Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 1186–1197

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Formant position based weighted spectral features for emotion recognition

Elif Bozkurt [a], Engin Erzin [a,*], Çiğdem Eroğlu Erdem [b], A.Tanju Erdem [c]

[a] *Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, 34450 Sariyer, Istanbul, Turkey*
[b] *Department of Electrical and Electronics Engineering, Bahçeşehir University, 34353 Beşiktaş, Istanbul, Turkey*
[c] *Department of Electrical and Electronics Engineering, Özyeğin University, 34662 Üsküdar, Istanbul, Turkey*

Available online 6 May 2011

Abstract

In this paper, we propose novel spectrally weighted mel-frequency cepstral coefficient (WMFCC) features for emotion recognition from speech. The idea is based on the fact that formant locations carry emotion-related information, and therefore critical spectral bands around formant locations can be emphasized during the calculation of MFCC features. The spectral weighting is derived from the normalized inverse harmonic mean function of the line spectral frequency (LSF) features, which are known to be localized around formant frequencies. The above approach can be considered as an early data fusion of spectral content and formant location information. We also investigate methods for late decision fusion of unimodal classifiers. We evaluate the proposed WMFCC features together with the standard spectral and prosody features using HMM based classifiers on the spontaneous FAU Aibo emotional speech corpus. The results show that unimodal classifiers with the WMFCC features perform significantly better than the classifiers with standard spectral features. Late decision fusion of classifiers provide further significant performance improvements.

Keywords: Emotion recognition; Emotional speech classification; Spectral features; Formant frequency; Line spectral frequency; Decision fusion

1. Introduction

Emotion-sensitive machine intelligence is a basic requirement for more natural human–computer interaction. In this sense, the orientation of research on emotional speech processing shifts from the analysis of acted towards spontaneous speech for advanced real-life applications in human–machine interaction systems (Ververidis and Kotropoulos, 2006; Batliner et al., 2003). The wide use of telecommunication services and multimedia devices will require human-centered designs instead of computer centered ones. Consequently, accurate perception of the user's affective state by computer systems will be crucial for the interaction process (Zeng et al., 2009). Examples of recent emotion-aware systems include call-center applications (Lee and Narayanan, 2005; Neiberg and Elenius, 2008; Morrison et al., 2007), intelligent automobile systems (Schuller et al., 2006) and interactive movie systems (Nakatsu et al., 2000).

Although extensively investigated, automatic emotion recognition from speech remains as an open problem in the field of human–computer interaction. Researchers mostly focus on defining a universal set of features that carry emotional clues and try to develop classifiers that efficiently model these features. Some commonly used speech features for emotion recognition are the mel-frequency cepstral coefficients (MFCC) (Vlasenko et al., 2007; Grimm et al., 2006), the fundamental frequency (F0, pitch), which has been referred as one of the most important features for determining emotion in speech (Nakatsu et al., 2000; Polzin and Waibel, 2000; Lee et al., 2004), and the resonant frequencies of the vocal tract, also known as formants (Nakatsu et al., 2000, 2004).

* Corresponding author.
E-mail addresses: ebozkurt@ku.edu.tr (E. Bozkurt), eerzin@ku.edu.tr (E. Erzin), cigdem.eroglu@bahcesehir.edu.tr (Ç. Eroğlu Erdem), tanju.erdem@ozyegin.edu.tr (A.T. Erdem).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.04.003

The contributions and scope of the paper can be stated under three items: (i) The main contribution is the introduction of novel spectrally weighted mel-frequency cepstral coefficient (WMFCC) features for emotion recognition from speech. Recently, Goudbeek et al. (2009) reported that emotion has a considerable influence on formant positioning. Based on this information, we propose WMFCC features by emphasizing spectral content of the critical spectral bands around formant locations. The spectral weighting is obtained from the normalized inverse harmonic mean function of line spectral frequency (LSF) features. Experimental results demonstrate the superiority of the proposed WMFCC features over traditional MFCC features. (ii) We experimentally evaluate various topologies of hidden Markov model (HMM) classifiers using different spectral and prosody features to gain insight about possible temporal patterns existing in certain feature sets for emotion recognition from speech. (iii) We evaluate the use of decision fusion methods to combine various classifiers with uncorrelated features of emotional speech. It is well-known that in classification systems, data fusion is effective when modalities are correlated, and late fusion is optimal when modalities are uncorrelated (Sargin et al., 2007). Experimental results show that decision fusion of classifiers is beneficial.

We evaluate the proposed WMFCC features and the combined classifiers on the spontaneous emotional speech corpus FAU Aibo (Steidl, 2009), which is an elicited corpus with clearly defined testing and training partitions ensuring speaker-independence and different room acoustics as in real-life. We achieve significant performance improvements over the best scoring emotion recognition systems in the Interspeech 2009 Emotion Challenge (Schuller et al., 2009) with the proposed WMFCC features and the decision fusion of classifiers. Furthermore, we observe evidence of temporal formant patterns in discriminating emotion related classes of speech signal.

The remainder of this paper is structured as follows. Section 2 defines the components of the proposed emotion recognition system. The employed spectral and prosody features together with the proposed WMFCC features are presented in Section 2.1. Section 2.2 defines the HMM based classifier for emotion recognition, and Section 2.3 presents the decision fusion method for HMM based classifiers. Experiments to assess the performance of the proposed system are discussed in Section 3. Finally, the concluding remarks are presented in Section 4.

2. Proposed system

A block diagram of the proposed automatic speech driven emotion recognition system is given in Fig. 1. This system consists of three main blocks: feature extraction, classification and late fusion of classifiers. The feature extraction block computes prosodic and spectral features including the proposed WMFCC features. The classification block includes HMM based classifiers. HMM based classifiers with several states are capable of modeling temporal clusters, where each state can represent a different distribution of observations. We target to capture emotion related patterns in syntactically meaningful chunks of speech segments using HMM based classifiers. Syntacti-

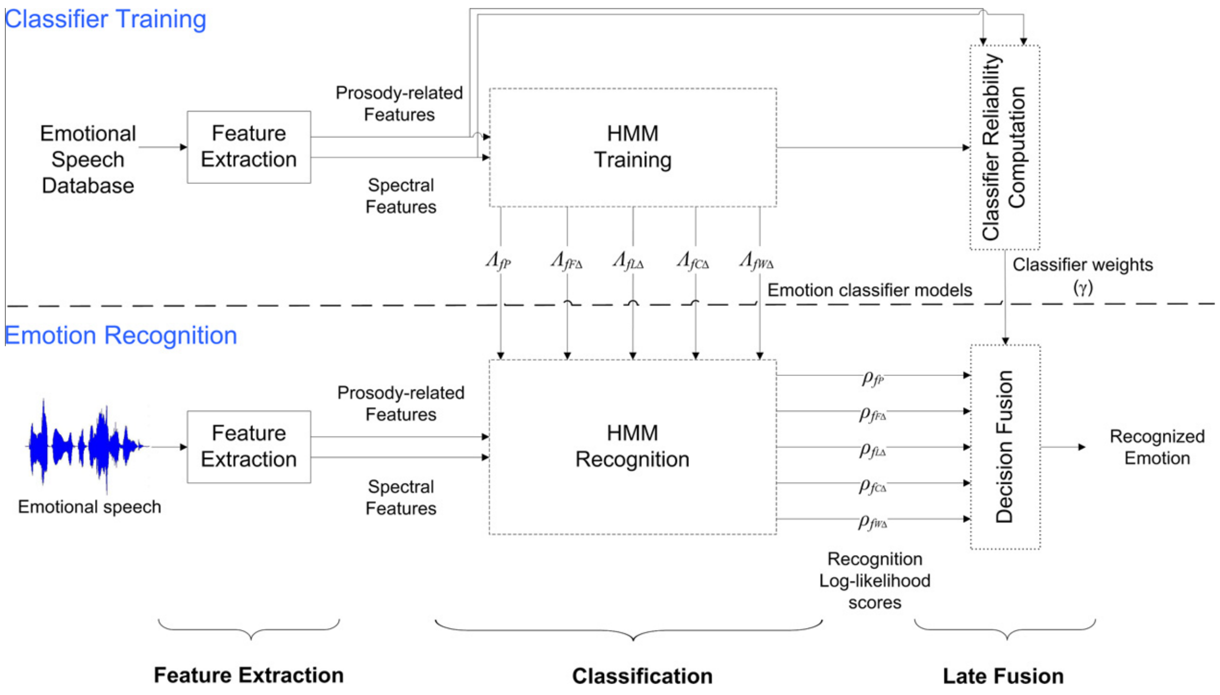

Fig. 1. Overview of the proposed emotion recognition system. The system is composed of classifier training and emotion recognition parts. Each spectral and prosody-related feature sequence, f, is used to train hidden Markov model sets, Λ_f, for all emotion classes. The highest log-likelihood scores, ρ_f, are evaluated through Viterbi decoding to be used in the decision fusion.

cally meaningful chunks of speech stands for a meaningful sequence of word(s) (Schuller et al., 2009). Finally, decision fusion is employed to benefit from different or uncorrelated feature sets. These three main blocks of the emotion recognition system are described in the following sections.

2.1. 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 consider only the acoustic content by using both prosody-related features and spectral features. The utilized features and the proposed formant position based weighted MFCC features are defined in the following sub-sections.

2.1.1. Prosody features

Voice characteristics at the prosodic level, including intonation, rhythm and intensity patterns, carry important clues for emotional states (Scherer, 1995). Hence, prosody features such as pitch and speech intensity can be used to model different emotions. For example, high values of pitch are correlated with happiness, anger, and fear, whereas sadness and boredom are associated with low pitch values (Scherer, 1995).

The pitch features for each syntactically meaningful chunk of speech are estimated for each 10 ms frame centered on 30 ms analysis window using the auto-correlation method (Deller et al., 1993). Since pitch values differ for each speaker and the system is desired to be speaker-independent, speaker normalization is applied. For each chunk of speech, we compute the mean pitch value over non-zero pitch values. Then, the mean pitch value is removed from the non-zero pitch values, which are computed for each frame. The regions between segments without a valid pitch (zero-value pitch segments) are filled with zero-mean and unit-variance Gaussian noise. No-pitch gaps creates deterministic constant observation segments, the filling of no-pitch gaps with Gaussian noise helps to maintain a proper training of statistical parameters of the HMM classifiers (Sargin et al., 2007). Then, pitch, first derivative of pitch and intensity values are used as normalized prosody features, which will be denoted as $\boldsymbol{f}_P$.

2.1.2. Formants

Formants are the resonant frequencies of the vocal tract filter. Accurate estimation of formant frequencies from speech is a challenging problem. State-of-the-art formant estimators locate candidate peaks of the spectra from short-time analysis of speech, and perform temporal tracking. In general, precision, robustness and computational efficiency are the main drawbacks of formant estimation. Goudbeek et al. (2009) reported that emotion has a considerable influence on formant positioning, especially on the placement of first two formants. Hence, we employ the first two formant frequencies for emotion recognition and represent them as $\boldsymbol{f}_F = [F1, F2]'$, where prime represents vector transpose. We also compute the first and second time derivatives of these two dimensional formant features using the following regression formula,

$$\Delta \boldsymbol{f}_F[n] = \frac{\sum_{k=-2}^{2} k \mathbf{f}_F[n+k]}{\sum_{k=-2}^{2} k^2}, \tag{1}$$

where $\boldsymbol{f}_F[n]$ is the formant feature vector at time frame n. Then we define the formant feature vector as $\boldsymbol{f}_{F\Delta} = [\boldsymbol{f}'_F \; \Delta\boldsymbol{f}'_F \; \Delta\Delta\boldsymbol{f}'_F]'$.

Formant frequencies are extracted using the PRAAT speech analysis software (Boersma and Weenink, 2010). Five formant frequencies are tracked and the maximum frequency of the highest formant is set to 8 kHz for all speakers. The time step between two consecutive analysis frames is selected as 10 ms within a 25 ms analysis window. Also, a high-pass filter with 50 Hz cut-off and 6 dB per octave tilt is used as pre-emphasis to compensate for spectral tilt.

2.1.3. MFCC features

The mel-frequency cepstral coefficient (MFCC) parametric representation is the most widely used spectral feature in automatic speech recognition. The MFCC features have also been used successfully in emotion recognition. We estimate the MFCC features using a 25 ms sliding Hamming window at intervals of 10 ms. We also include the log-energy, the first and the second time derivatives into the feature vector. The resulting dynamic feature vector is represented as: $\boldsymbol{f}_{C\Delta} = [\boldsymbol{f}'_C \; \Delta\boldsymbol{f}'_C \; \Delta\Delta\boldsymbol{f}'_C]'$.

2.1.4. LSF features

Line spectral frequency (LSF) representation of the linear prediction (LP) filter was introduced by Itakura (1975). While LSF features represent the spectral envelope, the localization of their angular values are closely related to formant frequencies. 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 pth order model for the vocal tract.

The LSF decomposition refers to expressing the pth 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,

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

The polynomials $P(z)$ and $Q(z)$ each have $p/2$ zeros on the unit circle, where phases of the zeros are interleaved in the interval $[0, \pi]$. Phases of the p zeros from the $P(z)$ and $Q(z)$ polynomials form the LSF feature representation for the LP model. Extraction of the LSF features, that is, estimation of the p zeros of the polynomials $P(z)$ and $Q(z)$, is also computationally effective and robust. The reader is referred to Itakura (1975) and Morris and

Clements (2002) for more theoretical information on LSF features.

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. It is well-known from this observation that a pair of LSF features are located at narrow-band formants (Morris and Clements, 2002). This relates LSF features to the location of narrow-band formants. This also implies that emotional influence on formant positioning can create measurable variations in the LSF features. Hence, similar to formant frequencies, the LSF features can be considered as candidate features for the emotion recognition task.

We represent the LSF feature vector, which is estimated over 20 ms frames centered on each 30 ms analysis window of speech with order $p = 16$, as $\boldsymbol{f}_L$. The LSF feature vector is extended to include the first and second time derivatives, and denoted as $\boldsymbol{f}_{L\Delta} = [\boldsymbol{f}'_L \ \Delta\boldsymbol{f}'_L \ \Delta\Delta\boldsymbol{f}'_L]'$.

2.1.5. Formant position based weighted MFCC features

In Section 2.1.4, we note that neighboring LSF features will be close to each other around formant frequencies. Using this fact, the inverse harmonic mean (IHM) weighting function was introduced for weighted quantization of LSF parameters (Laroia et al., 1991). The IHM weighting function is defined in (Laroia et al., 1991) as,

$$\omega_i = \begin{cases} \frac{1}{f_L^{i+1} - f_L^i} & i = 1, \\ \frac{1}{f_L^i - f_L^{i-1}} + \frac{1}{f_L^{i+1} - f_L^i} & i = 2, 3, \ldots, p-1, \\ \frac{1}{f_L^i - f_L^{i-1}} & i = p, \end{cases} \tag{3}$$

where f_L^i is the ith line spectral frequency for pth order filter and ω_i is the corresponding IHM weight. Note that the IHM weighting function is inversely proportional to the distance between neighboring LSF features. Hence, the IHM weight is expected to be higher when two consecutive LSF features are closely located, or equivalently when the LSF features are located at the narrow-band formants.

Furthermore, the IHM weights, ω_i, are associated with each LSF, f_L^i, for all the p LSF features. In order to define a weighting function for MFCC calculation, we define a mapping from the IHM weighting function to a weighting function for the mel-scaled critical bands. 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 the normalized IHM weightings,

$$v_i = \frac{\bar{\omega}_n \left(m_i - f_L^{n-1}\right) + \bar{\omega}_{n-1}\left(f_L^n - m_i\right)}{f_L^n - f_L^{n-1}} \quad i = 1, 2, \ldots, N_B, \tag{4}$$

where $\bar{\omega}_n$ is the normalized IHM weight, N_B is the number of critical bands, $f_L^{n-1} < m_i \leqslant f_L^n$ and the boundary line spectrum frequencies are defined as $f_L^0 = 0$ and $f_L^{p+1} = \pi$. We define the normalized critical band weights to retain a unity sum as,

$$\bar{v}_i = \frac{v_i}{\sum_j v_j} \quad i = 1, 2, \ldots, N_B \tag{5}$$

and the normalized IHM weights are defined as,

$$\bar{\omega}_i = \left(\frac{\omega_i}{\sum_j \omega_j}\right)^{\alpha} \quad i = 1, 2, \ldots, p, \tag{6}$$

where α is non-negative control parameter. Note that in logarithmic domain α becomes a constant multiplier to the IHM weights at all frequencies. As a result, the relative gain margin of critical band weights around formants gets larger for small α values, and gets lower for higher α values. Sample critical band weighting functions for different control parameter values are plotted in Fig. 2. The underlying speech frame has four visible formants, and the proposed critical band weighting function can successfully locate these formant positions.

Extraction of the WMFCC features is performed similar to the standard MFCC calculation and extraction steps are summarized in Fig. 3. Each analysis frame is first multiplied with a Hamming window and transformed to the frequency domain using the Fast Fourier Transform (FFT). Mel-scaled triangular filter-bank energies, e_i, which are located at critical band frequencies, m_i, are calculated using the square magnitude of the spectrum. Then, for each analysis frame, LSF feature vectors and corresponding normalized inverse harmonic mean weights are calculated, which are mapped to the critical band frequency values. Next, critical band energies are weighted by the normalized critical band weights and represented in logarithmic scale. The proposed WMFCC feature, f_W^j, is derived using discrete cosine transform (DCT)

$$f_W^j = \frac{1}{N_B} \sum_{i=1}^{N_B} \log(\bar{v}_i e_i) \cos\left((i - 0.5)\frac{j\pi}{N_B}\right), \quad j = 1, 2, \ldots, N, \tag{7}$$

where N is the number of WMFCC feature components that are extracted. Lastly, WMFCC vectors are extended to include the first and the second order derivatives, as well. Note that the proposed WMFCC feature vector is formed by an early data fusion of spectral content and LSF-derived formant location information. The control parameter, α, tunes the contribution of the LSF-derived formant location information in this data fusion.

A summary of the feature set representations of the proposed WMFCC feature together with spectral and prosody features are given in Table 1.

2.2. HMM-based classification

Hidden Markov models have been deployed with great success in automatic speech recognition to model temporal

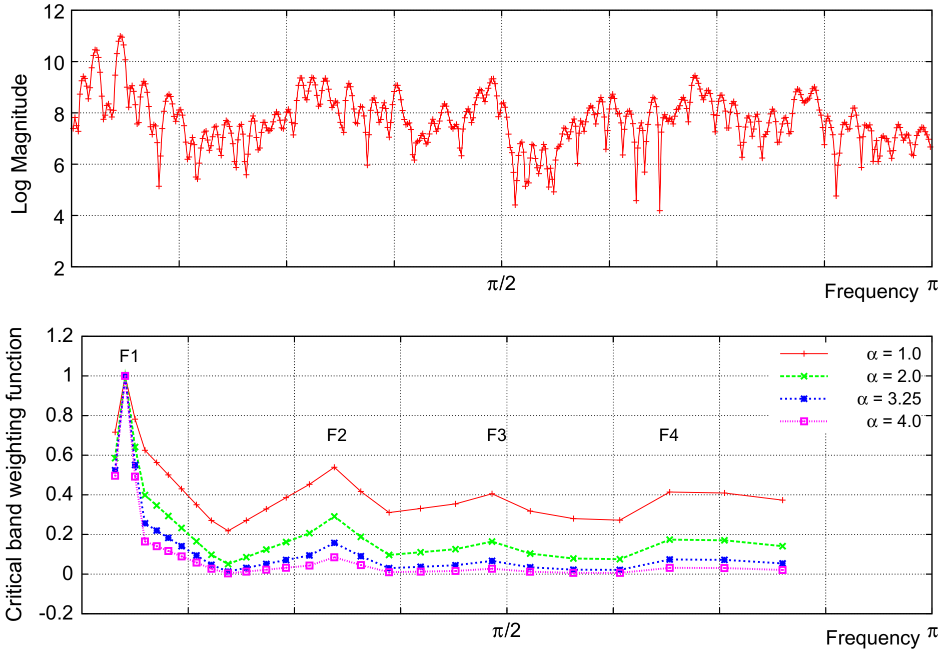

Fig. 2. (Top): The log-magnitude spectrum of a voiced speech frame. (Bottom): Sample inverse harmonic mean (IHM) based critical band weighting functions, $\bar{\omega}_i$, for various α values of the voiced speech frame given in the top sub-plot.

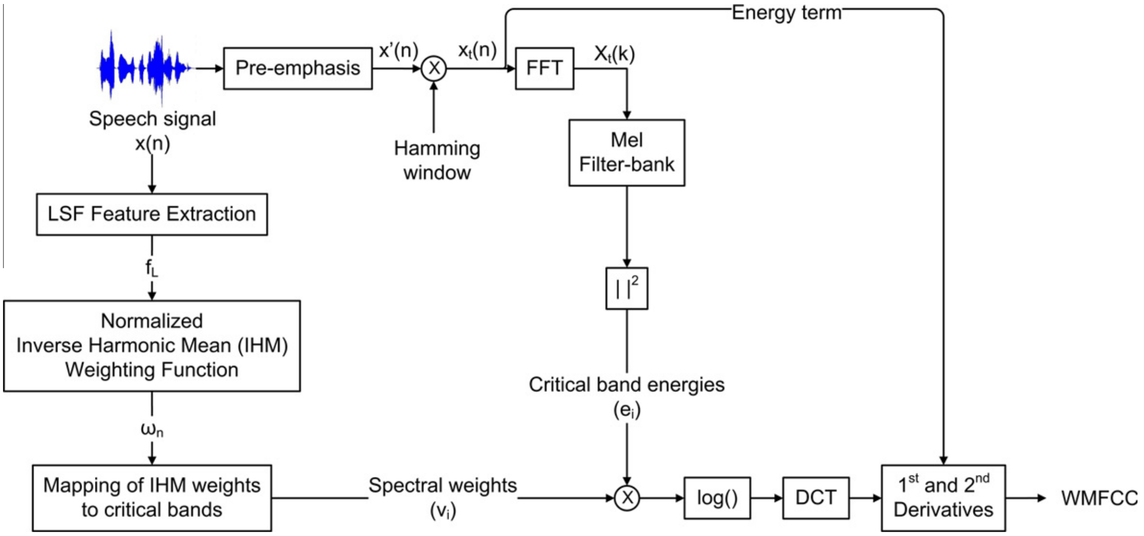

Fig. 3. Extraction of the proposed WMFCC features.

Table 1
Summary of the feature set representations.

f_P	Mean normalized pitch, pitch derivative and intensity
$f_{F\Delta}$	First two formants F1 and F2 with dynamic features
f_C	MFCC features
$f_{C\Delta}$	MFCC and dynamic features
f_L	LSF features
$f_{L\Delta}$	LSF and dynamic features
$f_{W\Delta}$	WMFCC and dynamic features

information in the speech spectrum, and they have been used similarly for emotion recognition as well (Schuller et al., 2003). We model the temporal patterns of the emotional speech utterances using HMM. We target to make a decision for a syntactically meaningful chunk of speech segment, where in each segment typically a single emotional evidence is expected. Furthermore, in each speech segment emotional evidence may exhibit temporal patterns. Hence, we employ N states left-to-right HMM, λ_e, (i.e., $a_{ij} = 0$ for $j \neq i, i+1$ where a_{ij} is the transition probability from state q_i to state q_j) to model emotion class e. Note that the set of HMMs with feature f for all emotions are represented as $\Lambda_f = \{\lambda_e\}$ in Fig. 1. Furthermore, feature observation probability distributions are modeled by M Gaussian mixture components with diagonal covariance matrices for each state. Topology parameters N and M

are determined through a model selection method and discussed in Section 3.

In the emotion recognition phase, the likelihood of a given speech segment is computed using a given HMM and the Viterbi algorithm for each emotion class as:

$$\rho(e) = \log P(\boldsymbol{f}|\lambda_e), \tag{8}$$

where $\rho(e)$ is the log-likelihood of the feature sequence $\boldsymbol{f}$ for the given HMM represented by λ_e. The utterance is classified as expressing the emotion ρ^*, which yields the highest likelihood score,

$$\rho^* = \arg\max_e \rho(e). \tag{9}$$

2.3. Decision fusion

Decision fusion is used to compensate for possible misclassification errors of a classifier based on one modality using other available modalities. In decision fusion, scores resulting from each unimodal classification are combined to arrive at a final conclusion, which is expected to give a more reliable overall decision. Decision fusion is especially effective when contributing modalities are not correlated and the resulting partial decisions are statistically independent.

We consider a weighted summation based decision fusion technique to combine different classifiers (Erzin et al., 2005; Kittler et al., 1998). The HMM 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,

$$\bar{\rho} = \left[1 + e^{-\left(\frac{\rho-\mu}{2\sigma}\right)}\right]^{-1}, \tag{10}$$

where μ and σ are the mean and the standard deviation of the log-likelihood ρ over emotion classes, respectively.

For two different feature sets, we have two HMM based classifiers with two likelihood score sets for each emotion. Let us denote these two normalized log-likelihoods as $\bar{\rho}_{f_1}(e)$ and $\bar{\rho}_{f_2}(e)$ for the emotion class e. The decision fusion then reduces to computing a single set of joint likelihood values, ρ_e, for each emotion class e. Assuming the two classifiers are statistically independent, we fuse the two classifiers, $f_1 \oplus f_2$, by computing the weighted average of the normalized likelihood scores,

$$\rho_e = \gamma\bar{\rho}_{f_1}(e) + (1-\gamma)\bar{\rho}_{f_2}(e), \tag{11}$$

where the value γ is independent of the emotion e and it is selected in the interval $[0, 1]$ to weight the contributions of the HMM classifiers. The γ weight is expected to be higher when the reliability of the first classifier with $\bar{\rho}_{f_1}(e)$ likelihood is higher, and vice versa (Erzin et al., 2005). In this study, we optimize the fusion weight γ over a subset of the training database to maximize the recognition rate.

3. Experimental results

In our experimental studies, we use the FAU Aibo corpus (Steidl, 2009), which is provided during the Interspeech 2009 Emotion Challenge (Schuller et al., 2009). The corpus consists of 9 h of German spontaneous emotional speech of 51 children (between ages 10–13), which is recorded while the children are interacting with Sony's pet robot Aibo. The audio recordings of the FAU Aibo corpus were sampled at a rate of 16 kHz, and have been manually segmented into syntactically meaningful chunks consisting of one or more words. A total of 18,216 chunks have been generated, each of which has been annotated with one of the five emotion class labels: anger, emphatic, neutral, positive or rest (non-neutral, but not belonging to the other categories). Another annotation was also carried out by choosing one of the two emotional classes: negative or idle (consisting of all non-negative states).

The number of instances are given in Tables 2 and 3, for the five-class and two-class tasks, respectively. In this study the training corpus of the Interspeech 2009 Emotion Challenge is split into two to form the training and development sets as given in these tables. The training and test sets of the FAU AIBO corpus have been recorded at two different schools, which provide speaker independence and different room acoustics as seen in most real life settings. Emotion class sizes in the FAU AIBO corpus are rather unbalanced. Therefore, the primary measure to evaluate the classification results is the unweighted average (UA) recall rate, which is the arithmetic average of the individual recall rates of each emotion class. The individual recall rate is defined as the proportion of utterances belonging to an emotion class which are correctly identified. In our experiments all the statistical classifiers are trained on the training set. The development set is used to optimize fusion weight γ, and the test set, as defined in the Interspeech 2009 Emotion Challenge, is used in the evaluation of the emotion recognition system.

3.1. Evaluation of the HMM classifiers

We first evaluate the topology of the HMM based classifiers by varying the number of states (N) and the number of mixture components per state (M) for various feature sets. Since the syntactically meaningful chunks in the FAU Aibo corpus are labeled with a single emotional class and are short in duration, we do not expect complex emotion-related temporal patterns in them. Therefore, we consider HMMs with up to 3-states to model emotion related temporal patterns, if they exist.

Two-class Emotion Recognition Results: The unweighted average (UA) recall rates for the two-class task are shown in Fig. 4. Fig. 4(d) shows the three UA rates for the prosody feature sets, where we observe an almost steady recall rate just under 62% with a 3-state HMM classifier and when the number of Gaussian mixtures per state is greater than 60. The 1-state and 2-state HMM classifiers on the

Table 2
FAU Aibo corpus instances for the five-class task.

	Anger	Emphatic	Neutral	Positive	Rest	Total
Train	571	1509	4338	400	547	7365
Develop	310	584	1252	274	174	2594
Test	611	1508	5377	215	546	8257

Table 3
FAU Aibo corpus instances for the two-class task.

	Negative	Idle	Total
Train	2359	5006	7365
Develop	999	1595	2594
Test	2465	5792	8257

other hand exhibit performance fluctuations around 60% UA recall rate. Performance of the formant features as given in Fig. 4(c) are located at the next tier between the 63% and 66% UA recall rate range. We observe that a 3-state HMM classifier for the formant features performs significantly better than a 1-state and a 2-state classifier.

Recall rates for the proposed WMFCC features and the spectral MFCC features are given respectively in Fig. 4(a) and (b), and they are observed at the top tier between the 66% and 70% UA rate range. Recall rate for the WMFCC feature is obtained using $\alpha = 3.1$, and its performance is in general higher than the performance of the MFCC features. Another observation is that, the recall rates of the 1-state HMM classifier using spectral features are low when the number of mixture components is low, but the recall rates increase when the number of Gaussian mixtures per state is increased. We observe that both a 1-state and a 2-state HMM classifier using spectral features perform better than a 3-state HMM classifier, when the number of mixture components is high. We should note that the recall rates for the LSF features are slightly lower than the recall rates of MFCC features, and they are not plotted in Fig. 4. However, they are presented and discussed in Section 3.3.

Five-class Emotion Recognition Results: The unweighted recall rates (UA) for the five-class emotion recognition task are given in Fig. 5. We observe that HMM classifiers using prosody and formant features have significantly lower recall rates than classifiers using spectral features. When spectral features (WMFCC or MFCC) are used, 1-state HMM classifiers have lower recall rates at low mixture components but the recall rates increase as the number of mixture components increase. Eventually, 1-state HMM classifiers perform better than or equivalent to 2- and 3-state HMM classifiers at high numbers of mixture components. The 3-state HMM classifier with the prosody features attains a steady recall rate for mixture components greater than 60 in Fig. 5(d) as observed with the two-class task. The 2-state and 1-state HMM classifiers catch up with this steady recall rate with the increased number of Gaussian components.

One can interpret the above results considering the total number of Gaussian mixture components per state in the HMM topology as follows. A 2-state HMM classifier has twice as many total number of mixture components as a 1-state HMM classifier. Hence, 2- and 3-state HMM classifiers have significantly higher total number of mixture components than a 1-state HMM classifier. At low numbers of mixture components per state, this creates an advantage for 2-state and 3-state HMM classifiers with spectral features. However, when the number of mixture components is sufficiently high, saturation occurs, and all HMM classifiers with spectral features perform similar. This observation indicates that no significant temporal pattern of emotional speech spectra, which we can model with a 2- or 3-state HMM topology, exists. On the other hand, this is not the case with formant features for the two-class task. The 3-state HMM classifier performs significantly better, which indicates possible temporal patterns for the formant features of emotional speech for the two-class task.

3.2. Performance of the proposed WMFCC features

We evaluate the proposed WMFCC features using HMM classifiers for a range of control parameter values, α, which were defined in (6). Fig. 6 presents the emotion recognition performance of the proposed WMFCC features for various α values together with the standard MFCC features. In the top plot, we see the UA recall rates for the two-class emotion recognition task using 1-state HMM classifiers with 112 and 120 mixture components per state. Recall rates of the standard MFCC features are shown with horizontal lines (since they do not depend on α) to set the baseline performances. We can see that the recall rates using the proposed WMFCC features are mostly higher than the recall rates using the standard MFCC features. The highest recall rate using WMFCC features is observed when α is around 3.1. The bottom plot in Fig. 6 presents UA recall rates of the five-class emotion recognition task using 2-state HMM classifiers with 80 and 96 mixture components.

We can observe that, for both the two-class and the five-class tasks, UA recall rates of the proposed WMFCC features are higher than the recall rates of the standard MFCC features for values of α in the [2, 4] interval. The highest recall rates with WMFCC features are obtained for α values in the [3, 4] interval.

3.3. Performance of unimodal classifiers and decision fusion

We compare the highest UA recall rates achieved with unimodal and decision fusion of HMM classifiers using spectral and prosody features for the two-class and five-

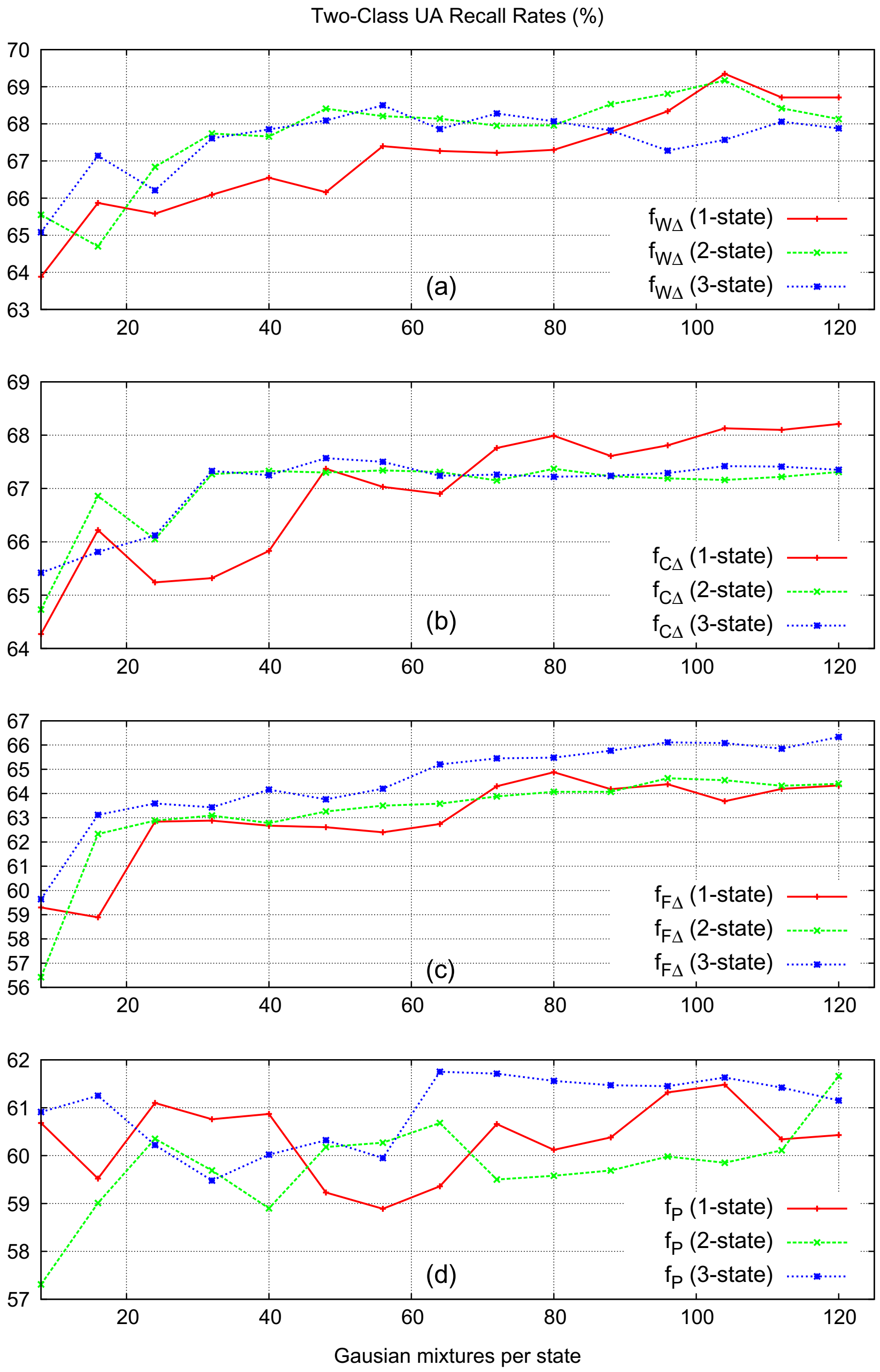

Fig. 4. Unweighted recall (UA) rates of HMM classifiers with different number of states and feature sets as a function of Gaussian mixtures per state for the two-class emotion recognition task. The feature sets include: (a) the proposed WMFCC features, (b) the spectral MFCC features, (c) the formant features, and (d) the prosody features.

class emotion recognition tasks as given in Tables 4 and 5, respectively.

Classifiers with the proposed WMFCC features score significantly higher than classifiers with the standard MFCC features for both the two-class and the five-class emotion recognition tasks. The highest UA recall rates achieved with the WMFCC features are 69.35% and 41.52% for the two-class and five-class tasks, respectively. In order to assess the statistical significance of the classifiers with the proposed WMFCC features and with the standard MFCC features, we performed McNemar's test (Dietterich, 1998), which is a paired success/failure trial using the binomial model. The McNemar's value for the five-class task is computed as 48.37 and for the two-class task as 193.03, which are both significantly greater than the statistical significance threshold $\chi^2_{(1,.95)} = 3.8414$.

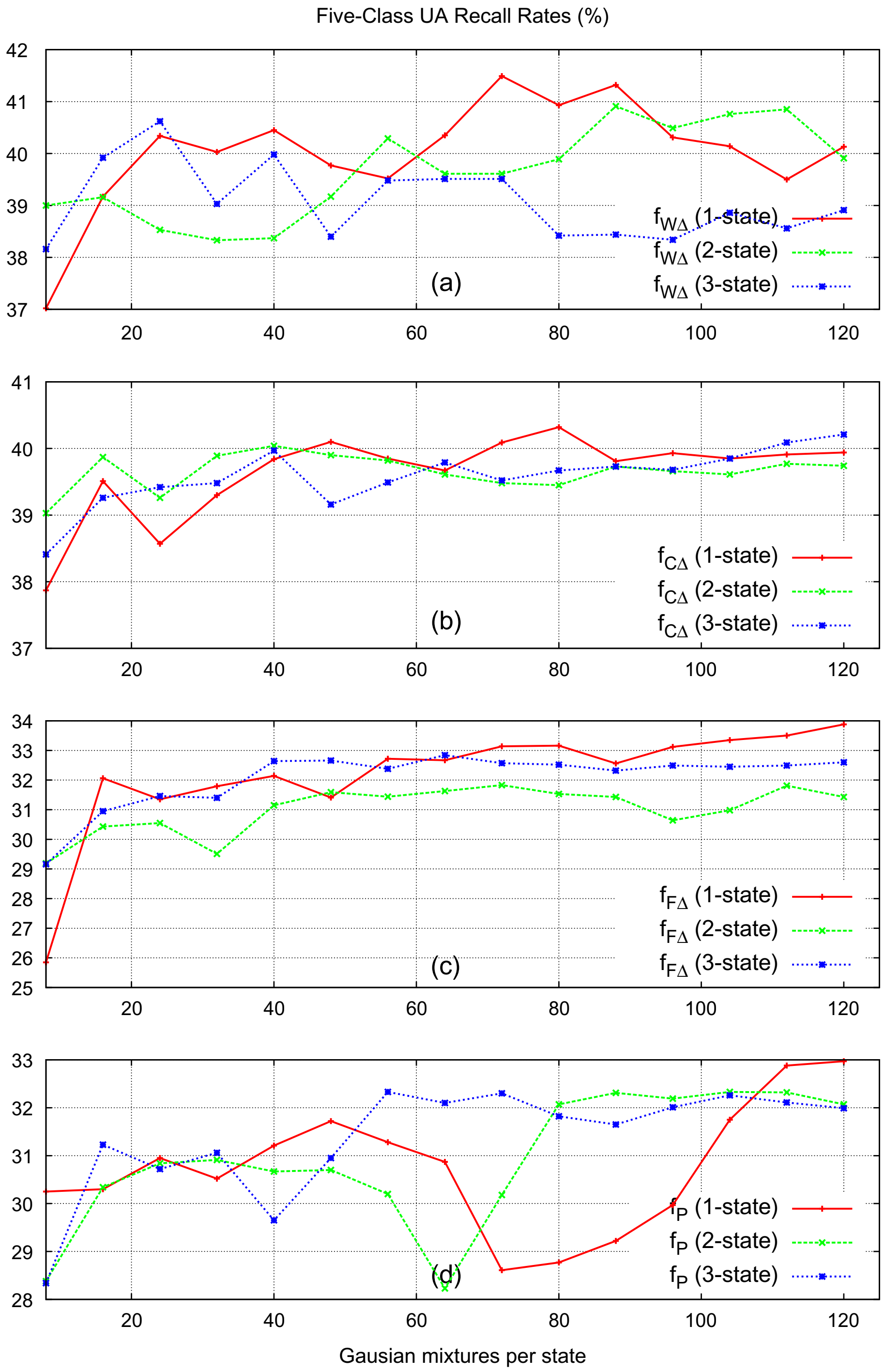

Fig. 5. Unweighted recall (UA) rates of HMM classifiers with different number of states and feature sets as a function of Gaussian mixtures per state for the five-class emotion recognition task. The feature sets include: (a) the proposed WMFCC features, (b) the spectral MFCC features, (c) the formant features, and (d) the prosody features.

In the five-class emotion recognition task (Table 5), HMM classifiers with prosody and formant features perform significantly lower (32.97% UA and 33.88% UA, respectively) than classifiers with spectral features (>39% UA). However, in the two-class task (Table 4), classifiers with formant features perform higher (66.33% UA) than the classifiers with LSF spectral features (65.06% UA). Hence, we can say that formant features are competitive for the two-class task.

Decision Fusion: Decision fusion of HMM classifiers is performed for various combinations of formant, MFCC, LSF and WMFCC features. The fusion weight, γ, is optimized over the development set prior to decision fusion evaluations on the test data. We observed that the fusion

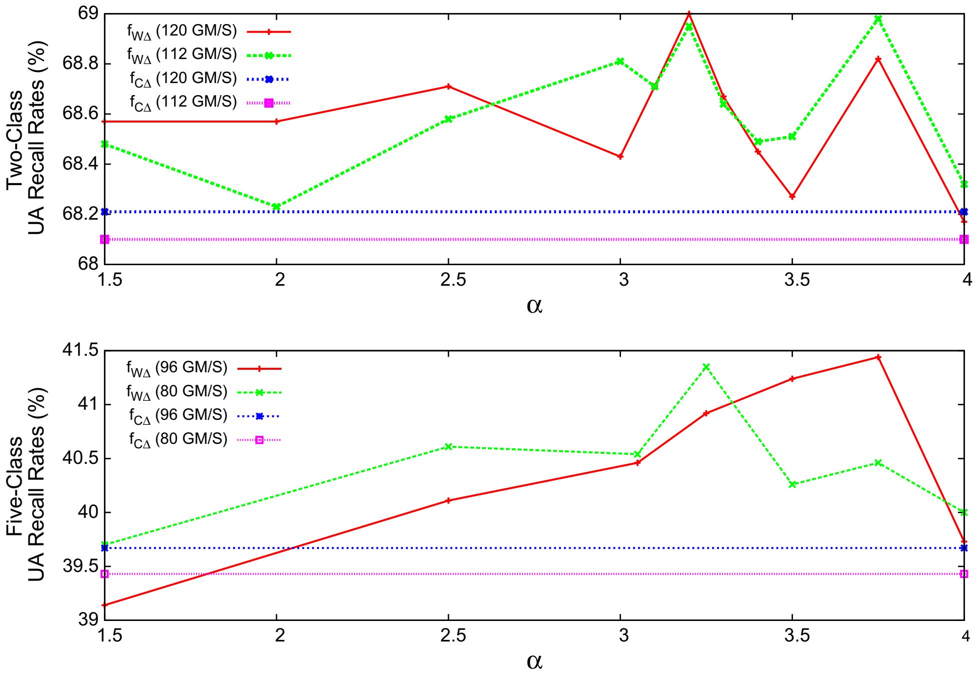

Fig. 6. Unweighted recall (UA) rates for the two-class and five-class emotion recognition tasks. The spectral features MFCC and the proposed WMFCC for various α values are evaluated with HMM classifiers with two different numbers of Gaussian mixture components per state.

Table 4
The highest two-class emotion recognition rates of unimodal classifiers and decision fusion of classifiers.

Unimodal/multimodal classifiers	UA recall rate (%)	γ
f_P	61.75	–
$f_{L\Delta}$	65.06	–
$f_{F\Delta}$	66.33	–
$f_{C\Delta}$	68.21	–
$f_{W\Delta}$	69.35	–
$f_{W\Delta} \oplus f_{C\Delta}$	70.06	0.86
$f_{W\Delta} \oplus f_{L\Delta}$	69.37	0.82
$f_{W\Delta} \oplus f_{F\Delta}$	70.32	0.84
$(f_{W\Delta} \oplus f_{C\Delta}) \oplus f_{F\Delta}$	70.55	0.72
Challenge Best (Dumouchel et al., 2009)	70.29	–

Table 5
The highest five-class emotion recognition rates of unimodal classifiers and decision fusion of classifiers.

Unimodal/multimodal classifiers	UA recall rate (%)	γ
f_P	32.97	–
$f_{F\Delta}$	33.88	–
$f_{L\Delta}$	39.17	–
$f_{C\Delta}$	40.32	–
$f_{W\Delta}$	41.52	–
$f_{W\Delta} \oplus f_{C\Delta}$	42.89	0.84
$f_{W\Delta} \oplus f_{L\Delta}$	43.09	0.98
$(f_{W\Delta} \oplus f_{C\Delta}) \oplus f_{L\Delta}$	43.59	0.91
Challenge best (Kockmann et al., 2009)	41.65	–

weight, γ, attains high values for both the two and the five-class tasks. This indicates that the first classifier in the fusion is weighted more, hence it is the more reliable classifier in the fusion. In this case, the classifiers with the proposed WMFCC features are found more reliable than the classifiers with competing feature sets at all decision fusion combinations. We can observe in Tables 4 and 5 that the recall rates show significant improvements after decision fusion for both the two-class and five-class tasks. Statistical significance of the experimental results can be investigated by modeling emotion recognition decisions as Bernoulli events. Then statistical significance of the experimental results can be measured with the expected standard deviation $\sqrt{p(1-p)/M}$, where p is the average recognition accuracy and M is the total number of decisions. The expected standard deviation of the recognition accuracy is 0.50% with an average recognition accuracy $p = 0.7$.

In the two-class task (see Table 4), decision fusion of the classifiers using the proposed WMFCC features and the classifiers using formant, MFCC and LSF features yield improvements over unimodal recall rates. The highest improvement is observed for the classifier fusion using WMFCC and formant features, which gives a recall rate of 70.32%. The decision fusion of three classifiers using WMFCC, MFCC and formant features yields the highest recall rate of 70.55%. These observations indicate that the WMFCC, MFCC and formant feature sets may contain some uncorrelated information, which appears as a performance improvement in the late fusion.

In the five-class task (see Table 5), decision fusion of two classifiers with the WMFCC and LSF features yields a recall rate of 43.09%, which shows a significant improvement over unimodal classifiers (max. 41.52%). Furthermore, decision fusion of three classifiers with WMFCC, MFCC and LSF features yields a recall rate of 43.59%. Inclusion of the LSF feature set brings improvements in

Table 6
Confusion matrix of the two-class task for the highest scoring decision fusion, $(\boldsymbol{f}_{W\Delta} \oplus \boldsymbol{f}_{C\Delta}) \oplus \boldsymbol{f}_{F\Delta}$.

	IDL	NEG	Total
Idle	3804	1988	5792
Negative	606	1859	2465

Table 7
Confusion matrix of the five-class task for the highest scoring decision fusion, $(\boldsymbol{f}_{W\Delta} \oplus \boldsymbol{f}_{C\Delta}) \oplus \boldsymbol{f}_{L\Delta}$.

	A	E	N	P	R	Total
Anger	312	175	62	18	44	611
Emphatic	190	906	299	39	74	1508
Neutral	579	1397	2519	449	433	5377
Positive	11	11	74	88	31	215
Rest	89	85	174	94	104	546

the late fusion. Similar to our conclusion for the two-class task, the above observations indicate that the WMFCC, MFCC and LSF feature sets contain some uncorrelated information for the five-class task.

The confusion matrices of the two-class and five-class tasks for the highest scoring decision fusion of classifiers are given in Tables 6 and 7, respectively. Negative emotion class in the two-class task consists of approximately 30% of the test data. The best classifier achieves recall rates of 75.4% and 65.7% for the negative and idle emotion classes, respectively. In the five-class task, neutral emotion class consists of approximately 65% of the test data. The best classifier achieves recall rates of 60.1%, 51.1%, 46.8%, 40.9% and 19.0% for the emphatic, anger, neutral, positive and rest emotion classes, respectively.

We can see from the above experiments that late decision fusion in general brings significant performance improvements to both the two-class and the five-class emotion recognition tasks. We should also note that the achieved recall rates are better than the highest recall rates reported at the Interspeech 2009 Emotion Challenge.

4. Conclusion

We introduced novel formant position based weighted mel-frequency cepstral coefficient (WMFCC) features for speech-driven emotion recognition. The spectral weighting is performed with the inverse harmonic mean function of LSF features, which is known to give more emphasis to the formant frequency regions. The experimental evaluations on the spontaneous emotional speech corpus FAU Aibo show statistically significant performance improvements with the WMFCC features as compared to the MFCC features. We also investigated decision fusion of classifiers, which yielded further improvements in recall rates. The achieved recall rates using decision fusion of classifiers are better than the highest recall rates reported at the Interspeech 2009 Emotion Challenge (Schuller et al., 2009).

We also evaluated different HMM classifiers with various spectral and prosodic feature sets for the classification of spontaneous emotional speech to gain insight about possible temporal patterns existing in emotional speech. We observed that a 1-state HMM classifier, that is the Gaussian mixture model (GMM) classifier, performs as good as 2- or 3-state HMM classifiers using spectral features with large number of mixtures. This suggests that, there are no significant temporal patterns of emotional speech spectra that can be modeled by 2- and 3-state HMM classifiers. However, a 3-state HMM classifier is observed to perform better with the formant features for the two-class task. This indicates existence of temporal formant patterns in discriminating negative and idle emotional classes in the two-class task. The formant features also contribute to improve recall rates with the decision fusion in two-class task. We should also note that the spontaneous nature of the FAU Aibo corpus creates a challenging classification problem for the emotion recognition task. This is a possible reason that the formant features do better for the two-class but worse for the five-class task. The two-class partition of the FAU Aibo corpus is observed to reduce the confusion between classes over the temporal formant patterns.

Future work on emotion recognition from speech could concentrate on integrating different clues from emotional speech, such as formants, spectral content, intonation, and linguistic content. Early data and/or late decision fusion of these clues are expected to improve the recall rates and deserve more investigation.

Acknowledgements

This work was supported in part by the Turkish Scientific and Technical Research Council (TUBITAK) under projects 106E201 (COST2102 action) and 110E056. The authors would like to acknowledge and thank the anonymous referees for their valuable comments that have significantly improved the quality of the paper.

References

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., 2003. How to find trouble in communication. Speech Comm. 40, 117–143.

Boersma, P., Weenink, D., 2010. Praat: doing phonetics by computer (version 5.2.01), http://www.praat.org/.

Deller, J., Hansen, J., Proakis, J., 1993. Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, New York.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 7 (10), 1895–1924.

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., Bouafaden, N., 2009. Cepstral and long-term features for emotion recognition. In: Interspeech 2009: 10th Annual Conference of the International Speech Communication Association 2009, Vols. 1–5.

Erzin, E., Yemez, Y., Tekalp, A.M., 2005. Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. IEEE Trans. Multimedia 7, 840–852.

Goudbeek, M.B., Goldman, J.P., Scherer, K., 2009. Emotion dimensions and formant position. In: Proc. Interspeech, Brighton, UK.

Grimm, M., Mower, E., Kroschel, K., S.Narayanan, 2006. Combining categorical and primitives-based emotion recognition. In: Proc. 14th Eur. Signal Process. Conf.

Itakura, F., 1975. Line spectrum representation of linear predictive coefficients of speech signals. J. Acoust. Soc. Amer. 57, S35.

Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20, 226–239.

Kockmann, M., Burget, L., Cernocky, J., 2009. Brno university of technology system for interspeech 2009 emotion challenge. In: Interspeech 2009: 10th Annual Conference of the International Speech Communication Association 2009, Vols. 1–5, pp. 316–319.

Laroia, R., Phamdo, N., Farvardin, N., 1991. Robust and efficient quantization of speech LSP parameters using structured vector quantizers. In: [Proc.] ICASSP 91: 1991 Internat. Conf. Acoust. Speech, and Signal Process. Toronto, Ont., Canada, pp. 641–644.

Lee, C., Yildirim, S., Bulut, M., Kazamzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. Emotion recognition based on phoneme classes. In: International Conference on Spoken Language Processing, Jeju Island.

Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13, 293–303.

Morris, R.W., Clements, M.A., 2002. Modification of formants in the line spectrum domain. IEEE Signal Process. Lett. 9, 19–21.

Morrison, D., Wang, R., Silva, L.C.D., 2007. Ensemble methods for spoken emotion recognition in call-centers. Speech Comm. 49, 98–112.

Nakatsu, R., Nicholson, J., Tosa, N., 2000. Emotion recognition and its applications to computer agents with spontaneous interactive capabilities. Knowledge-based Syst. 13, 497–504.

Neiberg, D., Elenius, K., 2008. Automatic recognition of anger in spontaneous speech. In: Interspeech (2008), ISCA, Brisbane, Australia. pp. 2755–2758.

Polzin, T., Waibel, A., 2000. Emotion-sensitive human–computer interfaces. In: Interspeech (2008), ISCA, Belfast.

Sargin, M., Yemez, Y., Erzin, E., Tekalp, A., 2007. Audiovisual synchronization and fusion using canonical correlation analysis. IEEE Trans. Multimedia 9, 1396–1403.

Scherer, K.R., 1995. How emotion is expressed in speech and singing. In: Proceedings of XIIIth International Congress of Phonetic Sciences, pp. 90–96.

Schuller, B., Lang, M., Rigoll, G., 2006. Recognition of spontaneous emotions by speech within automative environment. In: DAGA, Braunschweig. pp. 57–58.

Schuller, B., Rigoll, G., Lang, M., 2003. Hidden markov model based speech emotion recognition. In: Proc. Internat. Conf. Acoustics, Speech Signal Process. (ICASSP).

Schuller, B., Steidl, S., Batliner, A., 2009. The interspeech 2009 emotion challenge. In: Interspeech (2009), ISCA, Brighton, UK.

Steidl, S., 2009. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. publisherLogos Verlag, Berlin.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features, and methods. Speech Comm. 48, 1162–1181.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In: Proceedings of Affective Computing and Intelligent Interaction, Lisbon, Portugal, pp. 139–147.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31, 39–58.