Use of Line Spectral Frequencies for Emotion Recognition from Speech

Elif Bozkurt, Engin Erzin
Koç University
Istanbul, Turkey
ebozkurt/eerzin@ku.edu.tr

Çiğdem Eroğlu Erdem
Bahçeşehir University
Istanbul, Turkey
cigdem.eroglu@bahcesehir.edu.tr

A. Tanju Erdem
Özyeğin University
Istanbul, Turkey
tanju.erdem@ozyegin.edu.tr

Abstract

We propose the use of the line spectral frequency (LSF) features for emotion recognition from speech, which have not been been previously employed for emotion recognition. Spectral features such as mel-scaled cepstral coefficients have already been successfully used for the parameterization of speech signals for emotion recognition. The LSF features also offer a spectral representation for speech, moreover they carry intrinsic information on the formant structure as well, which are related to the emotional state of the speaker ***(buraya bir referans koyalim m?)****. We use the Gaussian mixture model (GMM) classifier architecture, that captures the static color of the spectral features. Experimental studies performed over the Berlin Emotional Speech Database and the FAU Aibo Emotion Corpus demonstrate that decision fusion configurations with LSF features bring a consistent improvement over the MFCC based emotion classification rates.*

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 or for developing toys that would advance human-toy interactions one step further by emotionally responding to humans. In this relatively new field of emotion recognition from speech, 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 [2]. Recently, the *INTERSPEECH 2009 Emotion Challenge* [7] avails spontaneous and emotionally rich FAU Aibo Emotion Corpus.

There are many studies in the recent literature on the problem of feature extraction and classifier design for emotion recognition [6, 9, 1]. Vlasenko, et. al. [9] aim to recognize emotions using a speaker recognition engine and introduce fusion of frame and turn-level based emotion recognition. They parameterize EMO-DB recordings with MFCC features and frame energy in addition to delta and acceleration coefficients. Turn-level analysis applies functionals to selected set of Low-Level-Descriptors (LLD) and their first-order delta coefficients. Speaker normalization and feature space optimization are applied prior to giving a final decision by SVM (support vector machines). This method assumes that emotion of the speaker does not change within one speaker turn, which may not always be valid. Schuller, et. al. [6] also study emotion recognition from speech using EMO-DB, yet under noise influence. They employ a large feature set including MFCC, spectral flux, and jitter features that is fed to a feature selection process. They classify the feature vectors with an SVM using 10-fold stratified cross validation.

In this study, we propose the use of LSF features together with the widely used MFCC features for emotion recognition. In this work, we use GMM based emotion classifiers to model the color of spectral features. Our method includes two main contributions: (i) use of LSF features, which are good candidates to model prosodic information since they are closely related to formant frequencies, and (ii) investigation of classifier fusion over different spectral features.

2 Speech-Driven Emotion Recognition

In the next subsections, we first present the speech features that we use for emotion classification. In particular, we propose the utilization of line spectral fre-

quency (LSF) features. Next, we present our emotion classification method using GMM classifiers. Finally, we discuss decision fusion of various classifiers to improve the emotion recognition performance.

2.1 Extraction of the Speech Features

In the following, we represent the spectral features of speech using mel-frequency cepstral coefficients (MFCC) and line spectrum frequency (LSF) features.

MFCC Features: Spectral features, such as mel-frequency cepstral coefficients (MFCC), are expected to model the varying nature of speech spectra under different emotions. We represent the spectral features of each analysis window of the speech data with a 13-dimensional MFCC vector consisting of energy and 12 cepstral coefficients, which will be denoted as $\mathbf{f}_C$.

LSF Features: Line spectrum frequency (LSF) decomposition has been first developed by Itakura [4] for robust representation of the coefficients of linear predictive (LP) speech models. The LSF features have not been previously used for emotion recognition from speech. In this paper we investigate their performance for emotion recognition.

LP 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. 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,

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

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 [5], and makes them good candidates to model emotion related prosodic information in the speech spectra. We represent the LSF feature vector of each analysis window of speech as a p dimensional vector $\mathbf{f}_L$.

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 the short-term spectra over time. The MFCC feature vector is extended to include the first and second order derivative features, and the resulting feature vector with dynamic components is represented as: $\mathbf{f}_{C\Delta} = \left[\ \mathbf{f}_C^T \quad \Delta\mathbf{f}_C^T \quad \Delta\Delta\mathbf{f}_C^T\ \right]^T$, where ^{T} is the vector transpose operator. Likewise, the extended LSF feature vector including dynamic components is denoted as $\mathbf{f}_{L\Delta}$.

2.2 Emotion Classification Using Gaussian Mixture Models

Gaussian Mixture Models (GMM) can be used to represent any continuous probability density function to a good approximation. GMMs have been successfully employed in many classification problems including emotion recognition [9]. In this paper, the probability density function of the feature space for each emotion is modeled with a GMM.

The representation of a probability density function under the GMM assumption is a weighted combination of M component Gaussian densities which is given by

$$p(\mathbf{f}) = \sum_{m=1}^{M} w_m p(\mathbf{f}|m) \tag{2}$$

where $\mathbf{f}$ is the feature vector and w_m is the mixture weight associated with the m-th Gaussian component. The weights are in $[0, 1]$ interval with their sum being equal to 1. The conditional probability $p(\mathbf{f}|m)$ is modeled by a Gaussian distribution with a component mean vector μ_m, and a diagonal covariance matrix Σ_m. The GMM for a given emotion is estimated with an expectation-maximization based iterative training process using the training set of feature vectors [10].

In the emotion recognition phase, the likelihood of the features of a given speech utterance is maximized over all emotion GMM densities. Suppose we are given a sequence of feature vectors for a speech utterance, $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$, where $\mathbf{f}_i$ is estimated from an analysis window of speech, and T is the total number of analysis windows in a speech utterance. Let us define the log-likelihood of this utterance for emotion class e using a GMM density model (γ_e) as,

$$\rho_{\gamma_e} = \log p(\mathbf{F}|\gamma_e) = \sum_{t=1}^{T} \log p(\mathbf{f}_t|\gamma_e), \tag{3}$$

where $p(\mathbf{f}_t|\gamma_e)$ is probability of feature $\mathbf{f}_t$ given the GMM-based probability density for the emotion class e as defined in (2). Then, the emotion that maximizes the class conditional log-likelihood probability of the utterance is selected as the recognized emotion:

$$e^* = \arg\max_{e \in E} \rho_{\gamma_e}, \tag{4}$$

where E is the set of all emotions and e^* is the recognized emotion.

2.3 Decision Fusion for Classification of Emotions

Decision fusion is used to compensate for possible misclassification errors resulting from a given modality classifier with other available modalities 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 aren't correlated and resulting partial decisions are statistically independent.

We consider a weighted summation based decision fusion technique to combine different classifiers [3] for emotion recognition. The GMM classifiers with MFCC and LSF features output likelihood scores for each emotion and utterance. The 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 $[0, 1]$ interval for all utterances [3]. After normalization, we have two likelihood score sets for the GMM classifiers for each emotion and utterance.

Let us denote normalized log-likelihoods of MFCC and LSF based GMM classifiers as $\bar{\rho}_{\gamma_e(C)}$ and $\bar{\rho}_{\gamma_e(L)}$ 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, which will be denoted by $\gamma_e(C) \oplus \gamma_e(L)$, by computing the weighted average of the normalized likelihood scores

$$\rho_e = \alpha \bar{\rho}_{\gamma_e(C)} + (1 - \alpha)\bar{\rho}_{\gamma_e(L)}, \qquad (5)$$

where the parameter α is selected in the interval $[0\ 1]$ to maximize the recognition rate on the training set.

3 Experimental Results

In our experiments, we use the Berlin Emotional Speech dataset (EMO-DB) [2] and the FAU Aibo Emotion Corpus [8]. The EMO-DB contains 5 male and 5 female speakers producing 10 German sentences used in everyday communication. These utterances simulate seven different emotions, namely **H**appiness, **A**nger, **S**adness, **F**ear, **B**oredom, **D**isgust and **N**eutral. We used a total of 535 emotional speech recordings (including several versions of same sentences), which have a sampling rate of 16 kHz. The FAU Aibo Emotion Corpus is recently distributed through the INTERSPEECH 2009 Emotion Challenge [7] to include clearly defined test and training partitions with speaker independence and different room acoustics. The FAU Aibo corpus investigates a five-class emotion classification problem that includes classes **A**nger (subsuming angry, touchy, and reprimanding), **E**mphatic, **N**eutral, **P**ositive (subsuming, motherese and joyful), and **R**est.

The speech data is processed over 20 msec frames centered on 30 msec windows to extract LSF features with order $p = 16$, and over 10 msec frames centered on 25 msec windows to estimate the MFCC features. The emotion recognition results over EMO-DB are extracted using 5-fold stratified cross validation (SCV). An emotion recognition decision is taken for each test utterance recording. The final recognition results are obtained as the average of the five test trials. The results on the FAU Aibo corpus are obtained based on the training and test partitions as defined in the INTERSPEECH 2009 Emotion Challenge.

The feature sets defined in Section 2.1 are used with GMM based classifiers for the evaluation of emotion recognition. Unweighted recall (UA) rates are presented in Table 1 for feature sets $\mathbf{f}_{C\Delta}$, $\mathbf{f}_{L\Delta}$ and $\mathbf{f}_L$. Unweighted recall rate is the arithmetic average of individual recall rates of each emotion class. The MFCC with dynamic components, $\mathbf{f}_{C\Delta}$, based classifier yields the best recall rates for EMO-DB and the FAU Aibo corpus. Note that, the EMO-DB database has significantly higher recall rates than the FAU Aibo corpus. This is because the EMO-DB is an acted database whereas the FAU Aibo is a spontaneous emotional speech database.

Table 1. Emotion recognition rates of unimodal and fusion of classifiers

Classifiers	UA Recall [%]	
	7-class EMO-DB	5-class FAU Aibo
$\gamma(f_{C\Delta})$	**82.98**	**39.94**
$\gamma(f_{L\Delta})$	80.01	39.10
$\gamma(f_L)$	78.95	33.68
$\gamma(f_{C\Delta}) \oplus \gamma(f_{L\Delta})$	84.27	**40.76**
$\gamma(f_{C\Delta}) \oplus \gamma(f_L)$	**84.58**	40.47

The recall rates for two possible decision fusion structures are listed in the last two rows of Table 1. The fusion parameter α is set to 0.57 over an independent training corpus. Fusion of the classifiers with $\mathbf{f}_{C\Delta}$ and $\mathbf{f}_L$ features yields the best recall rates for EMO-DB

Table 2. Confusion Matrix of $\gamma(\mathbf{f}_{C\Delta})/(\gamma(f_{C\Delta}) \oplus \gamma(f_L))$ over EMO-DB database

	F	D	H	B	N	S	A
Fear	**72.6/82.7**	**1.4/0**	**15.9/7.2**	0/1.4	**7.1/5.7**	0/0	2.8/2.8
Disgust	0/4.2	88.8/84.8	0/0	**2.2/0**	4.4/4.4	**4.4/0**	0/6.4
Happiness	**5.5/4.2**	0/0	**66.4/70.4**	0/0	0/1.3	0/0	**28/23.9**
Boredom	0/1.2	**1.3/0**	**1.3/1.2**	82.7/81.4	**9.7/8.6**	5.0/7.4	0/0
Neutral	0/1.2	**1.3/1.2**	0/0	10.2/11.5	88.5/86.0	0/0	0/0
Sadness	0/0	0/0	0/0	3.2/3.2	**5.0/3.3**	**91.7/93.4**	0/0
Anger	0.7/0.7	0/0	**9.4/6.2**	0/0	0/0	0/0	**89.8/93.0**

database, whereas the fusion with $\mathbf{f}_{L\Delta}$ attains higher recall rates for FAU Aibo corpus. It is observed that in both of the decision fusion configurations the LSF features bring a consistent improvement over the state of art MFCC based emotion classification rates for both of the databases. Hence, the LSF features indeed carry emotion related clues which are independent of spectral MFCC features.

The confusion matrices for the GMM classifier with MFCC features and their dynamic components, and for the decision fusion of the GMM and LSF based classifiers over the EMO-DB database are given in Table 2. Recall rate improvements are given in bold. We can observe from the table that, sadness has the highest recognition rate (91.7%) and anger follows it with a 89.8% recognition rate for the unimodal classifier. We can also observe that interestingly for the unimodal classifier, anger and happiness have large confusion rates: 9.4% of the time anger was classified as happiness and 28% of the time happiness was classified as anger. We observed that, with the decision fusion of $\gamma(f_{C\Delta})$ and $\gamma(f_L)$ classifiers, fear, happiness, sadness, and anger emotions have recall rate improvements. Furthermore, the misclassification rate of happiness as anger decreases from 28% to 23.9%, and likewise, misclassification rate of anger as happiness decreases from 9.4% to 6.2%.

4 Conclusions

In this paper, 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. We demonstrate through experimental results on two different emotional speech databases 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.

References

[1] E. Bozkurt, E. Erzin, C. E. Erdem, and T. Erdem. Improving automatic emotion recognition from speech signals. In *10th Annual Conf. Int. Speech Comm. Assoc. (INTERSPEECH)*, pages 324–327, Brighton, UK, Sep. 2009.

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

[3] E. Erzin, Y. Yemez, and A. M. Tekalp. Multimodal speaker identification using an adaptive classifier cascade based on modality realiability. *IEEE Transactions on Multimedia*, 7(5):840–852, Oct. 2005.

[4] F. Itakura. Line spectrum representation of linear predictive coefficients of speech signals. *Journal of the Acoustical Society of America*, 57(1):35, 1975.

[5] R. W. Morris and M. A. Clements. Modification of formants in the line spectrum domain. *IEEE Signal Processing Letters*, 9(1):19–21, January 2002.

[6] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. In *Speech Prosody*, Dresden, Germany, May 2006.

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

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

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

[10] G. Xuan, W. Zhang, and P. Chai. Em algorithms of gaussian mixture model and hidden markov model. In *International Conference on Image Processing (ICIP)*, volume 1, pages 145–148, 2001.