International Journal of Fuzzy Logic and Intelligent Systems, vol. 2, no. 2, June 2002 pp. 122-126

Emotion Recognition Based on Frequency Analysis of Speech Signal

Kwee-Bo Sim*, Chang-Hyun Park*, Dong-Wook Lee*, and Young-Hoon Joo**

*** School of Electrical and Electronic Engineering, Chung-Ang University, Seoul, Korea**

**** School of Electronic and Information Engineering, Kunsan National University, Chonbuk, 573-701, Korea**

Abstract

In this study, we find features of 3 emotions (Happiness, Angry, Surprise) as the fundamental research of emotion recognition. Speech signal with emotion has several elements. That is, voice quality, pitch, formant, speech speed, etc. Until now, most researchers have used the change of pitch or Short-time average power envelope or Mel based speech power coefficients. Of course, pitch is very efficient and informative feature. Thus we used it in this study. As pitch is very sensitive to a delicate emotion, it changes easily whenever a man is at different emotional state. Therefore, we can find the pitch is changed steeply or changed with gentle slope or not changed. And, this paper extracts formant features from speech signal with emotion. Each vowels show that each formant has similar position without big difference. Based on this fact, in the pleasure case, we extract features of laughter. And, with that, we separate laughing for easy work. Also, we find those for the angry and surprise.

Keywords : Formant, Pitch, Slope, Quasi-period, Vocal tract

Ⅰ. Introduction

Recently, human being's technology has reached to a walking robot, a speaking robot etc. But, we shouldn't stop at this stage. If a machine can understand somebody's emotion, it can help him or her in several ways. Human being has a lot of emotions. That is, anger, happiness, sorrow, surprise, normal, gloominess, etc. From these, we select the most recognized emotions- Anger, happiness, surprise-. The reason of this selection is showed from Frank Dellaert's paper [2]. According to the paper, recognition error rate of anger is 1%. that of happy is 3%. that of sadness is 5%, etc. And we'll extract features of those. Thereafter, Based on that result, we do comparison about each emotion features. Generally, there are (1) words for conversation (2) Tone (3) Pitch (4) Formant Frequency (5) Speech speed, etc as the element for emotional recognition from speech signal. For human, it is natural that tone, words, speed, voice quality is easier elements rather than frequency to perceive other's feeling. Therefore, the former things are important elements for classifying feelings. And, already established methods mainly used the former things, but using formant is good for implementing as machine. Thus, our final goal of this research is implementing an emotional recognition system using pitch, formant, speech speed, etc from speech signal. A paper Detecting data which represent emotion features from the speech signal showed speech speed doesn't work at classification of emotions [1]. As another way of representing the time-varying signal characteristics of speech is via a parameterization of the spectral activity based on the model of speech production. Because the human vocal tract is like a tube, or concatenation of tubes, of varying cross-sectional area that is excited either at one end or at a point along the tube (corresponding to turbulent air at a constriction), acoustic theory tells us that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequenci

This work was supported by grant No. N09-A08-4301-05 (Development of the Key Technology for Autonomous Family Machine) from the project of Developing SIC(Super Intelligence Chip) and its Applications under the program of Next generation technologies in 2000 of Ministry of Commerce, Industry and Energy.

Ⅱ. Setting

At first, we made 10 people speak 10 given scripts with one's emotions. And wave format is set to 22.05Khz, 16bit, mono. for analysis, we used a program 'Praat' made by Dr. Paul Boersma. And, we should decide a segmentation method. From waveform, we look the magnitude in 0.15s interval. Then , obtain average in that interval. At that time, if average is below 0.02 (wave form is bounded +,-0.5) for several frames , it is defined as a segmentation point. Once the start point and end point of an utterance have been determined, we should estimate the pitch contour and the formant.

2.1. Analysis method

First of all, we take the segmentation process (explained that process before in briefly). Then, get the pitch slope at 0.05seccond intervals. So we can decide by the slope[6] and several rules representing the emotional features which emotion is conveyed.

III. The features of emotions

At this time, we treat 3 emotions (angry, happiness, surprise). When a man gets angry, the pitch contour of his speech is steeper than that of normal state. And, In his state, he will shout mostly. Then, his mouth cavity, tongue and nasal cavity are constricted or extended. The results will affect the frequency of sound. As it is, formant depends on it directly. For example, if mouth cavity is constricted, F1 will be lower than that of the normal state [3]. According to these points, analyzing the features of emotions is performed.

3.1. Happiness

When a human gets happy, he or she laughs on saying loudly. Also, it can be done without laughing. But, Most peoples saying and laughing are used together. So, our research is focused on finding the features of laughing. The sort of laughing is various and isnt defined as a specific vowel. So it is difficult to find a laughing word in the signal form. At first, Fig. 1 shows the position of vowels in F1, F2 (F1: 1st Formant, F2: 2nd Formant).

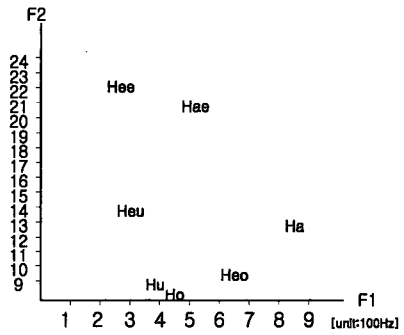

Fig. 1. F1, F2 position for some vowels

Vowels which are shown in Fig 1 are things generated when human laughs. And, voice signal has other elements in addition to presented formant. But, because F1 and F2 dominate the signal correspond to a vowel, we treat mainly F1 and F2. And in the above figure, the bound of each vowel is made to a tolerance of about 70~90Hz. Laughing has several sort. So, in briefly, we classify into 2 types. A broad laugh and A foxy laugh.

3.1.1. A broad laugh

At first, example for a broad laugh is Ha Ha Ha Ha , eu Heo Heo Heo etc. this sort of laugh shows explicitly pitch features.

Fig. 2 represents laugh ha ha ha ha which was performed by a 27 years old man. Circles show how pitch of each syllable changes. Fig. 3 also shows similar contour(Fig. 3 represents heu hae hae hae hae by a 40 years old man).

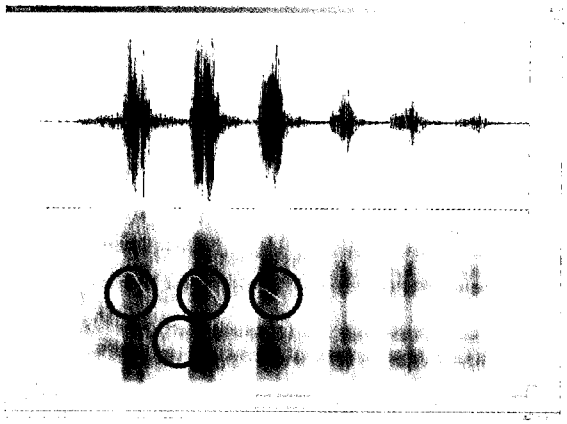

Fig. 2. Broad laugh ha ha ha ha

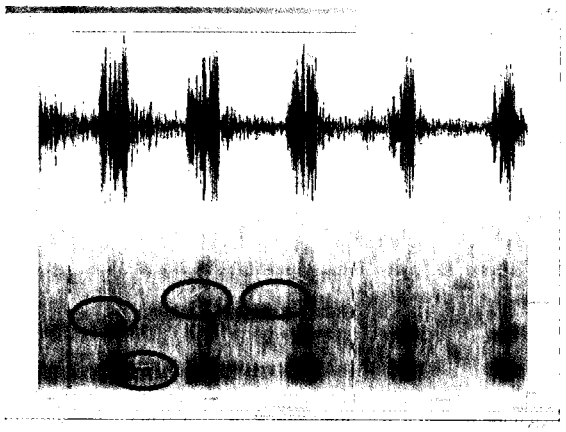

Fig. 3. Broad laugh heu hae hae hae hae

And, as another important feature, there are F1 and F2. Then, from above two figures, we can see F1, F2 keep at a fixed level. And, through the comparison of F1, F2 of Fig. 1 and those of Fig. 2 and 3, we can know formant of Fig. 2 is near to a vowel ha , and formant of Fig. 3 is between hae and heo . That is, it shows that we can know inversely which vowel pronounced.

3.1.2. A foxy laugh

Another laugh is classified by a foxy laugh in this paper. This is not split by syllable differently to a broad laugh. Of course, this sort of laugh also shows F1, F2 corresponded to vowel in Fig. 1. The feature of this laugh is ceaseless pitch contour. Besides, it is bounded in interval about 100Hz.. and, because a man is in excited state, his tone is high (from 350Hz to 500Hz). Fig. 4 shows features which stated in the above sentences (Fig. 4 hee hee hee hee hae hae hae by a 40 years old man). And, frames of Fig. 4 are shown as the standard for segmentation (lines in the upper window) and for foxy laugh (lines in the lower window). Also, Most of foxy laughs show Pitch over 370Hz

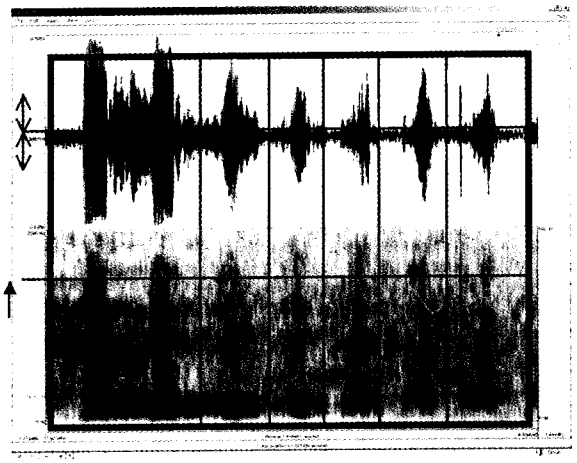

Fig. 4. Foxy laugh hee hee hee hee hae hae hae

3.2. Angry

We focus on the change of formant in this emotion. Of course, we dont exclude the pitch contour. It is always important element in the emotion recognition. Angry is also classified into 3 types. Question type, Shout type, Explanation-type. And, if anybody gets angry, then he/she will be excited and say with exaggerated gesture. At that time, physical change makes sound different to that of normal state. Fig. 5 shows A-type vowels have different F1, F2 to B-type vowels. The more tongue is distant from palate in case of A-type vowels, the higher F1 gets, the lower F2 gets. And, The more pharyngeal gets far from the glottis in case of B-type vowels, the lower F1 gets. The more anyones lips purse up, the more F2 contracts.

At that time, this truth can be used when we extract the feature of angry state. For example, Mol is pronounced in Fig. 6. A-type is pronounced normally, B-type is pronounced in the angry state. It shows that saying in the angry state has changed F1 and F2.

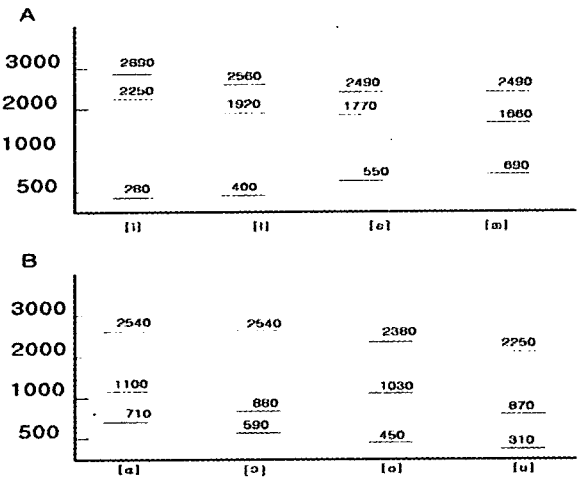

Fig. 5. Formants for some vowels

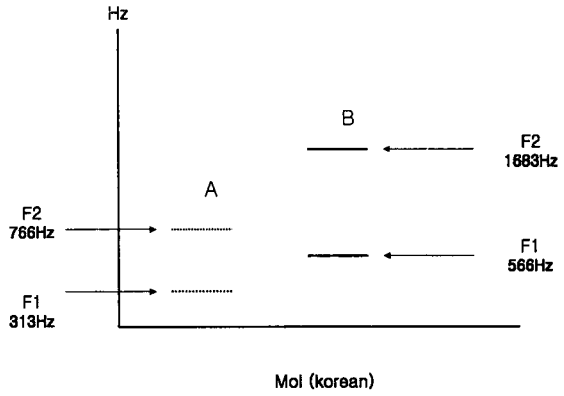

Fig. 6. Change of formants by physical change.

3.2.1 Question type

This type is shown often when people get angry. In this case, end segment of sentence is steeply risen. Of course, though pitch contour of normal state was such form. that of the angry state shows much steeper pitch contour than the normal state. Scripts of Table 1 represent when anyone gets angry in question type and when anyone gets any question. From above Table, we can see that each state is different explicitly. And, pitch contour before showing this pattern has low value and not sudden changes.

	Script	Difference
Angry	이제 눈에 뵈는게 없냐 (I-Jae Noon-e Bae-Neun-Ge Up-Nya)	135Hz
	내가 그렇게--- 한 사람인줄 알어 (Nae-Ga Gue-Rer-Ke...Han Sa-Ram-In-Jool Al-A)	128Hz
	니가 뭘 안다고 함부로 떠들어 (Ni-Ga Mol AnDaGo Ham-BuRo Tu dul Uh)	116Hz
Nomal	이거 얼마에요(I-Gu Ul-Ma-E-Yo)	51Hz
	이거 어때요(I-Gu-Ur-Tae-Yo)	11Hz
	뭐라고 해야되지 (Mo-Ra-Go Hae-Ya Doe Ji	46Hz
	그럼 어떻게 되는 거야 (Gue-Rum Ur-Tur-Ke Dae-Neun-Gu-Ya	33Hz

3.2.2 Explanation-type

Characteristic of this type is composed of long sentence. So, differently to a short case, this case shows large wave or shout rarely. Segment term of Fig. 7 is long. From this figure, Pitch contour represents a speaker emphasizes quasi -periodically (circles express the stressed point.).

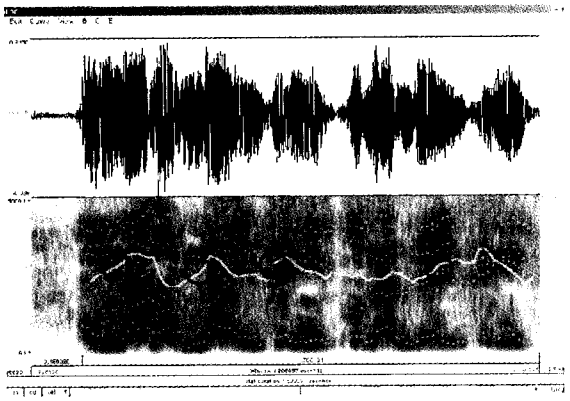

Fig. 7. Pitch contour for explanation-type

3.2.3 Shout type

This type shows a typical short sentence. And, because people shout shortly, this type gives us a simple pitch contour with formant changed. As explained in Fig 6, Angry ?Physical change ? Formant changed. That is, the more lips opened, the higher F1 and the more pharyngeal gets far from the glottis in case of B-type vowels, the lower F1. Also, Pitch contour of shout type have the falling form or mountain form (rising shortly and falling down instantly).

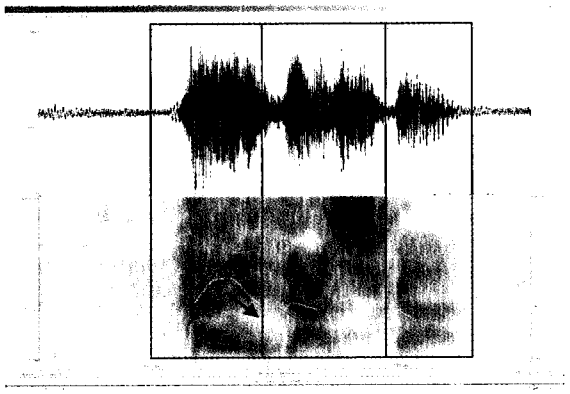

Fig. 8. Pitch contour for Shout type

Fig. 8. shows Ya, I-Za-Sic-A , first segment represents "Ya" shouted by speaker.

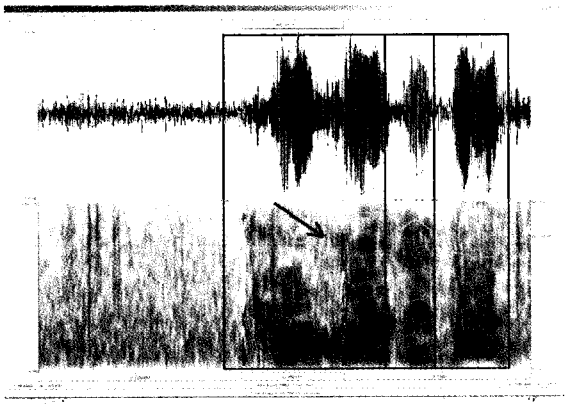

Fig. 9. Pitch contour for Shout type

Fig. 9 shows "Zerck-Ur-Do...", an arrow of first segment represents a stressed syllable.

3.3. Surprise

When the surprise occurred, peoples linguistic central nerves are paralyzed. So most people cant speak any sentences. But, after an instant, people can say as a response for any cause of surprise. Anyway, in this paper, we regard the surprise as a scream. As we find the characteristic of a scream, it is similar to that of a shout type in the pitch aspect. There is, As a same feature, falling down pitch contour. But, there is a different thing. Pitch contour of a shout type shows any vowel but, that of a surprise doesnt show explicit vowel (F1,F2). And, the feature of another scream represents N-type pitch contour. N-type pitch contour literally represents N form. N-type is often observed from ladies scream.

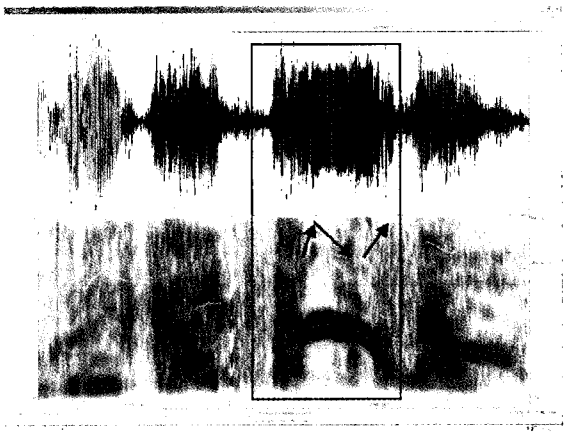

Fig. 10. Pitch contour for N type (surprise)

Ⅳ. Conclusion

First of all, we should assume that preprocess is completed. Preprocess is the classification of sex and the acquisition of reference pitch and F1/F2 for an object. Thereafter, segmentation process ?computation of slopes for pitch? check F1, F2 value? classification based on acquired data. In this paper, we extracted features of 3-emotions. These have been overlooked until now, but extraction of features from emotions is very important and should not be missed. So, we found features of typical emotions. But, features which we found are insufficient and in parts. Of course. this parts cover many aspects of each emotions. In future, we will simulate our emotion recognition system with these features.

References

[1] C. H. Park, D. W. Lee, Y. H. Joo, and K. B. Sim, "Detecting data which represent emotion features from the speech signal," *ICCAS 2001*, Che-ju island, Korea, fall, 2001.

[2] F. Dellaert, T. Pozin, and A. Waibel, Recognizing Emotion In Speech, *Technical Report*, Carnegie Mellon Univ.
[3] K. S. Lee and D. I. Seok, *Auditory*, Dae-Gu University, 1996.
[4] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice-Hall International,1993.
[5] T. L. New and F. S. Wei, Speech Based Emotion Classification, *Electrical and Electronic Technology*, Tencon, 2001.
[6] X. Lin, Y. Chen, S. Lim, and C.Lim, Recognition of Emotional State from Spoken Sentences, *Multimedia Signal Processing, IEEE 3rd workshop*, 1999.

Kwee-Bo Sim
Kwee-Bo Sim was born September 20. 1956. He received the B.S. and M.S. degrees in Department of Electronic Engineering from Chung-Ang University, Seoul Korea, in 1984 and 1986 respectively, and Ph. D. degree in Department of Electronic Engineering from the University of Tokyo, Japan, in 1990. Since 1991, he has been a faculty member of the School of Electrical and Electronic Engineering at the Chung-Ang University, where he is currently a Professor. His research interests are Artificial Life, Neuro-Fuzzy and Soft Computing, Evolutionary Computation, Learning and Adaptation Algorithm, Autonomous Decentralized System, Intelligent Control and Robot System, and Artificial Immune System etc. He is a member of IEEE, SICE, RSJ, KITE, KIEE, ICASE, and KFIS.

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

Dong-Wook Lee
Dong-wook Lee was born August 24, 1973. He received the B.S. M.S., and Ph.. D degree in Department of Control and Instrumentation Engineering from Chung-Ang University, Seoul, Korea, in 1996. 1998, and 2000 respectively. He is currently pursuing post-doc. course at Chung-Ang University. His research interests are Evolutionary Computation, Artificial Life, Artificial Immune System, and Evolvable Hardware. He is a member of KITE, KIEE, ICASE, and KFIS.

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

Chang-Hyun Park
Chang-Hyun Park was born April 10,1975. He received the B.S. in Department of Control and Instrumentation Engineering from Chung-Ang University, Seoul,Korea, in 2001. He is currently pursuing M.S. course at Chung-Ang University. His research interests are Evolutionary Computation, Artificial Life, Artificial Immune System, and Evolvable Hardware.

Phone : +82-2-820-5319
Fax : +82-2-817-0553
E-mail : 3rr0r@wm.cau.ac.kr

Young-Hoon Joo
Young-Hoon Joo received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Yonsei University, Korea, in 1982, 1984, and 1995, respectively. He worked with Samsung Electronics Company, Korea, from 1986 to 1995, as a Project Manager. He was with University of Houston, TX, from 1998 to 1999, as a Visiting Professor in the Department of Electrical and Computer Engineering. He is currently Associate Professor in the School of Electronic and Information Engineering, Kunsan National University, Korea. His major is mainly in the field of mobile robots, fuzzy modeling and control, computer vision, genetic algorithms, intelligent control, and nonlinear systems control. Prof. Joo is serving as the Associate Editor and Director for the Transactions of the KIEE (2000-2002) and Journal of Fuzzy Logic and Intelligent Systems(1999-2002). He is a member of KITE, KIEE, ICASE, and KFIS.

Phone : +82-63-469-4706
Fax : +82-63-469-4706
E-mail : yhjoo@kunsan.ac.kr