International Journal of Fuzzy Logic and Intelligent Systems, vol. 6, no. 2, June 2006, pp. 150-154

Pattern Recognition Methods for Emotion Recognition with speech signal

Chang-Hyun Park and Kwee-Bo Sim*

School of Electrical and Electronic Engineering, Chung-Ang University
221, Heukseok-Dong, Dongjak-Gu, Seoul, 156-756, Korea
Tel : +82-2-820-5319, Fax : +82-2-817-0553, E-mail : kbsim@cau.ac.kr

Abstract

In this paper, we apply several pattern recognition algorithms to emotion recognition system with speech signal and compare the results. Firstly, we need emotional speech databases. Also, speech features for emotion recognition are determined on the database analysis step. Secondly, recognition algorithms are applied to these speech features. The algorithms we try are artificial neural network, Bayesian learning, Principal Component Analysis, LBG algorithm. Thereafter, the performance gap of these methods is presented on the experiment result section.

Key words : Emotion Recognition, Speech Signal, Neural Network, Bayesian Learning, Principal component analysis, LBG algorithm

1. Introduction

Many researchers on presence show that users are responding socially and emotionally to the systems, the characters in the virtual environments(VEs), or the robots that they are interacting with. For example, emotions affect the user's perception of the VE(Ijsselsteijn 2002;Lombard and Ditton 1997), and the interaction with the VE, in turn, affects the user's emotions.(Dillon et al. 2000; Kalawsky 2000)[1]. With advances in scientific technology, the influences of machines have become increasingly significant. Robots that have been used only in the industry have come to be used at home as well. Then, because human beings are very emotional and emotion plays a crucial role in human interaction, the robots also should be developed in such a way that can identify commands through human emotions to enhance its use at home. Many researchers have studied this field with speech, facial expression, and physiological signal. Fatma et al(2003) implemented a prototype multimodal affective user interface and analysed the physiological signals associated with emotions[1]. V.Hozjan et al(2003) tried a context independent multilingual emotion recognition from speech signals and the artificial neural network was used for the classification[2]. In 1998, Chen and Tao distinguished 6 different emotions; happiness, sadness, danger, dislike, surprise and fear by pitches and RMS energy envelope from each sentence. However, these features could not acquire the adequate recognition of emotions Therefore, they improved the method by adding the recognition of facial expressions[3]. The study showed that when both audio information and facial expression are used together, the recognition performance would be improved. In J. Nicholson's study, he divided emotions into a conscious and unconscious emotional expression. His study was limited to the conscious emotional expression, which was far easier to recognize. He extracted features of 8 emotions - joy, tease, fear, sadness, disgust, anger, surprise and neutrality referring them as conscious emotion. Those features were categorized into prosodic and, phonetic features. He obtained the result from the sub neural networks to decision logic. For this research, two different cases were studied: the closed case using training data and the open case using new speech sounds. The closed case resulted in 70 percent of the recognition rate average while the opened case only showed 30 percent of the recognition rate average. The study showed that the recognition rate can be significantly high in the case where personal data basis was used. Nevertheless, the open case, the low recognition rate proved the difficulty of identifying the general emotion features [4]. Also, in other paper, sound factors like pitch and energy, content-related features were also used; such as profanities, discourse information, repetition of the same sub-dialog [5]. The references discussed the emotional features and the classification methods. Many of the studies used the pitch as the common features, by focusing on reflecting the patterns in the feature extracting area [6,7]. In this paper, we apply several pattern recognition algorithms to emotion recognition system with speech signal and compare the results. Firstly, we need emotional speech databases. Also, speech features for emotion recognition are determined on the database analysis step.

Manuscript received Feb. 25, 2006; revised May. 15, 2006.

* Corresponding author

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(KRF-2005-042-D00268).

Secondly, recognition algorithms are applied to these speech features. The algorithms we try are artificial neural network, Bayesian learning, Principal Component Analysis, LBG algorithm. Thereafter, the performance gap of these methods is presented on the experiment result section.

2. Feature Exteation

The emotion recognizer is composed of two parts. The first part is the feature extraction part from the speech and the second part is the pattern recognition part from the features. The features are the statistical value of pitch, loudness (intensity), Section Number, Increasing Rate (IR) and Crossing Rate (CR). As a pitch extraction method, we used an autocorrelation approach which is one of the most common approaches. The pitch value was sampled at every 0.1 second and the mean of the values was defined as the pitch mean. Also, the variance was obtained from the same data. Loudness (intensity) was obtained by using a magnitude estimation method. Section Number, IR and CR ware obtained from the methods of our former paper [8].

3. Emotion Recognitions

3.1 Artificial Neural Network(ANN)

In this study, we used an ANN in order to classify emotions. ANNs provide a general, practical method for learning real-valued, discrete-valued, and vector-valued functions from examples. Algorithms such as back propagation (BP) use gradient descent to tune network parameters to best fit a training set of input-output pairs. ANN learning is robust to errors in the training data and has been successfully applied to problems such as interpreting visual scenes, speech recognition, and learning robot control strategies [9]. The parameters of ANN used for this paper is following;

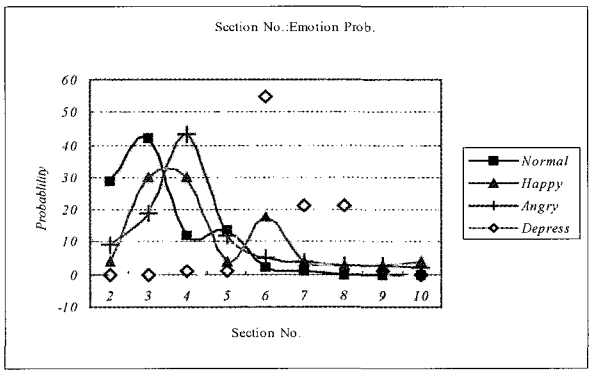
Table 1. Parameters of the ANN for this paper

Parameter	Values
Input Units	3~5
Hidden Units	11
Output Units	2
Learning Rate	0.003
Tolerance	0.25
Sigmoid Function	$1/(1+e^{-3x})$

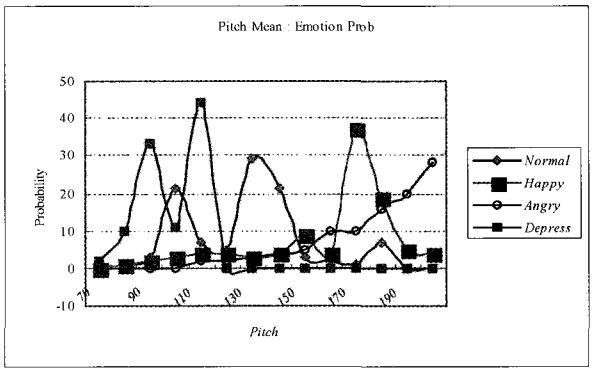
Since this problem had 2 binary outputs, an error tolerance of less than 25% was adequate for determining a correct output.

3.2 Bayesian Learning

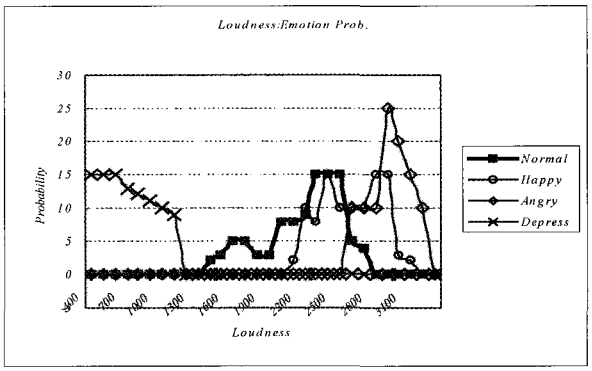
Bayes theorem provided a way to calculate the probability of a hypothesis is based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data. Therefore, the prior probability about the problem had to be obtained. 400 samples for the prior probability were used. More precisely, we observed the relationship between each emotion and the features (Pitch average, Loudness, Section No) from the samples.



(a) Section No.



(b) Pitch Ave.



(c) Lounness

Fig 1. The distribution of each emotion for (a) Section. No (b) Pitch Avg. (c) Loudness

Fig.1(a) represents the distribution probability of each emotion for the Sect. No. As the figure shows, sorrow can be clearly classified from other emotions. Normal, Happy, Angry each has a region with slight difference from one another. Fig.1(b) represents the distribution probability of each emotion for the pitch average. It shows that Sorrow, Normal, Happy, Angry in this figure have their own regions. Fig.1(c) shows the distribution probability of each emotion for the loudness. The distributions of Normal, sorrow, and angry have their own region, respectively. Therefore, this component can easily classify the emotions into 3 parts.

3.3 Principal Component Analysis

Principal Component Analysis (PCA) has been a popular feature extraction technique and also has been used as a pattern recognition technique. Specially, it has been employed in a number of computer vision applications ranging from face recognition through to texture flaw detection. Left of the Fig. 2 shows the procedure for training emotion data. The emotion feature vector is composed of pitch mean, loudness, section number, IR, CR. That is, 5×1 vector is collected and training vector set S is composed. Thereafter, the training is performed according to the process of left of the Fig 2. Also, the process to classify a new input data is following. Firstly, we get the eigen data elements, $\omega_k = u_k^T(\Gamma - \varphi)$. Secondly, we get the eigen data vector (Ω), $\Omega^T = [\omega_1, \omega_2, ..., \omega_{M'}]$. Lastly, we measure the distance between the training data and input data. $\varepsilon_k = \| \Omega - \Omega_k \|^2$.

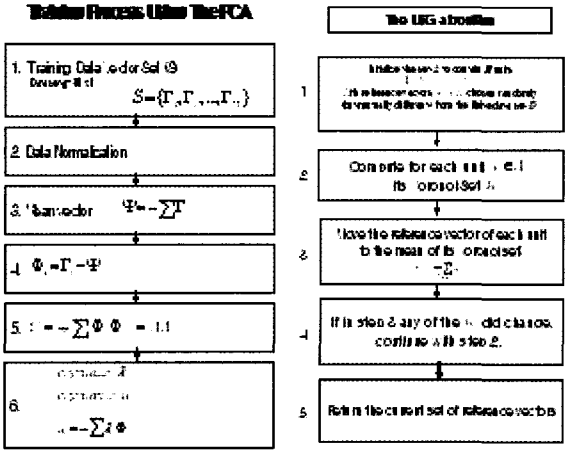


Fig 2. PCA and LBG algorithm

3.4 LBG Algorithm

The LBG algorithm works by repeatedly moving all reference vectors to the arithmetic mean of their Voronoi sets. The theoretical foundation for this is that it can be shown that a necessary condition for a set of reference vectors $\{\omega_c \mid c \in A\}$ to minimize the distortion error

$$\varepsilon(D, A) = \frac{1}{|D|} \sum_{c \in A} \sum_{\xi \in R_c} \| \xi - \omega_c \|^2 \quad (1)$$

is that each reference vector ω_c fulfills the centroid condition. In the case of a finite set of input signals and the use of the Euclidean distance measure the centroid condition reduces to

$$\omega_c = \frac{1}{|R_c|} \sum_{\xi \in R_c} \xi \quad (2)$$

where R_c is the Voronoi set of unit c. The complete LBG algorithm is shown in the right of the Fig 2. This algorithm is mainly used for 2-D data such as an image categorization. So, our training data types for it were 2-D. Then, because we got the several candidate parameters in the preprocessing stage, the evaluation of validity of each parameter was required. For that reason, some experiment for that was performed.

4. Emotional Speech Database

4.1 Emotional Speech Acquisition

10 male Graduate students, age range (24-31), were asked to speak 400 samples (10 subjects × 4 emotions × 10 sentences) of speech (subjects) with 4 emotions. They are ordinary Koreans but from different provinces. In this experiment, specific ages were not requested. Therefore a mean age was not calculated. The recording format was 11 KHz, 16 bit, mono and since the loudness depends on the gap between the subject and MIC. The gap was fixed to 10 Cm. The recorded speech was 30 common and simple everyday sentences and the length of the sentence was restricted to 6~10 syllables. Because the 30 prepared sentences should be verified whether they can be adopted as emotional data, we gave a question "Do you think which emotions were included or felt in each recordings." to 30 other people excepting the subjects who recorded emotional speech. As a result of the question, 10 sentences being resulted in 90% agreement were selected as final speech data and those were analyzed as representatives of each emotion.

5. Experiment Results

5.1 Recognition with ANN

Fig. 3 is the test result of applying the training results for 400 speech samples. To verify the usefulness of each feature, 4 feature sets were input to the ANN and tested. Of those sets, Feature 2 (Pitch average, Loudness, CR) was the most recognized set.

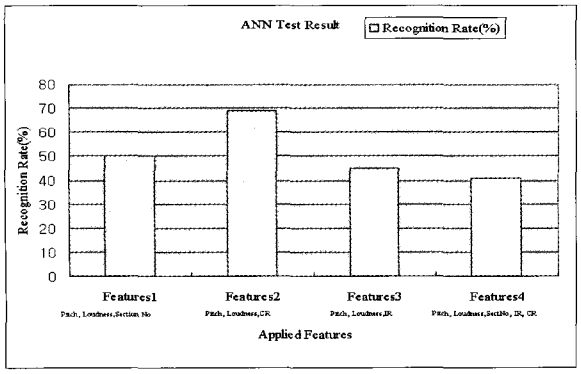


Fig 3. Test result of ANN

5.2 Recognition with Bayesian Learning

Table 2 is the result tested on 400 samples (each 100 samples to 4 subjects), which used the previously obtained distributions.

Table 2. BL experiment Result

	Normal	Happy	Angry	Depress	Average
S1	57%	40%	80%	70%	62%
S2	90%	73%	80%	94%	84%
S3	70%	51%	89%	91%	75%
S4	67%	56%	91%	85%	75%
Average	71%	55%	85%	85%	74%

Seeing the Table 2, we can think the recognition rate depends on the subject and emotions because the coincidence rate between the emotion expression pattern of a subject and the expression pattern consisting of the prior probabilities was different. Therefore, the greater the universality of the prior probability is, the better the performance.

5.3 Recognition with PCA

The input parameter type selected to the PCA algorithm was Loudness and Pitch mean. All training samples were not inputted to the PCA. Instead, each representative of 4 emotions was selected by LBG algorithm (That is, the center of each cluster). 100 test data was randomly selected. The experiment was tried 3 times (The PCA algorithm was implemented by MATLAB). So, the result is similar to that of the LBG. The error rate was calculated by *error rate=(Data No. in the other emotion category)/(Data No. of each emotion).*

Table 3. Error rate of PCA about emotion recognition

	Neutral	Angry	Sorrow	Happy
1st trial	68%	22%	18%	44%
2nd trial	73%	18%	15%	50%
3rd trial	70%	21%	20%	40%

Fig.4 is the result shown when the 3rd type data (Happy) were inputted. Because data of 3rd type were inputted, it was shown that the 3rd bar(Shown minimum Euclidean distance between input data and training data) is the smallest. Therefore this figure shows the correct solution.

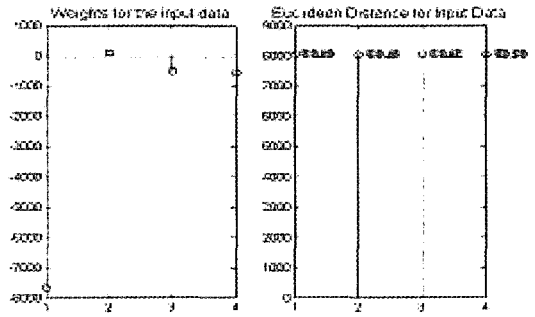


Fig 4. Recognition result by PCA when the happy(3rd bar on the right graph) data was inputted.

5.3 Recognition with LBG

The LBG algorithm adopted 3 pairs of training parameter type. The selected combinations were (a)Loudness and Pitch mean, (b) $\frac{Sect.no.}{Syllablesno.}$ and Pitch mean, (c) $\frac{Sect.no.}{Syllablesno.}$ and Loudness. Especially, the $\frac{Sect.no.}{Syllablesno.}$ means the speech tempo. We took some experiments for each pair.

Table 4. Error rate for 3 pairs of feature sets

	Neutral	Angry	Sorrow	Happy	Avg
Loudness, Pitch mean	71.4%	22%	20%	40%	38.2%
$\frac{Sect.no.}{Syllablesno.}$ Pitch mean	57%	71%	20%	60%	52%
$\frac{Sect.no.}{Syllablesno.}$ Loudness	71%	22%	20%	40%	38.2%

Table 4 shows the result of the experiment about what amount of validity those parameter types get. This table also shows the emotion recognition performance of the LBG algorithm. Because the data of sorrow and angry were biased, error rate of both was lower than other data. We could think that the LBG algorithm may be bad at this case. However, we also can't assure that the LBG algorithm is always not relevant to the emotion recognition. Fig. 5 shows the categorized data by matlab. 4 categories have each color and thick lined circles means the center of each category. The reason why the number of centers is not only 4 is because the final screen shows the trace of the center movement during the iterations. Also, the LBG algorithm was implemented by MATLAB.

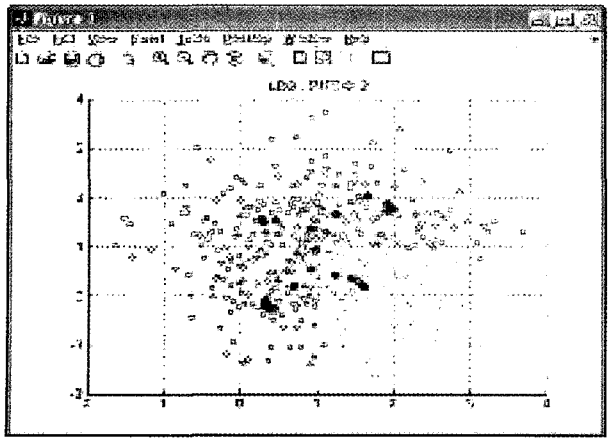


Fig 5. Finally Categorized Data by LBG algorithm

5. Conclutions

We presented the emotion recognition results by 4 sorts of pattern recognition algorithms. The feature extraction of the emotion recognition is felt not quite satisfied. However, we selected some feature sets by experiments and the best features showed about 26\% error rate. Also, the performance of the Bayesian learning and ANN was better than that of LBG and PCA. However,that result is very bad in the pattern recognition field but in the emotion recognition case, we think it is not the worst. Especially, from this paper, we can expect if we get some better emotion feature, we will get better performance. In contrary sense, if we don't get better feature, we can't expect the higher recognition rate. Truly, we have tried hard to search for features that represent the pitch contour and rhythm of speech but have not found those yet. In future work, we will focus on the that point and take an experiment with such features.

References

[1] F. Nasoz, K.Alvarez, C.L.Lisetti, and N.Finkelstein , "Emotion Recognition from physiological signals using wireless sensors for presence technologies," *Springer-verlag*, london, 2003.

[2] V.Hzjan and Z.Kacic, "Context-independent Multilingual Emotion Recognition from speech signals," *International Journal of Speech technology*, pp. 311-320, 2003.

[3] L.S.Chen, H.Tao, T.S.Huang, T.Miyasato and R.Nakatsu, "Emotion recognition from audiovisual information," *IEEE Second Workshop on Multimedia Signal Processing*, 1998.

[4] J.Nicholson, K.Takahashi, and R.Nakatsu, "Emotion Recognition in speech using neural networks," *Proc. Of ICONIP*, Vol.2, 1996.

[5] S.Batliner, K.Fisher, R.Hyber, J.Spilker, and E.Noth, "Desperately seeking Emotions:Actors, Wizards and Human Beings," *Proceedings of the ISCA Workshop on Speech and Emotion.*

[6] T.Moriyama and S.Ozawa, "Emotion Recognition and Synthesis System on Speech," *IEEE International Conference on Multimedia Computing and Systems*, Vol.1, 1999.

[7] D.Galanis, V.Darsinos, and G.Kokkinakis, "Investigating emotional speech parameters for speech synthesis," *Proc. of ICECS*, Vol.2, pp. 3-16, Oct, 1996.

[8] C.H.Park, K.S.Byun, and K.B.Sim, "The Implementation of the Emotion Recognition from Speech and Facial Expression System," *Proc. of ICNC*, Part 2, pp. 85-88, Aug, 2005.

[9] T.M.Mitchell, Machine Learning, *McGraw-Hill International Edition , Singapore*, 1997.

Chang-Hyun Park

Chang-Hyun Park received his B.S. and M.S. degrees in the Department of Control and Instrumentation Engineering from Chung-Ang University in 2001 and 2003 respectively. He is currently Doctor course in the School of Electrical and Electronics Engineering from Chung-Ang University, Korea. His research interests include emotion recognition, pattern recognition, machine learning, evolutionary computation, intelligent robot etc.

E-mail : 3rr0r@wm.cau.ac.kr

Kwee-Bo Sim

Kwee-Bo Sim received his B.S. and M.S. degrees in the Department of Electronic Engineering from Chung-Ang University, Korea, in 1984 and 1986 respectively, and Ph.D. degree in the Department of Electrical Engineering from The University of Tokyo, Japan, in 1990. Since 1991, he has been a faculty member of the School of Electrical and Electronics Engineering at Chung-Ang University, where he is currently a Professor. His research interests are in artificial life, emotion recognition, ubiquitous intelligent robot, intelligent System, computational intelligence, intelligent home and home network, ubiquitous computing and Sense Network, adaptation and machine learning algorithms, neural network, fuzzy system, evolutionary computation, multi-agent and distributed autonomous robotic system, artificial immune system, evolvable hardware and embedded system etc. He is a member of IEEE, SICE, RSJ, KITE, KIEE, KFIS, and ICASE Fellow. He is currently President of the KFIS.

Phone : +82-2-820-5319
Fax : +82-2-817-0553
E-mail : kbsim@cau.ac.kr