RECOGNIZING SIX EMOTIONAL STATES USING SPEECH SIGNALS

Bong-Seok Kang, Chul-Hee Han, Dae-Hee Youn, and Chungyong Lee

Department of Electrical and Computer Engineering
Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, KOREA
Email: kbs@radar.yonsei.ac.kr

ABSTRACT

This paper examines three algorithms to recognize speaker's emotion using the speech signals. Target emotions are happiness, sadness, anger, fear, boredom and neutral state. MLB(Maximum-Likelihood Bayes), NN(Nearest Neighbor) and HMM (Hidden Markov Model) algorithms are used as the pattern matching techniques. In all cases, pitch and energy are used as the features. The feature vectors for MLB and NN are composed of pitch mean, pitch standard deviation, energy mean, energy standard deviation, etc. For HMM, vectors of delta pitch with delta-delta pitch and delta energy with delta-delta energy are used. We recorded a corpus of emotional speech data and performed the subjective evaluation for the data. The subjective recognition result was 56% and was compared with the classifiers' recognition rates. MLB, NN, and HMM classifiers achieved recognition rates of 68.9%, 69.3%, and 89.1%, respectively, for the speaker dependent, and context-independent classification.

1. INTRODUCTION

The four most commonly accepted primary emotions of the human beings are happiness, sadness, anger, and fear[1][2][3][4]. We can recognize the emotional states through the facial expression, the speech, the gestures, the heart rate, the temperature and the blood pressure and so on[1]. Especially, the speech plays an important role in communicating human emotions. There are many researches on the correlates between the speech and the emotion. In 1972, Williams and Stevens found that emotions have several effects on the fundamental frequency contour of a speech[2]. In 1993, Murray compiled of a lot of studies made so far on emotions. Murray described that the most commonly referenced vocal parameters in the emotion literature are pitch, duration, intensity, and voice quality. He also noted that the acoustic correlates of primary emotions are cross- cultural[3].

This work is supported by KRISS (Korea Research Institute of Standards and Science).

Based on these researches, some recent studies suggest several methods for emotion recognition. Deb Roy used MLB (Maximum-Likelihood Bayes) after preprocessing by the Fisher linear discriminant method[4]. Frank Dellaert used MLB, KR(Kernel Regression), and KNN(K-nearest neighbors) with feature selection techniques[5]. They commonly used some averaged values of all speech frames. For the emotion recognition, we made the speech database containing speaker's emotions. We extracted pitch and energy from each frame of speech signals. Using these pitch and energy, we apply MLB, NN, and HMM to the emotion recognition.

The database construction procedure is described in Section 2. In Section 3 the feature extraction and selection from the emotional speech is presented. Section 4 explains methods to apply three pattern recognition algorithms to the emotional speech recognition. The experimental results are described in Section 5. Finally we summarize our major findings and results in Section 6.

2. DATABASE CONSTRUCTION

To experiment emotion recognition, we built an emotional speech database. Target emotions are happiness, sadness, anger, fear, boredom and neutral state. Total 8 speakers, 4 men and 4 women, who are amateur announcers of YBS(Yonsei Broadcasting System) were involved. We selected the five Korean words which have no emotional bias. The speakers uttered these words with simulated six emotions. They watched the edited movie to induce natural emotions. We prepared the movie scenes that arouse emotions of happiness, sadness, anger, or fear. We picked the movie scenes by surveying. If the speakers wanted something to arouse emotions, we played these ready-made scenes. The details of recording process are as follows:

In the first step, we let the speakers utter the five object words with one emotion. They repeated them 13 times with 5 second-interval. After done, we continued the

recording with the next emotions. Finally, we recorded total 3120 words.

Recording was performed in the studio of Yonsei University. DAT(Digital Audio Tape) was interconnected with the console of outer studio. For our experiment, the recorded data were sampled by16kHz.

3. FEATURES FOR EMOTION RECOGNITION

Emotion affects pitch, energy, and speech rate of the speaker [6]. When speaker is happy or angry, pitch and energy are high, and speech rate increases[3]. When speaker is sad or bored, pitch and energy are low, and speech rate decreases. Figure 1 shows the energy levels in the subinterval of speech signals. Figure 2 gives the pitch levels.

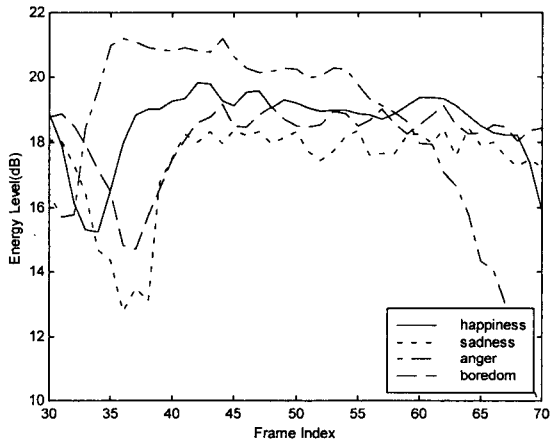

Figure 1. Energy Levels of a syllable, /-ger-/.

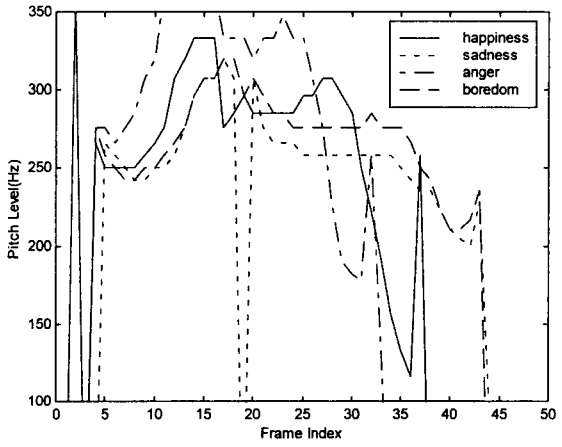

Figure 2. Pitch Levels of a syllable, /-ger-/.

As figures show, the emotions of happiness, sadness, anger, and boredom are different in energy levels and pitch. Based on these facts, we selected energy mean, energy max, pitch mean, and pitch max as the features. Figures also show that emotions vary in the pitch and energy ranges. Based on this fact, we considered pitch standard deviation and energy standard deviation. These 6 features were used as input parameters for MLB and NN classifiers.

We also noted that pitch and energy contours for the same syllable are various in the different emotions. Especially this is distinct in the Figure 2. Therefore, we suggest two feature vectors to reflect this characteristic. One is delta pitch with delta-delta pitch, and the other is delta energy with delta-delta energy. These feature vectors are useful to distinguish emotional states with neighboring pitch mean and energy mean such as happiness and anger, or sadness and boredom.

We computed pitch and energy in each frame. Log energy is used to estimate energy, and the pitch extraction algorithm in EVRC (IS-127) is applied to estimate pitch[7]. In our experiment, sampled speech signals with 16KHz sampling frequency are analyzed using 20ms Hamming window with 50% overlap. Energy and pitch are computed in each frame.

4. PATTERN RECOGNITION ALGORITHM

Considered are emotion recognition systems that are speaker-dependent and context-independent. We made the training data set with 300 utterances for each person and the testing data set with 90 utterances.

4.1 MLB (Maximum-Likelihood Bayes) Algorithm

Under the assumption that the probability distribution function of each feature (i.e. pitch mean, pitch max, pitch standard deviation, energy mean, energy max, and energy standard deviation) is Gaussian, we determined the decision boundary by the Bayes decision method[8] using the training data set. After we had tested recognition performance for each feature with the testing data set, we applied the Bayes decision method for all combinations of six features to find the optimal feature set. As a result, final feature set was composed of pitch mean, pitch standard deviation, pitch max, energy standard deviation, and energy max.

4.2 NN(Nearest Neighbor) Algorithm

After we generated ten reference patterns per each emotion using training data set, we used Nearest

Neighbor method[9] was applied to find an optimal combination of all six features. The reference patterns were produced using LBG clustering algorithm[10] and Euclidean distance was exploited for the distance measure. We computed the distances between reference patterns and unknown input. The class including the nearest reference pattern is selected. In this case, the optimal feature set consisted of pitch mean, pitch standard deviation, pitch max, energy mean, and energy max.

4.3 HMM(Hidden Markov Model) Algorithm

The emotions were modeled by using 8 states. Discrete HMM of left-to-right structure was used[11]. Two feature vectors, delta energy with delta-delta energy and delta pitch with delta-delta pitch, were considered. The codebook size was 64 for the energy and 256 for the pitch.

5. EXPERIMENTAL RESULTS

5.1 Subjective Evaluation Results

We have conducted a subjective evaluation by untrained listeners in order to estimate the quality of the constructed data, and then the result were compared with the recognition performance of our emotion recognition system. The recognition accuracy rate of the subjective evaluation was 56.0%. The resulting confusion matrix is shown in Table 1. From the table, it can be seen that happiness and anger are confusing as well as sadness and boredom.

Table 1. Subjective Evaluation Results [%], where N, H, S, A, F, and B indicate Neutral, Happiness, Sadness, Anger, Fear, and Boredom, respectively.

Class	N	H	S	A	F	B	Accuracy rate
Neutral	65.6	0.6	0.8	28.6	0.4	4.4	66.6
Happiness	28.4	39.0	2.2	23.6	1.8	5.0	39.0
Sadness	18.4	0.6	44.0	4.5	13.7	18.8	44.0
Anger	6.2	8.4	1.2	81.8	0.7	1.7	81.8
Fear	18.1	2.7	14.4	5.9	53.0	5.8	53.0
Boredom	12.6	0.9	17.3	11.3	5.3	52.6	52.6
total							56.0

5.2 The recognition results by MLB classifier

When MLB classifier was applied, the recognition accuracy rate was 68.9%. The resulting confusion matrix is shown in Table 2. As the result of subjective evaluations, happiness and anger are confusing as well as sadness and boredom.

Table 2. The recognition results by MLB classifier [%], where N, H, S, A, F, and B indicate Neutral, Happiness, Sadness, Anger, Fear, and Boredom, respectively.

Class	N	H	S	A	F	B	Accuracy rate
Neutral	73.8	2.1	5.5	3.5	0.7	14.5	73.8
Happiness	2.7	61.3	2.0	15.3	6.0	12.7	61.3
Sadness	4.0	2.7	52.7	4.0	12.7	24.0	52.7
Anger	2.0	4.7	1.3	89.3	0.7	2.0	89.3
Fear	2.0	4.0	9.33	6.0	70.7	8.0	70.7
Boredom	3.7	3.7	11.9	5.2	9.7	65.7	65.7
Total							68.9

5.3 The recognition results by NN classifier

When NN classifier was applied, the recognition accuracy rate was 69.3%. The resulting confusion matrix is shown in Table 3. From the table, it can be seen that happiness and anger are confusing as well as sadness and boredom.

Table 3. The recognition results by NN classifier [%], where N, H, S, A, F, and B indicate Neutral, Happiness, Sadness, Anger, Fear, and Boredom, respectively.

Class	N	H	S	A	F	B	Accuracy rate
Neutral	78.6	6.9	4.8	1.4	0.7	7.6	78.6
Happiness	2.7	68.0	2.0	10.7	10.0	6.7	68.0
Sadness	7.3	2.7	62.7	2.7	10.7	10.0	62.7
Anger	2.7	10.7	3.3	78.7	2.7	2.0	78.7
Fear	4.0	7.3	4.0	6.0	65.3	13.3	65.3
Boredom	5.2	3.7	17.9	2.5	8.2	61.9	61.9
total							69.3

5.4 The recognition results by HMM classifier

We got accuracy rate of 89.3% in experiment using HMM classifier. HMM classifier produced higher accuracy by 20.4% than MLB classifier, and by 22.7% than NN classifier. Nonetheless, happiness and anger are still confusing as well as sadness and boredom.

Table 4. The recognition results by HMM classifier [%], where N, H, S, A, F, and B indicate Neutral, Happiness, Sadness, Anger, Fear, and Boredom, respectively.

Class	N	H	S	A	F	B	Accuracy rate

Neutral	96.7	1.7	1.7	0.0	0.0	0.0	96.7
Happiness	0.0	85.8	0.8	9.2	2.5	1.7	85.8
Sadness	0.8	3.3	82.5	0.0	3.3	10.0	82.5
Anger	0.0	4.2	0.8	92.5	1.7	0.8	92.5
Fear	0.0	1.7	0.8	0.0	93.3	4.2	93.3
Boredom	0.0	0.8	11.7	0.0	2.5	85.0	85.0
Total							89.3

6. CONCLUSIONS

In this paper we designed the emotion recognition systems for speech signal using MLB, NN, and HMM. The performance of these systems is investigated for six emotions, such as happiness, sadness, anger, fear, boredom and neutral. The mean, standard deviations, and max of pitch and energy were used as the features for MLB and NN. The pitch and energy per frame were used for HMM.

The recognition result of the subjective evaluation is 56%. MLB, NN, and HMM classifiers achieved the accuracy of 68.9%, 69.3% and 89.1%, respectively, for speaker dependent and text-independent classification. HMM yields the best result. 'Happiness and anger' and 'sadness and boredom' are very confusing in every experiment. The result of the subjective evaluation and the performance of our systems showed very similar error patterns in these points.

We suggested two feature vectors included characteristic of the pitch and energy contours. These feature vectors are delta pitch with delta-delta pitch and delta energy with delta-delta energy. HMM system with neighboring level of pitch mean and level of energy mean is very useful to distinguish emotional states, such as happiness and anger or sadness and boredom.

After all, using pitch and energy contour is desirable to decrease error rate rather than using statistics of pitch and energy. By using the contour the error rate caused from happiness and anger or sadness and boredom is decreased.

7. REFERENCES

[1] Rosalind W. Picard, *Affective Computing*, The MIT Press, 1997.

[2] Williams, W.E., and K.N. Stevens, "Emotions and Speech: Some Acoustical Correlates," in *J. Accoust. Soc. Am*, vol.52, no.4, pp.1238-1250, April 1972.

[3] Iain R. Murray and John L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," in *J. Accoust. Soc. Am*, vol. 93, no. 2, pp. 1097-1108, Feb. 1993.

[4] Deb Roy and Alex Pentland, "Automatic spoken affect analysis and classification," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pp. 363-367, Killington, VT, Oct. 1996.

[5] Frank Dellaert, Thomas Polzin, and Alex Waibel, "Recognizing emotion in speech," in *Proceedings of the Fourth International Conference on Spoken Language Processing,* vol.3, pp.1970-1973, Oct. 1996.

[6] Janet E. Cahn, "Generating expression in synthesized speech," Masters thesis, MIT Media Laboratory, May 1989.

[7] QUALCOMM Inc., Proposed TIA/EIA/PN-3292 Standard - Enhanced Variable Rate Codec, *Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, Official Ballot Version, April 1996.

[8] R.O. Duda, and P.E. Hart, *Pattern classification and scene anlaysis*, John Wiley & Sons Inc., 1973.

[9] Earl Gose, Richard Johnsonbaugh, and Steve Jost, *Pattern Recognition and Image Analysis*, Prentice Hall PTR, 1996.

[10] Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantization," *IEEE Trans. Commun.* vol. COM-28, no. 1, pp. 84-95, Jan. 1980.

[11] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, Feb. 1989.