2005 IEEE International Workshop on Robots and Human Interactive Communication

Robust Emotion Recognition Feature, Frequency Range of Meaningful Signal*

Eun Ho Kim Kyung Hak Hyun Yoon Keun Kwak

Department of Mechanical Engineering

Korea Advanced Institute of Science and Technology

Guseong Dong, Yuseong Gu, Daejeon, Republic of Korea

kimeunho@kaist.ac.kr crommo9@kaist.ac.kr ykkwak@kaist.ac.kr

***Abstract* – Although the literature in emotion recognition from voice emphasizes that the recognition of emotions is generally classified in term of primary (or basic) emotions. However, they fail to explain the rationale for their classification. In addition, for the more exact recognition, more features to classify emotion are needed. But there are only a few features such as energy, pitch, and tempo. Hence, rather than using primary emotions, we classify emotions in emotional groups that have the same emotional state. We also propose a new feature called the frequency range of meaningful signal for emotion recognition from voice. In contrast to other features, this feature is independent of the magnitude of a speech signal and it is robust in a noisy environment. We also confirm the usefulness of this proposed feature through recognition experiments.**

Index Terms – Emotion recognition, frequency range of meaningful signal, emotional group, emotional sate, HRI

I. INTRODUCTION

Recently, robots have been developed as human-friendly robots. Human-friendly robots must be capable of doing several tasks, such as manipulating, perceiving, serving, and communicating. Communication is an important capability, not only in linguistic terms but also in emotional term. In the field of human-robot interaction (HRI), the ability of the robot to recognize emotions in human is a challenging issue, especially. When the recognition is based solely on voice, which is a fundamental mode of human communication.

This paper is organized as follows: In section 2, we focus on the emotional state and the emotional group in order to explain the rationale of classification from the viewpoint of psychology. In section 3, we propose a new feature for emotion recognition called the frequency range of meaningful signal (FRMS). In section 4, we compare the proposed feature with other existing features, such as energy and pitch. In section 5, we show the experimental results using FRMS to recognize emotions. Finally, in section 6 we summarize our conclusions.

II EMOTIONAL STATE

A. Necessity of an emotional state in HRI

For emotion recognition, we must first find a suitable classification of human emotions. Other researchers have classified emotions into a few primary emotions (see Table Ⅴ) but their classifications have lacked adequate explanations of validated reasons [1, 2].

However, given the vast number and variety of human emotions and the number of primary emotion that should be considered, it is not surprising that they failed to find a rationale. The root of this failure lies in the attempt to base the classification on primary emotions. In the field of HRI, the designation of each human emotion is unnecessary because robots recognize human emotion by a binary number. Hence, the name of each emotional state is unimportant to robot. From this point of view, it is better not to base the classification on primary emotions but rather on the human state, which we define as emotional state, including the physiological, behavioral, and cognitive state.

B. Emotional state

In 2000, Hiroyasu Miwa defined the emotional space as having three levels: the activation level, the pleasant level, and the certainty level. [3] Using the same approach, we can map emotions in the emotional space having infinite dimensions and infinite indices.

Psychology of emotion is can be categorized in terms of physiological, behavioral and cognitive psychology [4]. Each theory explains how humans feel and recognize emotions. According to these theories, humans recognize their own emotions by changes in physiology or behavior,

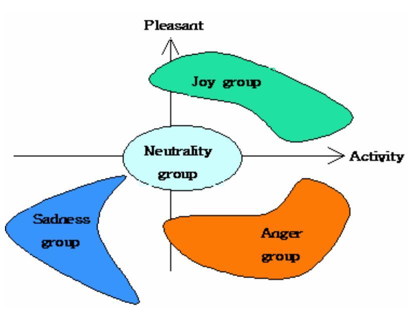

Fig. 1 Emotional state space. sentence

* This research (paper) was written as part of the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Science and Technology of Korea..

and, although the same changes may occur, humans experience other emotions that stem from other cognitions or thoughts for given a situation. By the same approach, infinite dimensions of emotion can be categorized into three groups, and we call each component of these categories an emotional state.

Behavioral and cognitive indices are not, however, recognized by using voice or facial expressions. Hence, the emotional space for emotion recognition has only physiological indices. Of the physiological indices, we propose the "activity" and the "pleasant" as the emotional states for emotion recognition. In Hiroyasu Miwa's model of emotional space, we removed a certain index because it is in the cognition category thereby necessitating artificial intelligence for recognition emotion.

With this model of emotional space, an infinite variety of dimensions can be compressed into two dimensions (see Fig. 1). In addition, an infinite numbers of emotions can be classified into four emotion groups namely joy, sadness, anger, and neutrality. For example, Hiroyasu Miwa defines six primary emotions into groups of joy (happiness), sadness (disgust, sadness), anger (anger, fear), and neutrality (neutrality). Hence, we don't recognize each primary emotion but rather each group of emotions that has the same emotional state. Now we will omit the group for brevity of expression.

Moreover, in terms of HRI, the emotional state has an advantage. To produce suitable reactions, the robot needs more information on the human state than on the name of emotion. For example, it is easier to design the robot's behavior with consideration of states of high activity and pleasant than design with consideration of joy. In addition, when we design the robot's reactions, there are more known of human emotions than human states. Hence, it is more complex to base the design on information regarding human emotions.

III. Frequency Range of Meaningful Signal

A. *What is the frequency range of meaningful signal?*

In general, human speech has a long frequency range. However, the important frequency range or meaningful frequency range is from 100 Hz to 5000 Hz [5]. We have

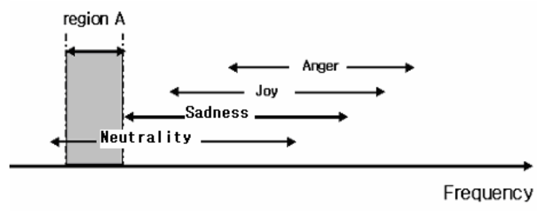

Fig. 2 Frequency range of meaningful signal

observed that this range varies for different people and different emotions, as illustrated in Fig 2. Using this point, we can recognize human emotion from voice. This new feature is called the FRMS.

Fig. 3 and Fig. 4 show four original emotional speech signals and low-pass filtered speech signals with a cut-off frequency of 160 Hz which occurs in the A region of Fig. 2. After passing the low-pass filter, the magnitudes of all the emotions decrease. Furthermore, the envelopes of all the emotions vanish except for neutrality, indicating, that only neutrality has meaningful signals under the cut-off frequency.

The main issue of this proposed feature is not the energy but rather the envelope. From Figs. 3 and 4, it might be confusing after the low-pass filtering, whether the energy of neutral emotion state is greater than that the energy of other emotional state. However, from Table Ⅰ we can determine from a two sample Z test that the P-value of the energy is 0.0892 and that the P-value of the envelope is about zero. Thus, an alternative hypothesis that neutrality has the same energy as sadness can be accepted with a 0.05 significance level whereas the hypothesis that neutrality has the same FRMS as sadness must be rejected [6]. From this result, we know that the FRMS differ in term of the energy feature. Furthermore, if we use the filters while moving the cut-off frequency, we can recognize each emotion by determining how much of the meaningful signal remains.

B. *Measurement of the remaining meaningful signal*

To use the FRMS feature, we must measure how much of the low-pass filtered speech signal remains as a meaningful signal. As you can see in figs.3 and 4, the meaningful signal can be represented by the envelope of a speech signal. Hence, the remaining meaningful signal can be measured by calculating the correlation between the

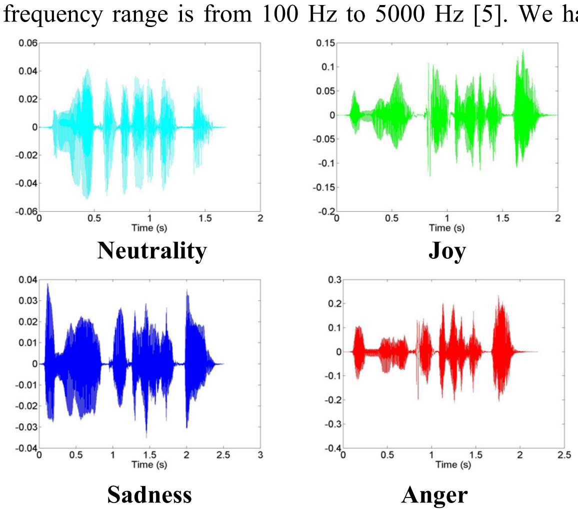

Fig. 3 Original speech signal for four emotions

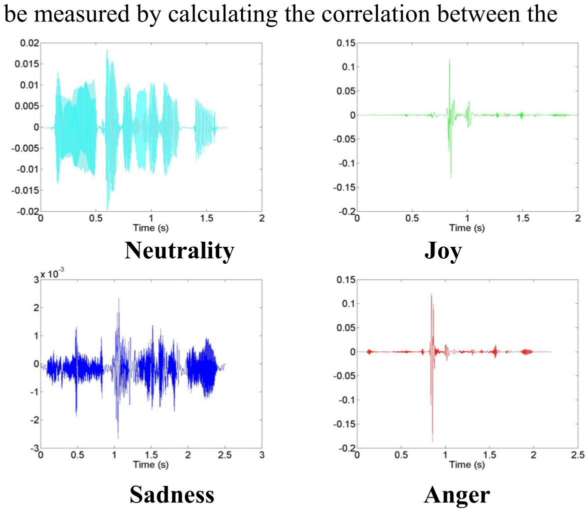

Fig. 4 Low-pass filtered speech signal for four emotions

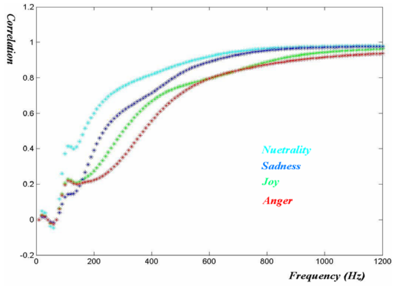

Fig. 5 Relation between frequency and correlation

envelope of the original speech signal and the envelope of the low-pass filtered speech signal.

Fig. 5 shows the relation between the frequency and correlation for a mean value of 135 sentences, with respect to the emotions of the first female speaker. From the figure, each emotion clearly has different characteristics; that is, there are differences in the correlation values of each cut-off frequency and in the rapidly increasing point, witch indicates the starting point of the meaningful signal. As you can see, this starting point (about 100 Hz) corresponds to the previously mentioned frequency range of 100 Hz~5000 Hz.

IV. COMPARISION WITH OTHER FEAUTRES

A. General features of energy and fundamental frequency

Most of the research on classifying emotions from voice has focused on the energy and the fundamental frequency (pitch). Hence, we compare our proposed feature, FRMS, with energy and pitch.

Energy can be extracted from equation (1). Furthermore, by using Karhunen-Loeve (KL) expansion, we can reduce the nine-dimensional energy vector (X_E) to two-principal dimension vector using [7], as follows:

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \qquad w(m) = 1 \quad 0 \le n \le N-1$$
$$= 0 \quad \text{otherwise} \qquad (1)$$
$$X_E = \left[\overline{E} \;\; E_{max} \;\; E_{std} \;\; \overline{\Delta E} \;\; \Delta E_{max} \;\; \Delta E_{std} \;\; \overline{\Delta\Delta E} \;\; \Delta\Delta E_{max} \;\; \Delta\Delta E_{std}\right]$$

To extract the pitch, we used the simple inverse filtering tracking Algorithm [8]. We also used the KL expansion to reduce the nine-dimensional pitch vector (X_P) to three-principal dimension vector [7]. Finally we used the train data and test data in the same way as in the FRMS experiments.

Table Ⅱ compare the performance of these two features and the FRMS. As a result, the overall recognition rate is almost the same for each of the three features.

TABLE I
MEAN AND STANDARD DEVIATION OF THE ENVELOPE AND ENERGY FOR EACH EMOTION

	Envelope		Energy	
	Mean	Std.	Mean	Std.
Neutrality	0.8496	0.0236	0.0089	0.0060
Joy	0.5991	0.0708	0.0200	0.0146
Sadness	0.6762	0.077	0.0065	0.0049
Anger	0.5252	0.0817	0.0189	0.0095

TABLE Ⅱ
COMPARISON OF THE FRMS FEATURE WITH ENERGY AND PITCH

	FRMS	Energy	Pitch
Neutrality	74.7	65.8	93.7
Joy	70.6	67.5	64.4
Sadness	82.5	82.5	64.6
Anger	75.0	83.4	93.2
Overall	75.7	74.8	79.0

Moreover, the recognition characteristics of the FRMS are better than others in the uniformity of recognition rate for each emotion.

B. What is the advantage of the FRMS feature?

Aside from the recognition rate, the key advantages of the FRMS are that it is not dependent on the magnitude of the speech signal and it is robust in noisy environments.

In practice, when the magnitude of a speech signal changes due to the magnitude of a speaker's voice, the distance between the speaker and the microphone or the characteristics of microphone, the recognition performance deteriorates. Hence, compensation for this deterioration is a significant study theme.

However, the FRMS is concerned only with the envelope of the signal and not the magnitude. Further, the FRMS focuses on the relationship between the original speech signal and the low-pass filtered speech signal. As such, the FRMS is independent of the distance or the magnitude of the voice, thereby making it powerful in practical use. To verify this independence, we performed recognition experiments using the same training data as the original magnitude but with half and double magnitude test data. As anticipated, recognition experiments of original, half and double magnitude speech data yielded almost the same results, as shown in Fig. 6. Furthermore, the results confirm that the FRMS feature has the advantage of being able to be used in practice without any pre-processing for distance compensation.

Another advantage of the FRMS feature is its robustness in noisy environments. Because most noises have a frequency higher than the cut-off frequency, high-frequency noises tend to disappear after low-pass filtering. In addition, if noises are weaker than the main voice, then most of these noises will not affect the envelope of the main voice. Hence, the FRMS can be robust in a noisy environment. To verify this robustness, we performed recognition experiments on voice noise and white noise speech data using the same training data as previously used. The voice noise data consisted of the main voice and two other set of emotional voice data from a different speaker with approximately half of the magnitude. The white noise data consisted of the main voice and white noise at about a quarter of the magnitude of the main voice. The experimental results confirm the robustness of the FRMS feature (see Fig. 6).

To compare the advantages of the FRMS feature with other features (energy and pitch), and to verify our half and double magnitude data, as well as the voice noise data

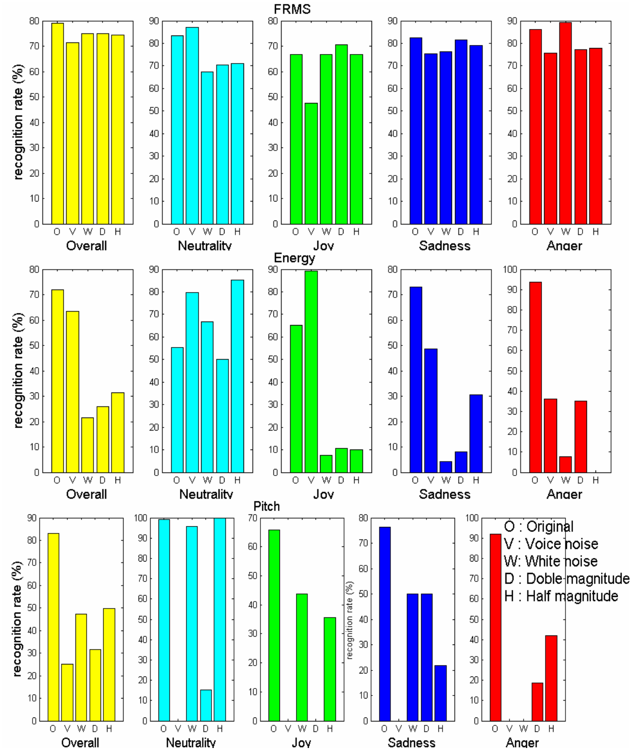

Fig. 6 Experiment results of the managed,
and a comparison of the FRMS features with energy and pitch

and white noise data, we performed the recognition experiment on the managed sample data using the FRMS, energy and pitch.

Fig. 6 shows the results of experiment. As anticipated, the overall recognition rate of the FRMS decreased slightly. However, the amount of decrease is sufficiently smaller than the other features to be considered robust and independent. More importantly, in most case, the overall recognition rate for each emotion deviates by no more than 10 percent and there is negligible change in the recognition characteristics that are the distinctiveness of the emotion. This result demonstrates the independence and robustness of the FRMS feature.

For the energy and pitch, the overall recognition rate decreased greatly (see Fig. 6), and the most emotions could be distinguished as only one or two emotions. For example, when the half-magnitude data was used with the energy feature the high-energy emotions were open mistaken for low-energy emotions; similarly, for the double-magnitude date the low-energy emotions were mistaken for high-energy emotion. Furthermore, with respect to the recognition of the voice and white noise data, the energy feature misrecognized all emotions as high-energy emotions, while the pitch feature open mistook all emotions as neutrality or some cases as sadness in recognition of all the managed data that are half magnitude, double magnitude, voice noise and white noise data. In other words, the energy and pitch features both failed in term of the recognition characteristics. From this failure, we conclude that the energy and pitch features depend on the magnitude of the voice or the distance, and are therefore not robust in a noisy environment.

TABLE Ⅲ
RESULTS OF RECOGNITION USING THE FRMS FEATURE

% / Recog.	Male			
	Neutrality	Joy	Sadness	Anger
Neutrality	80.2	4.8	10.5	4.6
Joy	1.7	80.0	5.1	13.1
Sadness	8.4	8.0	80.6	3.1
Anger	2.7	20.8	0.2	76.4
Overall	79.3			
% / Recog.	**Female**			
	Neutrality	Joy	Sadness	Anger
Neutrality	69.1	5.0	21.1	4.8
Joy	8.2	61.1	4.8	25.9
Sadness	8.8	5.9	84.4	0.8
Anger	3.4	19.6	3.4	73.5
Overall	72.0			

V. EXPERIMENTS

A. Database

Given that in many languages the fundamental tendencies of sounds are expressed in similar ways, our results in recognizing the emotions of Korean language speakers can generally be applied to speakers of other languages. For this reason, we used a database produced by Professor C.Y. Lee of Yonsei University's Media and Communication Signal Processing Laboratory with the support of the Korea Research Institute of Standards and Science. This data covers the four emotions of neutrality, joy, sadness and anger; and its principles are as follows [9]:

- easy pronunciation in a neutral, joyful, sad and angry state;
- 45 dialogic sentences that express a natural emotion;

The original data is stored in the form of 16 kHz and 32bits over 30dB S/N and margined with no sound for about 50 ms in the beginning and end of the utterance. To use the data in MATLAB, we changed the data for the training and experiments into 16 bits format through quantization with pulse code modulation filter.

To verify how accurately the database reflects a speaker's emotions, experiment were conducted at Yonsei University on the subjective emotion recognition of human [9]. Table Ⅳ shows the results of these experiments. The recognition rate was unequal for each emotion, and, for the recognition characteristic, sadness was well recognized but joy was not.

TABLE Ⅳ
HUMAN PERFORMANCE OF EMOTION RECOGNITION FOR THE DATABASE

% / Recog.	Male			
	Neutrality	Joy	Sadness	Anger
Neutrality	83.9	3.1	8.9	4.1
Joy	26.6	57.8	3.5	12.0
Sadness	6.4	0.6	92.2	0.8
Anger	15.1	5.4	1.0	78.5
Overall	78.1			

TABLE Ⅴ
COMPARISON WITH OTHER FEATURES

Features	Emotions classified	Average accuracy (%)
Pitch contour (Xiao Lin et al.,1999)	Normal, Emotive	82
Energy by pattern recognition (Dellaert et al.,1996)	Happiness, Sadness, Anger, Fear	79.5
Energy by neural network (Nicholson et al.,1999)	Joy, Teasing, Fear, Sadness, Disgust, Anger, Surprise, Ne utral	50
LFPC by Tin Lay New et al., 2003	Anger, Sadness, Joy, Neutrality	78.1
Proposed method (FRMS)	Anger, Sadness, Joy, Neutrality	75.7

B. Experimental method

The databases consist of 5400 sentences (that is, 45 dialogic sentences times three repetitions times four emotions times ten speakers comprising five males and five females). The same set of sentences was used for all four emotions. For training, each experiment used 30 sentences (that is ten dialogic sentences times three repetitions for each speaker). Based on this training database, the recognition experiments were conducted on the remaining 80 percent of the data.

Fig. 7 shows a schematic diagram of the experiment. The FRMS extraction was performed by calculating the correlation between the original speech signals and the low-pass filtered speech signal from 10 Hz to 1200 Hz at every 10 Hz. To reduce the processing time, we used cubic interpolation. The cubic interpolation reduces the number of passes in the low-pass filtering process from 120 to 16. In addition, when we used the KL expansion, the dimensions of the correlation vectors $X_{FRMS} \in R^{120}$ were reduced to six for males and eight for females [7]. For the classification, we used the Bayes classifier (that is, normal distribution based quadric classifier) [10].

C. Experiment result

Table Ⅲ shows the experiment results for the male and female speakers. From these results, we deduce that sadness has the best recognition rate and joy has worst recognition rate. And these recognition characteristics are same as the results of the recognition by humans. The average recognition rate is approximately 76 percent (that is, 79 percent for males and 72 percent for females).

Finally, we compared the performance of the proposed feature with the results of the other researcher obtained for

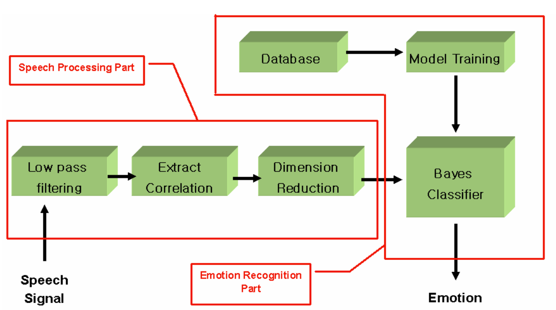

Fig. 7 Schematic diagram of the experiment

other features. Table Ⅴ summarize these results [1, 2]. Notice that each result differs in the number and type of emotions classified and the size of the database used. Nevertheless, it provides a crude comparison of the proposed FRMS feature.

Ⅵ. CONCLUSIONS

We have defined the emotional state for emotion recognition in the HRI field, and from this definition have explained the rationale of the classification of human emotions from the viewpoint of psychology. We proposed a new feature called FRMS. From the results of experiments, we show that the recognition rate is approximately 76 percent. We also show that the FRMS feature is independent of the magnitude of the speech signal, and it is robust in noisy environments. These two points of the FRMS feature are remarkable advantages that enable emotions to be easily recognized without the problem of noise or distance between the speaker and microphone.

In the future, we will hope to verify the FRMS feature with human test, and we plan to compare the recognition rate of this method with that of various classification methods. A real filed test (not tested by the database) must be accomplished for the practice use of the FRMS feature.

REFERENCES

[1] Xiao Lin, Yanqiu Chen, Soonleng Lin, and Choonban Lim, "Recognition of Emotional state from Spoken Sentences," *IEEE Multimedia signal processing* pp. 469-473, 1999

[2] Nwe, T.L, Foo, S.W, De Silva, L.C, "Speech emotion recognition using hidden Markov Models," *Speech communication*, vol. 41, no. 4, pp.603-623, 2003

[3] Hiroyasu Miwa, Umetsu, T., Takanishi, A., Takanobu, H., "Robot Personalization based on the Mental Dynamics," *Intelligent Robots and Systems*, *IEEE/RSJ conf.* vol. 1, pp. 8-14, 2000

[4] Robert Plutchik, *Emotions and Life : perspectives from Psychology, Biology, and Evolution*, American Psychological Association press, 2003

[5] Cloria J. Borden, Katherine S. Harris, Lawrence J. Raphael, *Speech Science Primer*, 3rd ed., Williams & Wilkins press, 2000

[6] Ronald E. Walpole, et al, *Probability and Statistics for Engineers and Scientists,* 7th ed, Prentice Hall press, 2001

[7] Josef Kitter, "A method for determining class subspace," *Information Processing Letters,* vol. 6, issue 3, pp. 77-79, 1977

[8] John D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," *IEEE Trans.* vol. AU-20, No.5, pp. 367-377, 1972

[9] Kang, Bong-Seok, "Text Independent Emotion Recognition Using Speech Signals," *Yonsei Univ.* 2000.

[10] S.Kumar, *Neural Networks: a classical approach,* Mc Graw Hill press, 2004, pp.219-235.