RECOGNIZING EMOTIONS IN SPEECH USING SHORT-TERM AND LONG-TERM FEATURES

Yang Li and Yunxin Zhao

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

ABSTRACT

The acoustic characteristics of speech are influenced by speakers' emotional status. In this study, we attempted to recognize the emotional status of individual speakers by using speech features that were extracted from short-time analysis frames as well as speech features that represented entire utterances. Principal component analysis was used to analyze the importance of individual features in representing emotional categories. Three classification methods including vector quantization, artificial neural networks and Gaussian mixture density model were used. Classifications using short-term features only, long-term features only and both short-term and long-term features were conducted. The best recognition performance of 62% accuracy was achieved by using the Gaussian mixture density method with both short-term and long-term features.

1. INTRODUCTION

A major task of intelligent human-machine interaction is to empower computers with the so called "affective computing" ability [1] such that a computer can recognize a user's emotional status and respond to the user in an affective way. Obviously, speech is one of the important communication channels between a computer and a user and it can be used for recognizing a user's emotional status.

In order to classify a talker's emotional condition from speech, emotion relevant features need to be extracted from the speech signal. In our study, the speech features were categorized as short-term features and long-term features. Short-term features, such as formants, formant bandwidths, pitch and log energy, reflect local speech characteristics in a short time window. Long-term features, such as mean of pitch, standard deviations of pitch, time envelopes of pitch and energy, reflect voice characteristics over a whole utterance. We recorded an emotional speech corpus of 5 speakers in 6 different emotions. Our study of the emotional speech corpus indicates that both short-term and long-term features vary with emotions and therefore both types of features can be used for emotion classification. The emotion recognition task was talker-dependent and text-independent. Three classification methods were used: vector quantization (VQ), Gaussian mixture density (GMD) models, and artificial neural network (ANN). Encouraging classification accuracy was achieved on the emotional speech corpus.

This paper is organized as follows, in section 2, the emotion speech corpus and the emotion features are described; in section 3, the classifiers are described; in section 4, classification results are presented and in section 5, a conclusion is given.

2. DATA PREPARATION AND FEATURE EXTRACTION

2.1. Data Preparation

The emotional speech corpus was collected from 5 untrained student volunteers (3 males and 2 females). Each speaker recorded 20 sentences in each of the following six emotions: "Neutral", "Happy", "Angry", "Fearful", "Surprised", and "Sad". The text contents of these sentences were written in a way to stimulate a speaker to speak in the specified emotions. For example, one of the happy sentence was: "My dad bought me a new sport car!" while a fearful sentence was: "There are some wolves in our backyard!" All the sentences had distinct contents. In addition, there were also three sentences that were spoken in all the six emotions so that comparison can be made across emotions. For each emotion, 15 sentences were used as the training set, and 5 sentences were used as the test set.

2.2. Short-term Feature Extraction

Feature analysis was made by applying a shifting short-time window to the speech sample sequences, where the window size was 50 msec and the step size was 20 msec. For each analysis frame, the following features were extracted: the first four formants, the first four formant bandwidths, pitch, log energy, the normalized first-order autocorrelation coefficient. Temporal differences of these features were also used to represent the speech dynamics. As indicated by previous researchers, [2][3][4][5] pitch is the most important indicator of emotional status. In our study, we found that formants also changed with the talkers' emotions.

Figure 1 shows the variation of the first four formants for the first phone segment /AO/ in the sentence "All of us studied hard for the exam" spoken in six emotions. Note that since the same sentence was read in the six emotions, the effect of contextual variation can be ruled out. As shown in Fig. 1, the third and fourth formants vary significantly with different emotions where as the first and second formants are relatively invariant.

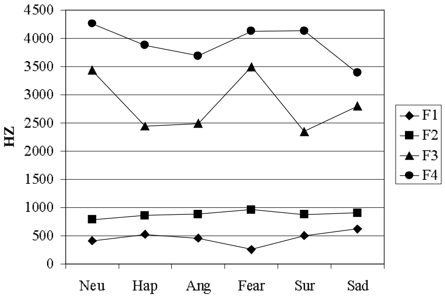

Figure 1: The formants of the vowel /AO/ in six different emotions.

2.3. Long-term Feature Extraction

The long-term features were extracted from the whole sequence of short-term features for each utterance, including mean of pitch, standard deviation of pitch, mean of log energy, standard deviation of log energy, as well as four indices that we introduce below for representing the time envelopes of pitch and energy. Referring a feature as either energy or pitch, the following average features are defined:

m_T -- average feature over the whole utterance
m_s -- average feature in the first half of the utterance
m_f – average feature in the second half of the utterance
m_1 – average feature in the beginning one third of the utterance
m_2 – average feature in the middle one third of the utterance
m_3 – average feature in the last one third of the utterance

In addition, we define f_i as the feature value in the frame i, and define f_i^* as the corresponding value of a fitting line for the feature envelope. The four indices were defined as:

Index-1 = $(ms – mf) / m_T$
Index-2 = $(2m_2 -(m_1+ m_3))/ 2m_T$
Index-3 = $(m_2- m_3)/ m_T$
Index-4 = $\sum_i (f_i - f_i^*)^2$

where Index-1 measures the increasing or decreasing trend of the feature sequences, Index-2 measures the bending shape of the envelope, Index-3 measures the final ending trend, and Index-4 measures the smoothness of the feature envelope.

2.4. Dimension Reduction and Feature Selection Using Principal Component Analysis (PCA)

Principal component analysis [7] was used for analyzing the importance of individual feature component in emotion classification. PCA examines the variance structure in the data and finds the directions in the feature space that have significant variations.

For the short-term features, the covariance matrices of emotional speech were first computed for individual phone units and then averaged to generate a single covariance matrix for PCA. Short-term features were selected according to the result of PCA so that the number of dimensions for short-term features was reduced from 22 to 12.

PCA can also be used for the long-term features. A study on the envelope indices (4 for pitch and 4 for energy) using PCA revealed that the six emotions in our corpus were somewhat paired: “Neutral” with “Sad”, “Happy” with “Surprised”, “Angry” with “Fearful”. Figure 2 shows the distribution of the indices in the two-dimension feature plane defined by the first two principal components.

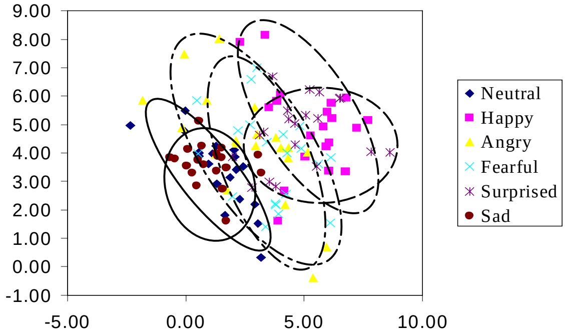

Figure 2: Envelope indices separate the six emotions to a certain extent. Emotions appear to be paired: “Neutral” with “Sad” (solid ellipse), “Happy” with “Surprised” (dash ellipse), “Angry” with “Fearful” (dash and dot).

Similar studies were made using PCA on the long-term features of mean of pitch, standard deviation of pitch, mean of log energy, standard deviation of log energy as well as using all the long-term features and the results are shown in Figure 3 and Figure 4 respectively.

It is clear that without the envelope indices, emotions can be separated easily into two general categories: unexcited (“Neutral” and “Sad”) and excited (“Happy”, “Angry”, “Fearful”, “Surprised”). Our envelope indices helped the classifier to further distinguish emotions within the excited categories and unexcited categories.

Long-term features were selected according to PCA so that the dimension was reduced from 12 to 8.

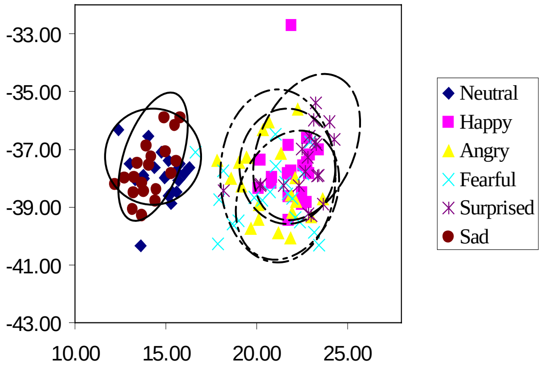

Figure 3: Mean and standard deviation of pitch and log energy projected into two-dimensional plane using PCA.

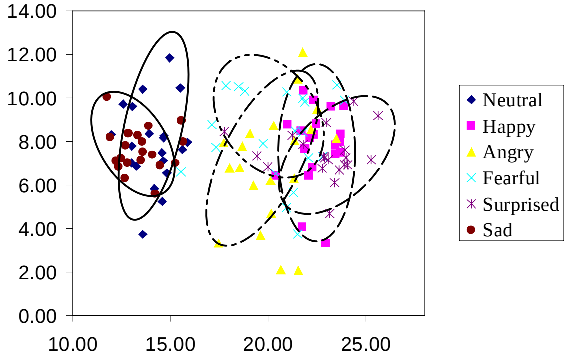

Figure 4: All long-term features were projected into two-dimensional plane using PCA.

3. CLASSIFICATION METHODS

Three types of classifier were used: VQ, ANN, GMD. In each method and for each emotion condition, 15 sentences were used as training data, and 5 sentences were used as test data.

In the VQ method, each short-term feature vector was concatenated by the long-term feature vector (repeated for each frame) extracted from the same sentence to form a long feature vector. Each feature component was first normalized by its standard deviation. Using the concatenated, normalized feature vectors, a codebook of size 8 was built for each emotional category. Classifications using only short-term features and using only long-term features were also conducted for comparison.

In the ANN method, a fully connected feed-forward 3-layer neural network was used. The same features as those used in the VQ method were used in ANN. Classifications using only short-term features and using only long-term features were also conducted. The number of neurons in the input layer was the same as the number of features. The output layer had 6 neurons with each one standing for a specific emotion. There were also 6 neurons in the hidden layer. Back propagation was used for training. For a specific emotion, the target value of the output neuron responsible for that emotion was 1, while the others were -1. In this way, errors used for back propagation could be calculated. In testing, the neuron in the output layer that had the largest value was picked as the classification result.

In the GMD method, for each emotion category, a Gaussain mixture model of size 4 was built for the short-term features and a Gaussain model was built for the long-term features. Classification results were then obtained by combining the likelihood scores from the two. The weighting of the two likelihood features were manually tuned.

4. EXPERIMENT RESULTS

The classification results were averaged over the 5 speakers and are summarized in Figure 5 for VQ, GMD, ANN using short-term features only, long-term features only, and both short-term and long-term features.

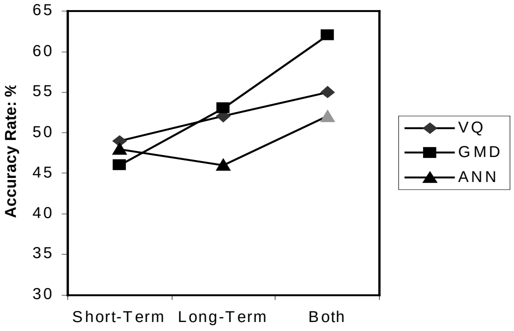

Figure 5: Emotion Classification Results

The best result was obtained using GMD with both long-term and short-term features. The detailed classification results for using the GMD method are given in Tables 1, 2 and 3.

	Neu.	Hap.	Ang.	Fear.	Sur.	Sad
Neu	0.4	0.1	0	0.2	0	0.3
Hap	0.05	0.6	0.1	0	0.25	0
Ang	0.05	0.2	0.45	0.1	0.2	0
Fear	0.25	0.05	0	0.4	0.1	0.2
Sur.	0	0.35	0.3	0.1	0.25	0
Sad	0.1	0	0.1	0.15	0	0.65

Table 1: GMD Results Using Short-term Features Only

	Neu.	Hap.	Ang.	Fear.	Sur.	Sad
Neu	0.4	0.1	0	0.25	0	0.25
Hap	0	0.6	0	0.05	0.35	0
Ang	0.05	0.1	0.5	0.05	0.1	0.2
Fear	0.25	0.05	0	0.35	0.05	0.3
Sur.	0	0.3	0.05	0.25	0.4	0
Sad	0.05	0.05	0	0	0	0.9

Table 2: GMD Results Using Long-term Features Only

	Neu.	Hap.	Ang.	Fear.	Sur.	Sad
Neu	0.45	0.1	0	0.2	0	0.25
Hap	0.05	0.85	0	0	0.1	0
Ang	0.05	0.1	0.5	0.05	0.1	0.2
Fear	0.2	0.05	0	0.45	0.05	0.25
Sur.	0	0.25	0.05	0.15	0.55	0
Sad	0.05	0.05	0	0	0	0.9

Table 3: GMD Results Using Both Short-term and Long-term Features

5. DISCUSSION

Although GMD achieved the best overall results and the best result for the long-term features, it did not perform well when only short-term features were used. Since the short-term features in general depend on the phonetic category of speech, they require more training data to reliably train the classifiers. Because VQ is basically a point-matching method, it is not strange that VQ performed better than GMD in this case. Comparing the results of Table 1, 2, and 3, we observe that the short-term features help to reduce the confusion between the emotional statuses of “happy” and “surprised”, as well as the confusion among “neutral”, “fearful” and “sad”.

During this study, we found the quality of training sentences to be very important. Though our speakers tried their best in acting different emotions, the resulted emotional speech were still not quite distinguishable for some emotions. Among the six emotions, “fearful” seems to be the most difficult one to act. Also, “surprised” and “happy” were often hard to tell by human listeners, which may reflect the real world situation.

In conclusion, our study demonstrates that it is possible to identify speakers’ emotions from speech for a limited number of emotional categories. Though certain emotions overlap with others, applications that require classification of only a small number of emotions are feasible.

ACKOWLEGEMENT

This research is part of the computer companion project conducted in the Beckman Institute of University of Illinois at Urbana-Champaign, funded by Yamaha Corporation. We appreciate the support from Professor G. McConkie and Professor T. Huang.

REFERENCES

1. R. W. Picard, "Affective Computing", The MIT Press, Cambridge, 1997

2. H. Levin and W. Lord, "Speech pitch frequency as an emotional state indicator," IEEE Trans. Systems, Man, and Cybernetics, vol. SMC-5, No.2, pp. 259-273, March 1975

3. L. A. Streeter, N. H. Macdonald, R. M. Krauss, W. Apple, K. M. Galotti, "Acoustic and perceptual indicators of emotional stree," J. Acoust. Soc. Am., vol.73, No.4, pp.1354-1360, April 1983

4. P. Lieberman, S. B. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," J. Acoust. Soc. Am., vol. 34, No. 7, pp. 992-927, July 1962

5. C. E. Williams, K. N. Stevens, "Emotions and speech: Some acoustical correlates", J. Acoust. Soc. Am., vol. 52, No.4 (part 2), pp. 1238 - 1250, April 1972

6. J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," Proc. Int. Conf. On Spoken Language Processing, vol3/SaA2S2, pp. 1267-1270

7. Everitt and Dunn, “Applied Multivariate Data Analysis”, Oxford University Press, New York, 1992

8. M. Slaney and G. McRoberts, “Baby ears: A recognition system for affective vocalizations,” Proc. IEEE Int. Conf. Acoustics. Speech, and Signal processing, Seattle, WA, 1998, pp. 985-988