Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Automatic Segmentation for Emotional Feature Extraction from Spoken Sentence

Kyung Hak Hyun *. Eun Ho Kim. **
Yoon Keun Kwak***

**Mechanical Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, Korea, (e-mail:cromno9@kaist.ac.kr)*
*** Mechanical Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, Korea, (e-mail:kimeunho@kaist.ac.kr)*
**** Mechanical Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, Korea, (e-mail:ykkwak@kaist.ac.kr)*

Abstract: Perception of speaker's emotion is one of interesting issues in human-robot interaction. Especially, friendly and instinctive interface between robots and humans is required for making service robots useful to inexpert interacting with robots. Among several mode in communications, speech is easiest method for human because speech is fundamental communication tool in human-human interaction. However, continuous speech is difficult to extract some features because speech is time-variant signal. In other words, segmentation is necessary to analyze speech signal easier. Researcher who were interested in phonetic information usually used 20~40ms windowing because they should extract features in short duration which gives an assumption about time-invariant in a frame. On other hand, emotions in speech are hard to be revealed in short duration because emotion does not change rapidly as phonetic feature does. Therefore, automatic segmentation for emotion feature extraction is proposed in this paper. Automatic segmentation is used for estimating boundaries between phonemes based on "spectral variation function method" and grouping phonemes based on "average energies per frames". In simulation result, it showed that automatic segmentation is useful for emotional feature extraction from spoken sentence.

1. INTRODUCTION

Human-Robot Interaction (HRI) is a promising issue in intelligent robots such as service robot, personal robot, and so on. For practical use of intelligent robots, a friendly and instinctive interface between humans and robots is required. The user who is interested in intelligent robot can be an expert on program or robot control, but usually he/she is not familiar with controlling and programming robots. Therefore, it is necessary for a personal robot to communicate naturally with an owner and reprogram itself based on the communication. There are several ways in which humans interact with each other; speech, eye contact, gesture, facial expression and etc. Among them speech is a fundamental communication method in human-human interaction. This is because people can readily exchange information without the need for any other tool through speech (Nwe et al., 2001).

Considerable effort was spent on developing friendly and instinctive interface using speech with emotion. According to Kostov and Fukud (2000), graphical user interface (GUI) will pay greater attention to the subjective user experience including the emotional impact of computer user in the future. For example, human computer interfaces could be made to respond differently according to the emotional sate of user. There is a good overview of socially interactive robots can be found (Fong et al., 2003) and also the importance of emotion recognition for human computer interaction is commonly agreed (Pantic and Rothkrantz, 2003). Therefore, it is useful and necessary that studies for speech signal processing including speech emotion recognition.

Although human easily understands context and emotions in speech, speech signal processing and language understanding automatically are not easy tasks in the-state-of-the-art. For speech recognition or speech emotion recognition, proper features should be extracted from speech signal. These features should be robust in noisy environments and distinguishable between different phonemes or emotions. In addition, proper classifier should be designed from analyzing training samples. Hidden Markov model (HMM), support vector machine (SVM), Gaussian mixture model (GMM), and neural network (NN) are representative classifier in speech processing areas. First of all, segmentation of speech signal, however, is needed. When speaker produces speech signal, the vocal tract of speaker is varied with time. In other words, the system of speech production is time-variant. Therefore, windowing is required for processing the signal frame by frame in speech processing area.

Many researchers noticed that the system can be assumed to be time-invariant in a short interval between 20ms ~ 40ms (Quatieri, 2002). So most researchers have used a window with fix frame size which is equal to 20ms ~ 40ms. However, there is a problem in fix frame method when applying to speech emotion recognition in sentence-based or continuous dialogue. It has several silence periods between phonemes and several frames which have overlaps with silence and signal. Furthermore, it has transient periods between phonemes. In a frame, the silence and transient periods

usually cause to extract distorted feature values which should be eliminated.

Therefore, we proposed automatic segmentation which finds boundaries between neighbouring phonemes and makes it easy to eliminate silence and transient periods in continuous speech from spoken sentence in this paper.

2. SYSTEM SETUP

2.1 Target System – T-ROT

T-ROT realizes a SilverMate service robot for elder people with capabilities of self-navigation, map-building, manipulating some objects and HRI. T-ROT is equipped with two stereo cameras, sixteen sound localization multi-microphone sensors, one sound recognition sensor, two manipulators, and touch screen & TFT LCD monitor, which allow T-ROT to interact with human, as presented in Fig. 1.

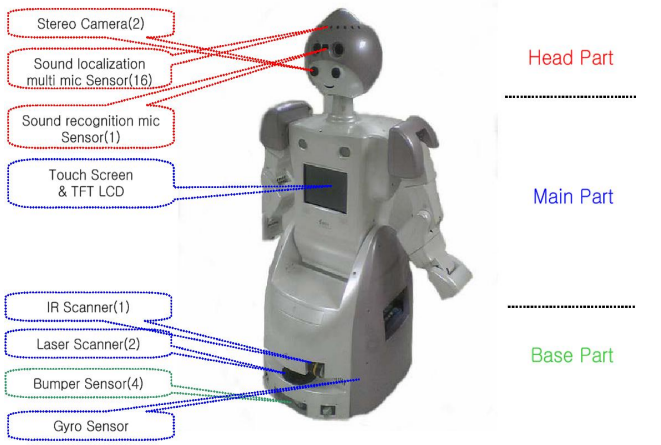

Fig. 1. Equipments of T-ROT

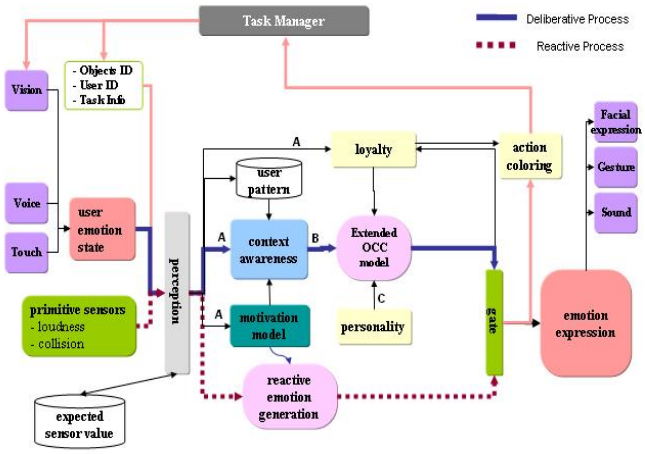

Fig. 2. Emotional Interaction Framework of T-ROT

2.2 Emotional Interaction Framework

Fig. 2. indicates the emotional interaction framework for T-ROT. The information of human emotional state can be extracted from various perceptual modalities such as vision, voice and touch sensor. According to the surrounding information such as user model, loyalty, and personality, robot's emotion is appropriately generated and expressed by face, gestures and voice. The robot's behaviors are also modified by user's emotional states. More details are described in (Lee et al., 2005).

3. AUTOMATIC SEGMENTATION

According to Sharma and Mammone (1996), most automatic segmentation uses the associated linguistic knowledge, such as the spoken text and/or phonetic string. On the other hand, Sharma and Mammone (1996) proposed blind speech segmentation which does not need any linguistic and phonetic information. Instead of linguistic and phonetic information, they use spectral variation function (SVF) method. SVF uses that peak is observed at boundaries between phonemes. However, it usually makes more segment boundaries than phonemic changes. Therefore, they combined SVF method with other algorithm such as convex hull method, level building dynamic programming (LBDP), and normal decomposition method.

Convex hull method detects the boundaries of syllables in a given speech signal (Mermelstein, 1975). The algorithm works on a subjective loudness function to find the number of syllabic units. Therefore, the result is quite different based on definition of loudness function.

LBDP is used to estimate the number of segmentation between minimum which comes from convex hull method and maximum from SVF method. In addition, normal decomposition method is used to detect the optimal boundaries between segmentation.

However, this method is complicated and costly because it needs to find optimal solutions at each sentence. Also, there are subjective factors to define loudness function in convex hull method and to define peak in SVF method. Therefore, we combined SVF method with "average energies per frame" which shows average energy values in segmentation.

The process of proposed method is presented as following:

First, the whole speech signal was captured from spoken sentence. After that, filtering was performed for noise reduction and cutting off the range that human can not perceive. Chebyshev filter type was used as bandpass filter ranging from 300 Hz ~ 3.2 kHz in this work. Fig. 3. shows the captured signal after filtering.

Second, we obtained the time derivative of cepstral coefficients to use SVF method. In SVF method, we used every local maximum as a peak to divide signal as much as possible. These peaks indicate where phonetic features changed a lot. Therefore, the candidates for segmentation

boundaries were chosen based on SVF. Fig. 4. shows the SVF and the peaks of SVF.

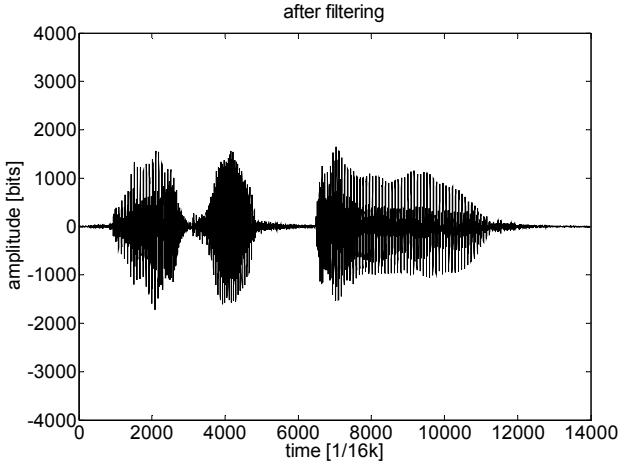

Fig. 3. The captured signal

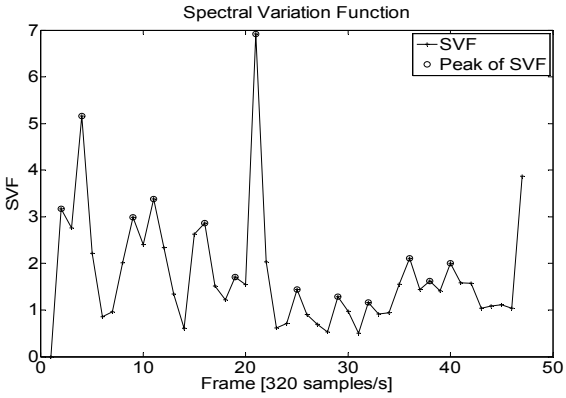

Fig. 4. SVF and peaks of SVF

Next, the position corresponding to peaks of SVF were used to divide speech signal (Fig. 5.). Then, each division was subdivided into frames using 20ms windowing. At each frame, the level of energy was calculated using Eq (1).

$$E(n) = \sqrt{\sum_{i=1}^{N} X^2(i)} \qquad (1)$$

Where n = frame index.
i = sample index.
X = speech signal
E = energy

After that, the sum of energy level in division was divided by number of frames in division. We used this value as a representative in a division and called it "average energies per frame". The average energies per frame was presented in Fig. 6.

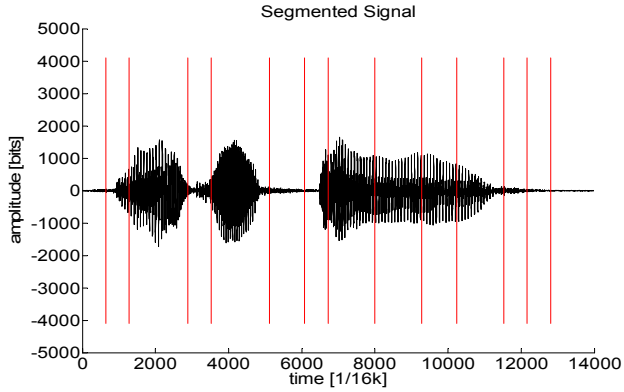

Fig. 5. Boundaries of segmentation

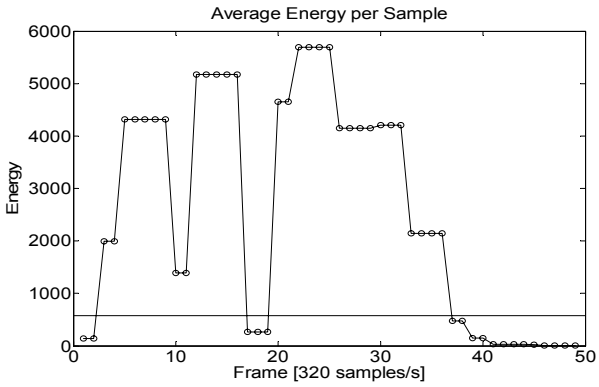

Fig. 6. Average energies per frame

Finally, we eliminated the signal below 10% of maximum value because it is almost same as silence period. As a result, we got the segmented speech signal is obtained (Fig. 7.).

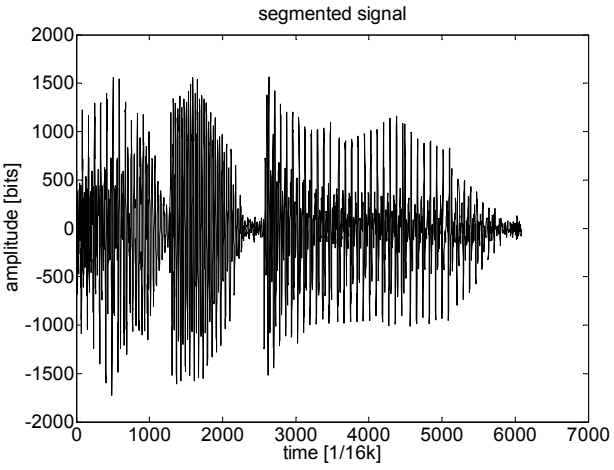

Fig. 7. Segmented speech signal

4. DATABASE

The Korean database was used in our experiment. The database was recorded in the framework of the G7 project in

Korea. In this database fifteen male speakers and fifteen female speakers were recorded.

4.1 Corpus of Database

The corpora contain short, medium, and long sentences that are context independent. The sentences are chosen upon consideration of the followings:

1) The sentence is able to pronounce in several emotional states.
2) The sentence can express the emotions naturally.
3) The database should contain all phonemes of Korean.
4) The database should contain several dictions such as honorific words.

4.2 Speech Materials

The corpora were recorded over four basic emotions. The recorded emotions are neutrality (N), joy (J), sadness (S), and anger (A). These emotions are defined in the previous section.

The database contains speech recorded in three iterations. All speech was recorded four times and the worst record among those was discarded. The aim of this step is to filter clumsy wording and maintain consistency of the speech.

The database contains 5400 sentences (10 subjects, 4 emotions, 45 contexts, 3 iterations). The subjective evaluation tests for all corpora were performed.

4.3 Recording Conditions

The subjects were amateur actors and actresses who have practiced emotional expression and they were selected according to their ability of expression. The recordings were made in a silent experimental environment with DAT. The sampling frequency was 48 kHz with quantization of 16 bits. Feature calculation used recordings that were decimated to a sampling rate of 16 kHz using an eight-order Chebyshev filter.

4.4 Subjective Evaluation Test

Subjective evaluation tests were made for the database. The subjective evaluation test included 30 listeners. The listeners were engineering students from Yonsei University in Korea.

Each listener decided which emotion corresponded to each utterance. The samples were played randomly. 10 listeners made decisions for each utterance.

The results of the subjective evaluation test showed, on average, 78.2% accuracy. The confusion matrix of the test is shown in table 1. The accuracy was highest for the emotion of sadness at 92.2%, whereas joy was the worst at 57.8%. Many errors were made in discriminating between joy and anger and between neutrality and anger. These results are caused by the difference in the basis for neutrality.

Table 1. Human Performance

[%]	N	J	S	A
N	83.9	3.1	8.9	4.1
J	26.6	57.8	3.5	12.0
S	6.4	0.6	92.2	0.8
A	15.1	5.4	1.0	78.5
Overall			78.2	

5. SPEECH EMOTION RECOGNITION

In this study, we evaluated performance of speech emotion recognition system with proposed automatic segmentation method. To evaluate, we used GMM and prosodic features such as pitch and energy which were generally used in both speech recognition and emotional speech recognition.

5.1 Pitch

Pitch is a fundamental frequency that comes from periodic vibration of vocal cords. Pitch contour depends strongly on the emotional state of speaker. For example, pitch of speaker in a stress changes rapidly and goes high. Usually it shows big difference between anger and sadness. In this paper, pitch value is estimated by Simple Inverse Filtering Tracking (SIFT) algorithm (Markel, 1972). The flow chart of SIFT is presented in Fig. 8.

5.2 Energy

Energy is a fundamental emotional feature because there are high correlations between energy and emotion. People usually speak loudly in a state of high arousal; on the other hand, they speak quietly when they feel sad or calm. To know the energy level in speech signal, we simply calculate the norm of speech signal in segmentation.

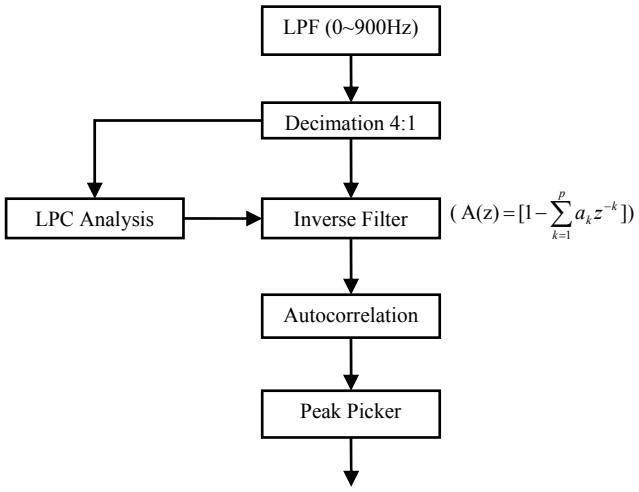

Fig.8. Flow chart of SIFT algorithm

5.3 Evaluation

For evaluation, k-fold cross-validation is conducted. Cross-validation is usually used for classification problem because it reduces the dependency of training samples and validation samples. In k-fold cross-validation, all samples in database are divided into k subsets. One of k subsets is used as training samples and the other subsets are used as validation samples. The cross-validation process is then repeated k times with each of the K subsets used exactly once as the validation data. Finally, the average of k results from the folds is used for a single estimation.

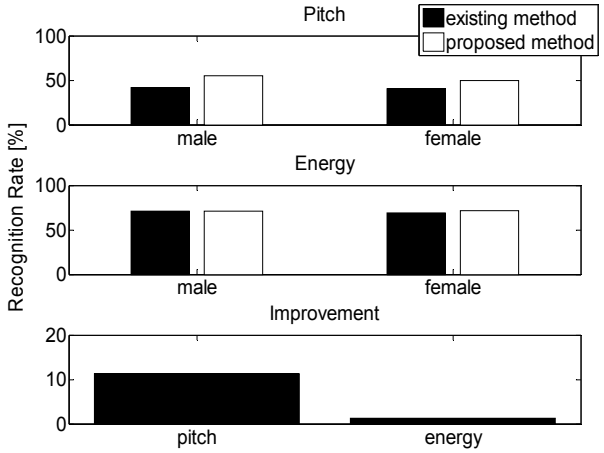

Fig.9. Simulation result

The simulation results are presented in Fig.9. As shown in Fig. 9, the recognition rate is improved in both case but further improved in pitch. This is because SIFT algorithm is more affected by segmentation. Energy is determined by envelop of signal and envelop does not change rapidly frame by frame. However, pitch changes rapidly and silence period makes more errors in pitch than errors in energy. In silence period, it is difficult to detect the periodicity of signal. Therefore, automatic segmentation is more needed in pitch detection.

6. CONCLUSION

In this paper, automatic segmentation algorithm for emotion feature extraction was proposed and it was applied to extract emotional features for speech emotion recognition. The proposed algorithm uses SVF method and average energies per frame, but it does not obtain an optimal solution for segmentation to reduce computational cost. To compare with fix frame method, we applied both fix frame method and proposed method to speech recognition problem using pitch and energy. The result showed proposed method was useful in speech emotion recognition problem. However, proposed method does not guarantee exact boundaries because it is not important in speech emotion recognition. Instead, proposed method discarded vague periods to focus on whole sentence rather than specific frame or period.

For future works, we plan to implement speech emotion recognition into service robot. It will show that speech emotion recognition is useful for friendly interacting between humans and robots.

Acknowledgement

This research was supported by the Brain Korea 21 Project in 2007 and performed as a part of the Intelligent Robotics Development Program, on of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

REFERENCES

Fong, T., Nourbakhsh, I., Dautenhahn K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems,* **42**, 143-166.

Kostov, V., Fukuda, S. (2000). Emotion in user interface, voice interaction system. In: *Proc. IEEE International Conference on System, Man, SMC 2000, and Cybernetics,* **2**, 798-803.

Lee, K.W., Kim, H.R., Yoon, W.C., Yoon, Y.S., Kwon, D.S. (2005). Designing A Human-Robot Interaction Framework For Home Service Robot. In: *Robot and Human Interactive Communication, RO-MAN 2005,* 286-293

Markel, J.D. (1972). The SIFT Algorithm for Fundamental Frequency Estimation. *IEEE Trans. On Audio and Electroacoustics,* **AU-20 (5)**, 367~377.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of Acoustical Society of America,* **58 (4)**, 880-883.

Nwe, T.L., Wei, F.S., de Silva, L.C. (2001). Speech Based Emotion Classification. In: *Proc. IEEE Region 10 International Conference, TENCON 2001,* **1**, 297-301.

Pantic, M., Rothkrantz, L.J.M. (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. In: *Proc. Proceedings of the IEEE*. 415-422.

Sharma, M., Marmmone, R. (1996). "Blind" Speech Segmentation: Automatic Segmentation of Speech without Linguistic Knowledge. In: *Proc. 4th International Conference on Spoken Language Processing, ICSLP 1996,* **2**, 1237-1240.

Quatieri, T.F. (2002). Discrete-Time Speech Signal Processing, Principles and Practice, Prentice Hall Signal Processing Series, Upper Saddle River, NJ 07458.