Improving Automatic Emotion Recognition from Speech Signals

Elif Bozkurt[1]*, Engin Erzin*[1]*, Çiğdem Eroğlu Erdem*[2]*, A. Tanju Erdem*[3]

[1]College of Engineering, Koç University, Istanbul 34450, Turkey
[2]Bahçeşehir University, Department of Electrical and Electronics Engineering, Istanbul, Turkey
[3]Özyeğin University, Faculty of Engineering, Istanbul, Turkey

ebozkurt/eerzin@ku.edu.tr, cigdem.erdem@bahcesehir.edu.tr, tanju.erdem@ozyegin.edu.tr

Abstract

We present a speech signal driven emotion recognition system. Our system is trained and tested with the INTERSPEECH 2009 Emotion Challenge corpus, which includes spontaneous and emotionally rich recordings. The challenge includes classifier and feature sub-challenges with five-class and two-class classification problems. We investigate prosody related, spectral and HMM-based features for the evaluation of emotion recognition with Gaussian mixture model (GMM) based classifiers. Spectral features consist of mel-scale 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 intensity. Unsupervised training of HMM structures are employed to define prosody related temporal features for the emotion recognition problem. We also investigate data fusion of different features and decision fusion of different classifiers, which are not well studied for emotion recognition framework. Experimental results of automatic emotion recognition with the INTERSPEECH 2009 Emotion Challenge corpus are presented.

Index Terms: emotion recognition, prosody modeling

1. Introduction

Recognition of the emotional state of a person from the speech signal has been increasingly important, especially in human-computer interaction. There are recent studies exploring emotional content of speech for call center applications [1] or for developing toys that would advance human-toy interactions one step further by emotionally responding to humans [2]. In this young field of emotion recognition from voice, there is a lack of common databases and test-conditions for the evaluation of task 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 [3], and the VAM database consist of audio-visual recordings of German TV talk show with spontaneous and emotionally rich content [4]. The *INTERSPEECH 2009 Emotion Challenge* [5] avails spontaneous and emotionally rich the FAU Aibo Emotion Corpus for the classifier and feature sub-challenges.

In this study we investigate various spectral and prosody features, mixture of different features and fusion of different classifiers for the INTERSPEECH 2009 Emotion Challenge. In this investigation, we use GMM based emotion classifiers to model the color of spectral and prosody features, and HMM based emotion classifiers to model temporal emotional prosody patterns. Spectral features consist of mel-scale 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 are recently employed for emotion recognition, our investigation includes the following novelties: (i) we use LSF features, which are good candidates to model prosodic information since they are closely related to formant frequencies, (ii) we employ a novel multi-branch HMM structure to model temporal prosody patterns of emotion classes, and (iii) we investigate data fusion of different features and decision fusion of different classifiers.

2. Feature Representations for Emotion Recognition

In the emotional state classification of a speaker, we use both prosody-related and spectral features of voice.

2.1. Prosody Features

It is well known that for different emotional states, speech signal carries different prosodic patterns [6]. 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 [6].

The pitch features of the emotional speech are estimated using the auto-correlation method [7]. Since pitch values differ for each gender and the system is desired to 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. Then, pitch, pitch derivative, and intensity values are used as normalized prosody features, which will be denoted as $\mathbf{f}_P$.

2.2. Spectral Features

Similarly, the spectral features, such as mel-frequency cepstral coefficients (MFCC), are expected to model the varying nature of speech spectra under different emotions. The first and second derivatives of the prosody and spectral features are also included in the feature set to model the temporal dynamic changes in the speech signal. We consider the line spectral frequency (LSF) representation as an alternative spectral feature. The LSF representation was introduced by Itakura [8]. Since LSF features are closely related to formant frequencies, they are good candidates to model prosodic information in the speech spectra.

The spectral features of each analysis window are represented with a 13-dimensional MFCC vector consisting of energy and 12 cepstral coefficients and will be denoted as $\mathbf{f}_C$. The 16th order LSF feature vector $\mathbf{f}_L$ is also estimated for each anal-

ysis window.

2.3. 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,

$$\Delta \mathbf{f}(i) = \frac{\sum_{k=1}^{K} [\mathbf{f}(i+k) - \mathbf{f}(i-k)]k}{2\sum_{k=1}^{K} k^2} \quad (1)$$

where the number of analysis windows in the regression computation is set to $2K + 1 = 5$. The MFCC feature vector is extended to include the first and second order derivative features, and the resulting dynamic feature vector is represented as $\mathbf{f}_{C\Delta} = [\mathbf{f}'_C \ \Delta\mathbf{f}'_C \ \Delta\Delta\mathbf{f}'_C]'$, where prime represents vector transpose. Likewise, the LSF feature vector with dynamic features is denoted as $\mathbf{f}_{L\Delta}$. We also combine the pitch-intensity and the MFCC features to form the feature vector $\mathbf{f}_{PC}$, and when the first and second order derivatives of this combined feature are also included, we have the feature vector $\mathbf{f}_{PC\Delta}$ for non-zero pitch segments.

2.4. HMM-based Features

Speech signal carries different temporal prosody patterns for different emotional states. The HMM structures can be used to model temporal prosody patterns, hence they can be employed to extract emotion-dependent clues.

We employ 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. 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, $\lambda_1, \lambda_2, \ldots, \lambda_B$. Let us define the likelihood of a speech utterance U for the i-th branch HMM as,

$$p_i = P(U|\lambda_i). \quad (2)$$

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 $\mathbf{f}_H$,

$$f_H(i) = \left[1 + e^{-(\frac{p_i - \bar{p}}{2\sigma} + 1)}\right]^{-1}, \quad (3)$$

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 $\mathbf{f}_H$ is a B dimensional vector. We refer to two possible set of features $\mathbf{f}_{HP}$ and $\mathbf{f}_{HPC}$ when the multi-branch HMM is trained over $\mathbf{f}_P$ and $\mathbf{f}_{PC\Delta}$ features, respectively.

3. GMM-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

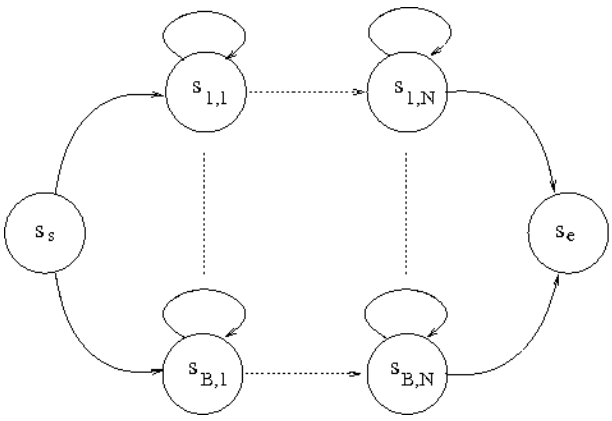

Figure 1: The multi-branch HMM structure.

given by

$$p(\mathbf{f}) = \sum_{k=1}^{K} \omega_k p(\mathbf{f}|k) \quad (4)$$

where $\mathbf{f}$ is the observation feature vector and ω_k is the mixture weight associated with the k-th Gaussian component. The weights satisfy the constraints,

$$0 \leq \omega_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{K} \omega_k = 1. \quad (5)$$

The conditional probability $p(\mathbf{f}|k)$ 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, $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$, let's define the log-likelihood of this utterance for emotion class e with a GMM density model γ_e as,

$$\rho_{\gamma_e} = \log p(\mathbf{F}|\gamma_e) = \sum_{t=1}^{T} \log p(\mathbf{f}_t|\gamma_e) \quad (6)$$

where $p(\mathbf{F}|\gamma_e)$ is the GMM probability density for the emotion class e as defined in (4). Then, the emotion GMM density that maximizes posterior probability of the utterance is set as the recognized emotion class,

$$\epsilon = \arg\max_{e \in E} \rho_{\gamma_e} \quad (7)$$

where E is the set of emotions and ϵ is the recognized emotion.

3.1. Decision Fusion

We consider a weighted summation based decision fusion technique to combine different classifiers [9]. 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 [9]. 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 $\bar{\rho}_{\gamma_e}$ and $\bar{\rho}_{\lambda_e}$ 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

$$\rho_e = \alpha\bar{\rho}_{\gamma_e} + (1-\alpha)\bar{\rho}_{\lambda_e} \quad (8)$$

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

4. EXPERIMENTAL RESULTS

We employ the FAU Aibo Emotion Corpus [10], which is distributed through the INTERSPEECH 2009 Emotion Challenge, in our experimental studies [5]. The FAU Aibo corpus includes clearly defined test and training partitions with speaker independence and different room acoustics. The recordings have a sampling rate of 16 kHz and they are processed over 20 msec frames centered on 30 msec windows for LSF features, and over 10 msec frames centered on 25 msec windows for all other features.

4.1. Evaluation of Classifiers

The challenge includes two different classification problems with five-class and two-class emotion classification targets. The five-class classification problem includes classes **A**nger (subsuming angry, touchy, and reprimanding), **E**mphatic, **N**eutral, **P**ositive (subsming motherese and joyful), and **R**est. Whereas, the two-class emotion classification task includes **NEG**ative (subsuming angry, touchy, reprimanding, and emphatic) and **IDL**e (all non-negative states).

All the feature sets as defined in Section 2 are used with the GMM based classifiers for the evaluation of emotion recognition. The GMM mixture components and the decision fusion parameter α are optimally selected to maximize emotion recall rate on a part of the training corpus. Recognition rates for the uni-modal GMM classifiers are given in Table 1. For the 2-class recognition problem f_L GMM and for the 5-class recognition problem $f_{PC\Delta}$ GMM classifiers have the highest accuracy as 65.25 % and 46.66 %, respectively.

Table 1: Emotion recognition rates with GMM based classifiers

Features	Recall [%]			
	2-class		5-class	
	UA	WA	UA	WA
$f_{C\Delta}$	66.36	62.09	39.94	41.29
$f_{L\Delta}$	66.05	60.24	39.10	41.78
f_L	63.36	65.25	33.68	40.39
$f_{PC\Delta}$	66.39	60.70	39.10	46.66

Decision fusion of different classifiers has been realized as defined in (8). The highest recognition rates for each decision fusion are listed in Table 2. Decision fusion of classifiers provides statistically significant improvement over unimodal classifiers. Among the decision fusion of GMM based classifiers, $\mathbf{f}_{PC\Delta}$ and $\mathbf{f}_{L\Delta}$ fusion yields the highest 5-class recognition rate, 47.83 %, with $\alpha = 0.57$. The confusion matrix of the decision fusion of these two GMM classifiers is given in Table 5. In addition, fusion of $\mathbf{f}_{C\Delta}$ and $\mathbf{f}_L$ has 64.44 % accuracy for the 2-class recognition problem when $\alpha = 0.64$.

Table 2: Emotion recognition rates after the decision fusion

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

Confusion matrices of different classifiers are given the following Tables 3-5.

4.2. Evaluation of Features

We employ 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. In the multi-branch HMM structure, branches are expected to capture temporal variations of different emotions. Then the branches are used to extract emotion dependent likelihoods, where we employ these likelihoods after a sigmoid normalization as the HMM-based features.

We experiment the HMM structure with different parameters by varying the 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 consider 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 train GMM classifiers using the HMM-based features. We observe that the f_{HPC} feature set with 3 states per branch and 12 Gaussian components per state yields the best results with a classification accuracy of 57.43 % and 27.48 % for 2 and 5-class classification respectively.

On the other hand, the decision fusion of two GMM classifiers with $f_{C\Delta}$ and $f_{L\Delta}$ features achieves 62.58 % and 43.71 % recognition rates for 2-class and 5-class classifications respectively. When we apply a second stage decision fusion to these results with HMM-based feature f_{HPC}, we obtain 63.03% and 44.17% recognition rates, respectively.

Table 3: 2-class confusion matrix of $\mathbf{f}_L$ GMM classifier

	NEG	**IDL**	**sum**
Negative	1446	1019	2465
IDLE	1850	3942	5792

Table 4: 5-class confusion matrix of $\mathbf{f}_{PC\Delta}$ GMM classifier

	A	**E**	**N**	**P**	**R**	**sum**
Anger	333	184	69	7	18	611
Emphatic	257	912	271	9	59	1508
Neutral	776	1601	2487	194	319	5377
Positive	21	16	105	43	30	215
Rest	118	116	181	53	78	546

Table 5: Confusion matrix of fusion of GMM classifiers with $\mathbf{f}_{PC\Delta}$ and $\mathbf{f}_{L\Delta}$ features

	A	**E**	**N**	**P**	**R**	**sum**
Anger	319	191	59	12	30	611
Emphatic	217	964	256	8	63	1508
Neutral	656	1638	2516	212	355	5377
Positive	19	18	94	50	34	215
Rest	105	110	185	46	100	546

5. Conclusions

We presented a speech-driven emotion recognition system for the INTERSPEECH 2009 Emotion Challenge. Emotion recognition is carried out using prosodic and spectral features, as well as the proposed HMM-based features, which are classified using GMM classifiers. Different feature and decision fusion strategies are tested. MFCC features perform better than prosody features since they capture rich spectral information. Similarly, the LSF features do well in emotion recognition. We also observed that the dynamic features improve overall recognition rates for all features.

The best 2-class and 5-class UA recall rates are achieved with the decision fusion of $f_{C\Delta}$, $f_{L\Delta}$ and f_{HPC} GMM classifiers at 67.90 % and 41.59 %, respectively. On the other hand the best WA recall for the 2-class recognition is observed as 65.25 % with the unimodal f_L GMM classifier. The best 5-class WA recall rate is achieved as 47.83 % with the decision fusion of $f_{PC\Delta}$ and $f_{L\Delta}$ GMM classifiers.

The feature sets that are consired for the emotion recognition task are observed to carry certain and not necessarily identical emotion clues. We observe some recognition improvements with the fusion of different classifiers. Extensive studies on the FAU Aibo Emotion Corpus or such natural emotional speech databases are needed for better modeling and evaluation of speech-driven emotion recognition systems. Furthermore, these databases should provide synchronous visual clues to enhance emotional face expression synthesis.

6. Acknowledgments

This work was supported in part by TUBITAK under projects 106E201 and TEYDEB 3070796, and COST2102 action. We would like to thank the INTERSPEECH 2009 Emotion Challenge team for their initiative and for kindly providing the challenge database and test results.

Table 6: Emotion recognition rates for the evaluation of features

HMM feature & Classifier Fusion	Recall [%] 5-class UA	5-class WA
$\gamma(f_{HP})$	24.56	21.30
$\gamma(f_{HPC})$	29.53	27.48
$\gamma(f_{C\Delta}) \oplus \gamma(f_{HP})$	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
$(\gamma(f_{C\Delta}) \oplus \gamma(f_{L\Delta})) \oplus \gamma(f_{HPC})$	41.59	44.17
	2-class	
$\gamma(f_{HPC})$	59.82	57.43
$(\gamma(f_{C\Delta}) \oplus \gamma(f_{L\Delta})) \oplus \gamma(f_{HPC})$	67.90	63.03

7. References

[1] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans.Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[2] P. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157–183, Jul. 2003.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520.

[4] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," 23 2008-April 26 2008, pp. 865–868.

[5] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech (2009), ISCA*, Brighton, UK, 2009.

[6] K. R. Scherer, "How emotion is expressed in speech and singing," in *Proceedings of XIIIth International Congress of Phonetic Sciences*.

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

[8] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of the Acoustical Society of America*, vol. 57, no. Suppl. 1, p. S35, 1975.

[9] E. Erzin, Y. Yemez, and A. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality realiability," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 840–852, Oct. 2005.

[10] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009.