m_i\right)}{f_L^n - f_L^{n-1}} \tag{4.3}$$

where N_B is the number of critical bands and the boundary line spectrum frequencies are defined as $f_L^0 = 0$ and $f_L^{p+1} = \pi$. The IHM based critical band weighting function is normalized to retain a unity sum as,

$$\overline{v_i} = \frac{v_i}{\sum_j v_j} \quad i = 1,2,\ldots,N_B \tag{4.4}$$

The proposed weighted MFCC features, WMFCC, f_W^i, are derived using discrete cosine transform (DCT) over weighted log-scaled filter-bank energies,

$$f_W^i = \frac{1}{N_B}\sum_{i=1}^{N_B} \overline{v}_i e_i \cos\left(\left(i - 0.5\right)\frac{j\pi}{N_M}\right) \quad j = 1,2,\ldots,N \tag{4.5}$$

where N is the number of WMFCC features that are extracted.

Sample IHM based critical band weighting functions for α control parameter values in the range [1, 4], are presented at the bottom part of Fig. 4.2, where four peak points labeled as F1-F4, denote the predicted first four formant positions. The upper part of the same figure visualizes the actual speech spectra where underlying speech frame has four visible formants corresponding to four highest peaks in the *log*-magnitude representation. From the figure, it is obvious that the proposed LSF based weighting function can successfully locate these formant positions.

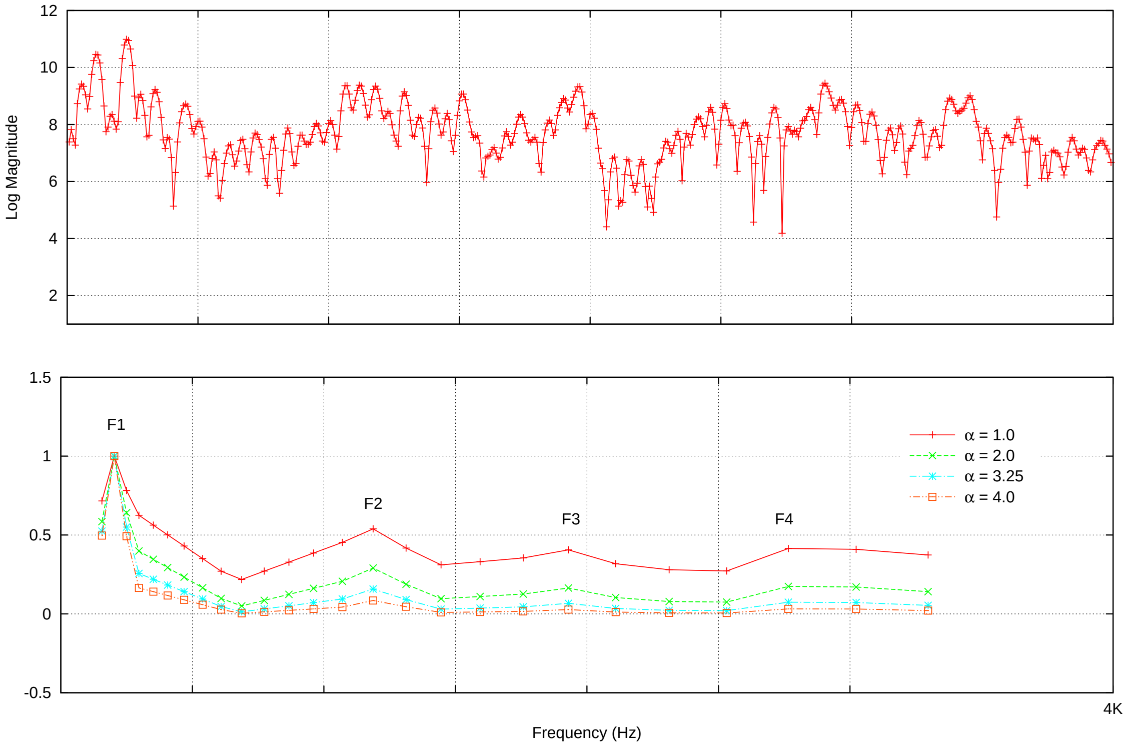

Fig. 4.2. *Top*: Actual speech spectra for a voiced speech frame in the *log*-magnitude representation. *Bottom*: Sample IHM based critical band weighting functions for various α values for the same voiced speech frame.

4.4. HMM-based Classification

HMM structures have been deployed with great success in automatic speech recognition to model temporal spectral information; they were also used similarly for emotion recognition [29]. We model the temporal patterns of the emotional speech utterances through HMM structures. The HMM structures are set to have N left-to-right states with M mixture components per state. One can expect that different emotions define differentiation in observation probability density functions with M mixtures over N states. Structural parameters N and M are determined through a model selection method where we choose the highest average recognition rate and non-zero transition probability as the selection criteria. In the emotion recognition phase, the likelihood of the features of a given speech utterance is computed over HMM structures for each emotion class. Then, the utterance is classified as expressing the emotion which yields the highest likelihood score.

4.5. Decision Fusion

The decision fusion is applied as detailed in Section 3.3.7.

4.6. Performance with HMM classifiers

4.6.1 Performance of the Proposed WMFCC Features

We evaluate the proposed formant position weighted MFCC features with HMM based classifiers for a range of control parameter, α, values. Note that, as defined in Chapter 4.3.5, $\alpha = 0$ case corresponds to $f_{C\Delta}$ features. While $\alpha \geq 1$, as α increases the emphasis of higher formants decrease in the proposed weighting function. In Table 4.2., we present recognition performance of WMFCC features modeled with 2 state HMMs, with mixtures [8, 120].

mixtures	**All Training Data UA Recognition Rates (%)**																	
	MFCCΔ	**Formant position based weighted MFCCΔ** *(for various α values)*																
	(0.0)	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**3.05**	**3.1**	**3.15**	**3.2**	**3.25**	**3.3**	**3.35**	**3.4**	**3.45**	**3.5**	**3.75**	**4.0**
8	39.02	39.39	39.55	39.26	38.96	39.74	39.40	39.00	39.58	38.14	37.74	37.44	38.29	37.80	39.50	39.60	39.50	38.71
16	39.87	39.15	39.90	39.17	40.64	39.69	38.96	39.17	39.71	38.44	39.40	39.34	38.90	39.46	39.43	39.69	39.14	38.32
32	39.89	39.71	39.81	38.34	39.63	39.44	39.23	38.44	38.18	38.92	38.49	39.67	39.73	39.41	38.67	38.28	39.24	38.48
48	39.90	38.80	38.90	38.74	39.31	39.41	39.41	39.18	39.21	39.29	39.40	40.44	40.09	40.37	39.36	38.96	39.91	39.42
64	39.58	39.71	40.18	39.40	39.48	40.59	39.58	39.61	39.15	40.21	39.35	39.03	39.50	39.08	39.36	40.35	39.82	39.93
80	39.43	40.25	39.70	39.96	40.61	39.66	40.54	39.89	40.67	40.50	41.35	40.14	39.88	40.83	40.20	40.26	40.46	40.00
96	39.67	39.63	39.14	39.69	40.11	39.28	40.04	40.53	41.09	39.97	40.92	39.42	40.56	40.70	40.75	41.24	41.44	39.73
112	39.79	39.75	38.33	39.40	39.97	40.46	40.09	40.85	40.32	40.70	40.59	40.08	40.93	41.49	40.03	41.06	40.80	39.68
120	39.69	39.55	38.45	39.10	38.46	40.75	40.13	39.90	39.93	40.22	40.62	40.10	40.24	40.62	40.39	39.96	40.86	39.78

Table 4.2. UA recognition rates for HMMs with number of states 2 and number of mixtures 8-120, modeling MFCC-ΔΔ and formant position based weighted MFCC-ΔΔ (WMFCC) feature sets. WMFCC feature vectors depend on the α weight value that ranges from 0.0 to 4.0, where $\alpha = 0$ case corresponds to standard MFCC definition. In the experiments all available training data is used for training. Mixtures are increased 2 by 2 and following every mixture increment re-estimation of models is applied for 12 times.

Fig. 4.3 visualizes the emotion recognition performances of WMFCC features for varying α values in Table 4.2. In the figure, α value ranges from 1.5 to 4 on the horizantal axis. On the vertical axis, recognition rates for left-to-right HMM classifiers with two states are presented. From top to bottom, HMM structures have 80, 96 and 112 mixture components, respectively. For ease of comparison of the performances, we visualize standard MFCC features' recognition rates with a horizantal line for each mixture component. Since the distribution of emotional classes in the FAU Aibo dataset is highly unbalanced, the performance is measured as unweighted recall (UA) rates that is the average recall of all classes. We observed significant performance improvements with respect to standard MFCC features for α values in [3, 4] interval. As seen in the figure, while the standard MFCC feature attains 39.43% recognition rate with the 80 mixture HMM classifier, the proposed WMFCC feature attains 41.35% recognition rate at $\alpha = 3.25$ value.

In order to evaluate MFCC and WMFCC based classifiers, we performed McNemars test, which is a paired success/failure trial using the binomial model. The McNemar's value is computed as 136.65, which is significantly larger than statistical significance threshold $\chi^2_{(1,0.95)} = 3.8414$.

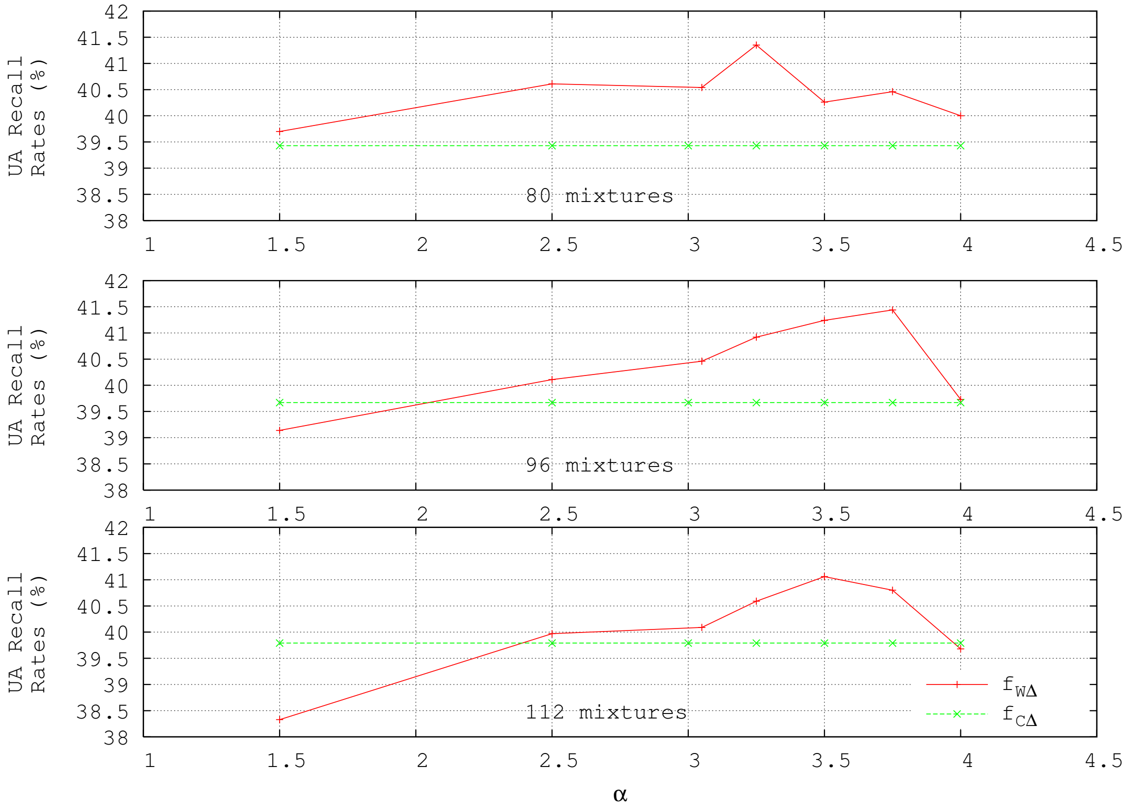

Fig. 4.3. Unweighted recall (UA) rates of WMFCC features for various *α* values

4.6.2. Performance Comparison of Prosody and Spectral Features

The spectral and prosody feature sets as defined in Chapter 4.3 are used with HMM based classifiers for the evaluation of emotion recognition. We employ the left-to-right HMM structure with various number of states and mixtures using the FAU Aibo training data for each feature set. The emotion recognition performances of features $\boldsymbol{f}_{P}$, $\boldsymbol{f}_{F\Delta}$, $\boldsymbol{f}_{L\Delta}$ and $\boldsymbol{f}_{C\Delta}$,for increasing number of mixtures and various number of states in the HMM structure are plotted in Fig. 4.4.

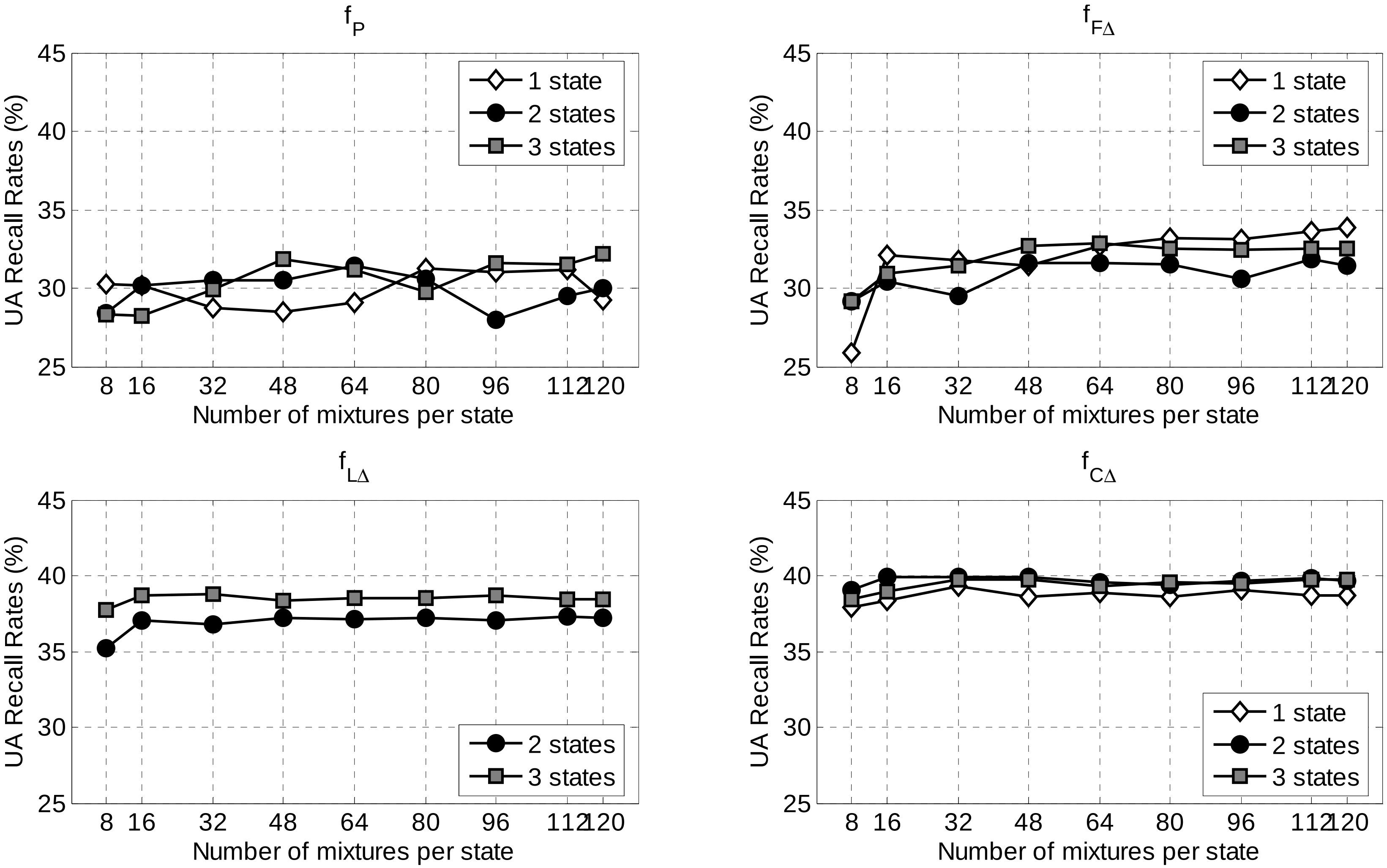

Fig. 4.4. Unweighted recall (UA) rates of spectral and prosody-related features modeled with 1-3 state HMMs for number of mixtures per state in the range [8, 120]

Observing Fig. 4.5 we summarize best performance UA value for each feature set in table 4.3. Prosody features perform best for 3 state HMMs with 120 mixtures as 32.18 %. First two formant features with dynamic parameters, modeled with single state and 120 mixtures HMM has UA value 33.89 %. 38.75 % is the highest UA rate for LSF features with dynamic parameters when modeled with 3 state and 72 mixture HMM structure. Finally, MFCC features with delta parameters has highest UA recall rate 39.90 % for HMMs with 2 state and 48 mixtures per state. WMFCC features overperform all features for HMM mixture value 80. The highest UA recall rate is achieved with WMFCC features as 41.35%. The prosody and formant features perform significantly lower than other spectral features.

Feature	f_P	$f_{F\Delta}$	$f_{L\Delta}$	$f_{C\Delta}$	$f_{W\Delta}$
UA recall rate (%)	32.18	33.89	38.75	39.90	41.35

Table 4.3. The highest UA recall rates for each feature set.

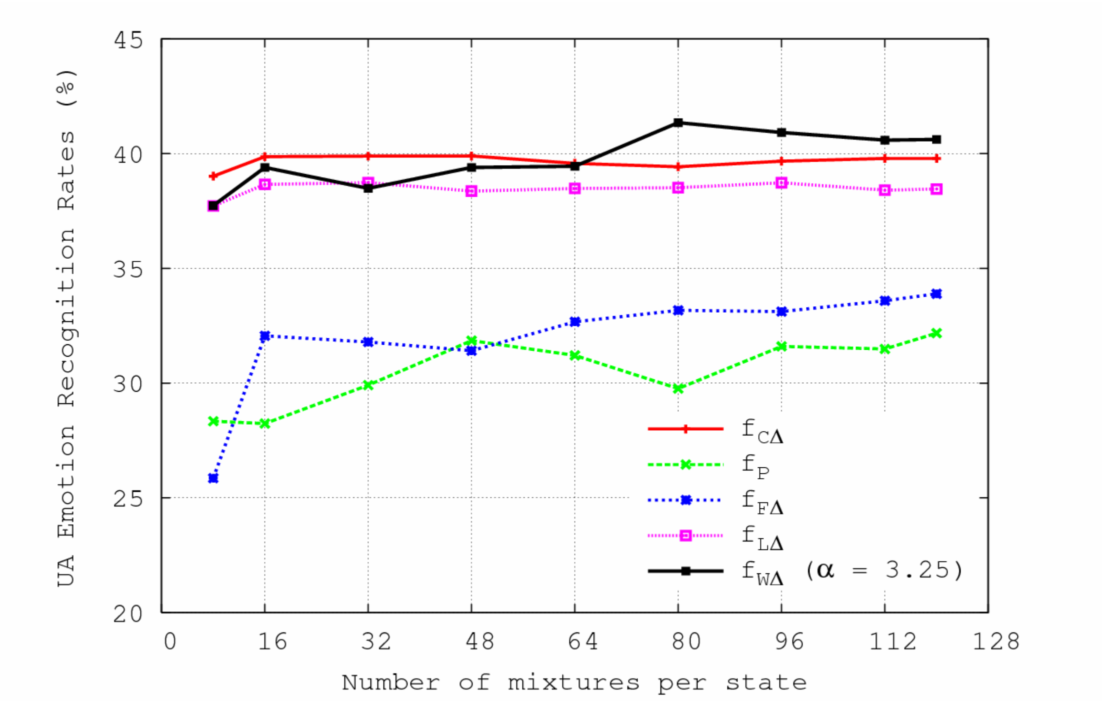

Fig. 4.5. Unweighted recall (UA) rates of spectral and prosody features for increasing number of mixture components per HMM state.

4.7. Performance of the Decision Fusion

Decision fusion of HMM classifiers is performed for MFCC, LSF and WMFCC spectral features. The highest recognition rates for each decision fusion are listed in Table 4.4. Among the decision fusion of pair of classifiers, $f_{W\Delta}$ and $f_{C\Delta}$ fusion yields the highest recognition rate, 42.63 %, with fusion weight β = 0.21. We use WMFCC features modeled with 2 state HMMs and 80 mixtures per state. The corresponding confusion matrix of the fusion of pair of classifiers is presented in Table 4.5. When this classifier fusion is further fused with $\boldsymbol{f}_{L\Delta}$ based classifier, the UA recall rate improved to 43.28 % with fusion weight β = 0.96.

Decision Fusion	UA recall Rate (%)	β
$f_{W\Delta} \oplus f_{C\Delta}$	42.63	0.21
$f_{W\Delta} \oplus f_{L\Delta}$	42.43	0.93
$(f_{W\Delta} \oplus f_{C\Delta}) \oplus f_{L\Delta}$	43.28	0.96

Table 4.4. The highest recognition rates after decision fusion.

	A	E	N	P	R
Anger	292	151	83	21	64
Emphatic	204	804	372	35	93
Neutral	559	1247	2557	400	614
Positive	7	7	80	80	40
Rest	72	74	165	86	149

Table 4.5. Confusion matrix for decision fusion of pair of classifiers with $\boldsymbol{f}_{W\Delta}$ and $\boldsymbol{f}_{C\Delta}$ features.

Chapter 5

CONCLUSIONS AND FUTURE WORK

We introduced a novel formant position based weighted Mel Frequency Cepstral Coefficient (WMFCC) feature set for a speech driven emotion recognition system. The experiments with the spontaneous emotional speech corpus FAU Aibo yield significant performance improvement with WMFCC features for high number of mixture components with left-to-right HMM structures. It’s expected that a high number of mixture components do a better job capturing emotion dependent variations in the spectral feature space. Furthermore, decision fusion of classifiers with different spectral features yields a 43.28% UA recall rate, which is significantly above the best scoring 41.65 % UA recall rate in the INTERSPEECH 2009 Emotion Challenge.

We investigate the contribution of the line spectral frequency (LSF) features to the speech driven emotion recognition task. The LSF features are known to be closely related to the formant frequencies, however they have not been previously employed for emotion recognition to the best of our knowledge. We demonstrate through experimental results on FAU Aibo emotional speech database that the LSF features are indeed beneficial and bring about consistent recall rate improvements for emotion recognition from speech. In particular, the decision fusion of the LSF features with the MFCC features results in improved classification rates over the state-of-the-art MFCC-only decision for both of the databases.

It is also interesting that in the challenge dataset proposed HMM-based features did not perform as accurate as spectral features. Nevertheless, after the two-stage decision fusion they brought significant improvement. Decision fusion strategy was more succesfull than unimodal training strategy in our experiments. Different decision fusion techniques can be tested for better results.

Further research should include feature pruning strategies to lower the confusion between emotional classes in FAU Aibo like spontaneous emotional speech datasets like FAU Aibo.

BIBLIOGRAPHY

[1] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 39–58, 2009.

[2] D. Ververidis and C. Kotropoulos, “Emotional Speech Recognition: Resources, Features, and Methods,” *Speech Comminication*, vol. 48, pp. 1162–1181, 2006.

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, “How to Find Trouble in Communication,” *Speech Comminication*, vol. 40, pp. 117–143, 2003.

[4] C. M. Lee and S. S. Narayanan, “Toward Detecting Emotions in Spoken Dialogs,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 293–303, 2005.

[5] D. Neiberg, and K. Elenius, “Automatic Recognition of Anger in Spontaneous Speech,” in *Proceedings of Interspeech 2008*, ISCA, Brisbane, Australia. pp. 2755–2758, 2008.

[6] D. Morrison, R. Wang and L. C. D. Silva, “Ensemble Methods for Spoken Emotion Recognition in Call-centers,” *Speech Communication*, vol. 49, pp. 98–112, 2007.

[7] B. Schuller, M. Lang, and G. Rigoll, “Recognition of Spontaneous Emotions by Speech within Automative Environment,” *DAGA*, Braunschweig, pp. 57–58, 2006.

[8] R. Nakatsu, J. Nicholson and N. Tosa, “Emotion Recognition and its Applications to Computer Agents with Spontaneous Interactive Capabilities,” *Knowledge-based Systems*, vol. 13, pp. 497–504, 2000.

[9] S. D’Mello, A. Graesser and R.W. Picard, “Toward an Affect Sensitive Auto Tutor,” *Intelligent Educational Systems*, vol. 22, pp. 53-61, 2007.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion Recognition in Human-Computer Interaction, ” *IEEE Signal Processing Magazine*, vol.18, no.1, pp. 32-80, 2001.

[11] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, “Audio-visual Affective Expression

Recognition Through Multistream Fused HMM," *IEEE Transactions on Multimedia,* vol. 10, no. 4, 2008.

[12] B. Vlasenko, B. Schuller, A. Wendemuth and G. Rigoll, "Frame vs. Turnlevel: Emotion Recognition from Speech Considering Static and Dynamic Processing," in *Proceedings of Affective Computing and Intelligent Interaction*, Lisbon, Portugal, pp. 139–147, 2007.

[13] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Automatic hierarchical classification of emotional speech," in *Proceedings of 3rd IEEE International Workshop on Multimedia Information Processing and Retrieval*, Taichung, Taiwan, May 2-5 2007, pp. 291–296.

[14] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. of the 2005 IEEE International Conference on Multimedia Expo (ICME-2005),* Amsterdam, Netherlands, 2005.

[15] J. Wilting, E. Krahmer, M. Swerts, "Real vs Acted Emotional Speech". Proc. INTERSPEECH' 2006, Pittsburgh, 2006.

[16] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Neimann, "Of all the Things Measure is Man, Automatic Classification of Emotions and Inter-labeler Consistency," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 2005),* pp. 317-320, 2005.

[17] C.M. Whissell, "The Dictionary of Affect in Language," Emotion: Theory, Research and Experience. The Measurement of Emotions, R. Plutchik and H. Kellerman, eds., vol. 4, pp. 113-131, Academic Press, 1989.

[18] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, V. Aharonson, "Patterns, Prototypes, Performance: Classifying Emotional User States," *Interspeech (2008), ISCA.*

[19] G. Furnas, T. Landauer, L. Gomes, and S. Dumais, "The Vocabulary Problem in

Human-System Communication," Comm. ACM, vol. 30, no. 11, pp. 964-972, 1987.

[20] T. Polzin and A. Waibel, "Emotion-sensitive Human-computer Interfaces," in *Proceedings of Interspeech 2000*, ISCA, Belfast, 2000.

[21] C. Lee, S. Yildirim, M. Bulut, A. Kazamzadeh, C. Busso, Z. Deng, S. Lee and S. Narayanan, "Emotion Recognition based on Phoneme Classes," in *Proceedings of International Conference on Spoken Language Processing*, Jeju Island, 2004.

[22] K. R. Scherer, "How Emotion is Expressed in Speech and Singing," in *Proceedings of XIIIth International Congress of Phonetic Sciences*, pp. 90–96, 1995.

[23] M. B. Goudbeek, J. P. Goldman and K. Scherer, "Emotion Dimensions and Formant Position," in *Proc. Interspeech 2009,* Brighton,UK, 2009.

[24] F. J. Tolkmitt, K. R. Scherer, "Effect of experimentally induced stress on vocal parameters," J*ournal of Experimental Psychology: Human Perception and Performance*, vol. 12 (3), pp. 302–313, 1986.

[25] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Trans. Biomedical Engineering 7, 829–837.

[26] M. Grimm, E. Mower, K. Kroschel and S.Narayanan, "Combining Categorical and Primitives-based Emotion Recognition," in *Proc. 14th European Signal Processing Conference*, 2006.

[27] C. Busso, S. Lee, and S.S. Narayanan, "Using Neutral Speech Models for Emotional Speech Analysis," *Interspeech 2007 – Eurospeech, 10th European Conference on Speech Communication and Technology*, Antwerp, Belgium, pp. 2225-2228, 2007.

[28] S. Steidl, 'AutomaticClassification of Emotion-related User States in Spontaneous Children's Speech', *PhD Thesis, Technical University of Erlangen-Nurnberg, Germany*, 2009.

[29] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov Model based Speech Emotion

Recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP),* 2003.

[30] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu and T. S. Huang, "Audio-visual Affect Recognition in Activation-Evaluation Space," in *Proc. of the 2005 IEEE International Conference on Multimedia Expo (ICME-2005),* Amsterdam, Netherlands, 2005.

[31] H. P. Espinoza, C.A. R. Garcia, L. V. Pineda, "Features Selection for Primitives Estimation on Emotional Speech," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP),* 2010.

[32] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, "'Feeltrace': An Instrument for Recording Perceived Emotion in Real Time," *Proc. ISCA Workshop Speech and Emotion,* pp. 19-24, 2000.

[33] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *In Proceedings of Interspeech*, pp. 1517–1520, Lisbon, Portugal, 2005.

[34] J. H. L. Hansen, 1996. NATO IST-03 (Formerly RSG. 10) speech under stress web page. URL http://cslrcolorado.edu/rspl/stress.html

[35] I. S. Engberg, A. V. Hansen, "Documentation of the Danish Emotional Speech database (DES)," Internal AAU report, Center for Person Kommunikation, Aalborg Univ., Denmark.

[36] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-visual Emotional Speech Database," *In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME),* Hannover, Germany, 2008

[37] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," *In Interspeech (2009),* ISCA, Brighton, UK, 2009.

[38] J. Deller, J. Hansen and J. Proakis, *Discrete-Time Processing of Speech Signals,* Macmillan Publishing Company, New York, 1993.

[39] F. Itakura, “Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals,” *Journal of the Acoustical Society of America*, vol. 57, no. 35, 1975.

[40] R.W. Morris and M. A. Clements, “Modification of Formants in the Line Spectrum Domain,” *IEEE Signal Processing Letters*, vol. 9, pp. 19–21, 2002.

[41] E. Erzin, Y. Yemez, and A. M. Tekalp, “Multimodal Speaker Identification using an Adaptive Classifier Cascade based on Modality Realiability,” *IEEE Transactions on Multimedia*, vol. 7, pp. 840–852, 2005.

[42] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 4.5.15),” available online at http://www.praat.org/, 2009.

[43] R. Laroia, N. Phamdo and N. Farvardin, “Robust and Efficient Quantization of Speech LSP Parameters using Structured Vector Quantizers,” in *Proceedings of 1991 International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 91),* Toronto, Ont., Canada. pp. 641–644, 1991.

[44] L. Rabiner and B. H. Juang, *Fundemantals of Speech Recognition*, Prentice Hall, 1993.

[45] Hidden Markov Model Toolkit version 3.4.1, available at link *http://htk.eng.cam.ac.uk/*