Improvement of Emotion Recognition from Voice by Separating of Obstruents

Eun Ho Kim, Kyung Hak Hyun and Yoon Keun Kwak

***Abstract*—Previous researchers in the area of emotion recognition have classified emotion from the whole voice. They did not consider that emotion features vary according to the phoneme. Hence, in the present work, we study the characteristics of phonemes in emotion features. Based on the results, we define the obstruents effect, which is a negative effect resulting from increased feature values. We then recognize emotion from the voice by separating obstruents rather than from the whole voice. By separating obstruents, we could improve the emotion recognition rate by about 4.3%.**

I. Introduction

Recentrly, emotional robots have been developed as human-friendly robots. Human-friendly robots require several capabilities such as manipulation, perceptivity, serving, and communication. Communication is a very important capability not only with regard to the communication of words but also emotions. In the field of human-robot interactions, recognition of emotion is an important task. In particular, recognition of emotion from voice is a critical challenge, because speech is the fundamental mode of human communication. Hence, in this paper, we focus on recognizing human emotion from voice. Through an analysis of the characteristics of phonemes in emotion features we propose the concept of separation of obstruents for improving the emotion recognition

II. Previous Work

A. Emotion feature

Voice features related with emotion can be mainly divided into two categories, prosodic and phonetic features. Prosodic features are not related with content of sentences and include such aspects as energy, pitch, and tempo. Previous study revealed that prosodic features are related with emotion [1]. For example, Xiao Lin [2] recognized emotion from voice using pitch. V. Kostov [3] developed a sentence independent system using the pitch, energy, and tempo. In 2004, Dimitrios Ververidis [4] extracted 87 static features from the dynamic features of frequency, pitch, and energy and evaluated the emotion recognition performance for each static feature.

Phonetic features contrast with prosody features such as linear prediction coefficient (LPC), which is using for modeling vocal tract, mel-frequency cepstrum coefficient, and so on.

B. Previous work

Most research on emotion recognition from voice focuses on speaker dependent and sentence independent systems and average recognition rates of 70% to 95% have been achieved. In the case of speaker independent system, the ASSESS system developed by S. McGilloway [5] displayed a 55% recognition rate. It is not unexpected that the emotion recognition rate decreases in the speaker independent system comparison with speaker dependent. Because even human have low recognition rate, roughly 60%, when attempting to recognize the emotion of unknown speakers [6].

In the case of the speaker dependent system SpeakSoflty [7], a recognition rate of 70% is achieved for five emotions using a neural network whereas MEXI [8] recognized five emotions at an 84% rate for the speaker dependent system and 60% for the speaker independent system based on fuzzy rule.

III. Obstruents Effect

A. Necessary to study the characteristics of each phoneme

In the field of emotion recognition from voice, most works do not investigate the characteristics of each phoneme or consider the characteristics of each phoneme, because the target system is a sentence independent system. Hence, to date, researchers in this field have extracted emotion features from whole sentences. They only separate silence from the whole sentence. However, emotion features except energy are used in the field of speech recognition and are related with frequency. Hence, emotion features vary with not only emotion but also at a phoneme.

If emotion features vary with phoneme rather than emotion, emotion recognition can easily fail because feature values changed by emotion are hidden by the phoneme. Hence, the system cannot readily detect the change of feature values changed by emotion. For example, in the spectral center, which is the average frequency weighted by the acoustic

Manuscript received March 15, 2006. This work was supported in part by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea

Eun Ho Kim is with the Mechanical Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, Korea (e-mail: galaxier@kaist.ac.kr).

Kyung Hak Hyun is with the Mechanical Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, Korea(e-mail: cromno9@kaist.ac.kr).

Yoon keun Kwak is with Mechanical Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, Korea (corresponding author to provide phone: 82-42-869-3212; fax: 82-42-869-5201; e-mail: ykkwak@ kaist.ac.kr).

1-4244-0565-3/06/$20.00 ©2006 IEEE.

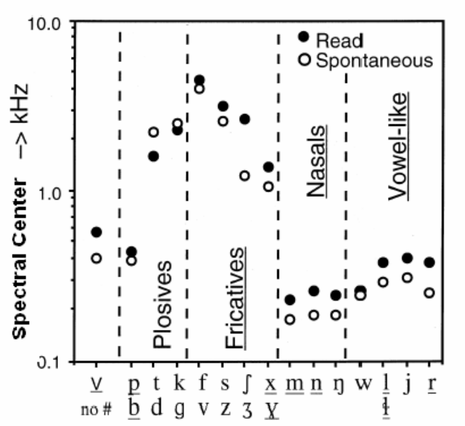

Fig. 1. Spectral center for each phoneme

power, generally sadness has a low value of about 360Hz whereas anger has a high value of about 550Hz (see Table 1). And ‘f’ has a high spectral center with a value of about 4000Hz (see fig. 1). Hence, the spectral center values are determined by the frequency of ‘f’, not by the emotion. In other words, if someone speaks a sentence with an emotional state then the spectral center of this speech is determined by how many ‘f’ sounds are uttered, not by the emotion state of the speaker.

Hence, even if the emotion recognition system is sentence independent it is necessary to study the variation of the emotion features by phonemes, because most emotion features vary by phoneme.

B. *Obstruents effect*

In 1999, Van Son [9] extracted the spectral center for each phoneme. From fig. 1 we see that obstruents (plosives and fricatives) have a high spectral center, from roughly 1 kHz to 4 kHz, in comparison with other phonemes. However, variation of the spectral center by emotions is lower than 200Hz [10]. Hence, obstruents have a negative effect on emotion recognition from voice.

Variation of the spectral center by emotion is countervailed by obstruents because obstruents increase the spectral center regardless of emotions. Hence, the emotion recognition rate is decreased by the obstruents. We define this negative effect as the obstruents effect.

Not only the spectral center but also spectral flatness measure (SFM) defined as the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum and LPC vary with the phoneme. SFM and LPC are also related with frequency, and hence obstruents have high SFM and LPC values (see Table 2). Thus, the obstruents effect affects not only the spectral center but also SFM and LPC. This means that the emotion recognition rate using SFM and LPC is also decreased by obstruents.

Fig. 2 indicates the LPC 1st order value of the 2nd female and 5th sentence. In fig. 2, the star denotes non-obstruents, which have a spectral value lower than 1 kHz whereas the circle denotes the obstruents, which have a spectral value higher than 1 kHz. From fig. 2, we can see that most obstruents are far from the median value. This is one example of the obstruents effect.

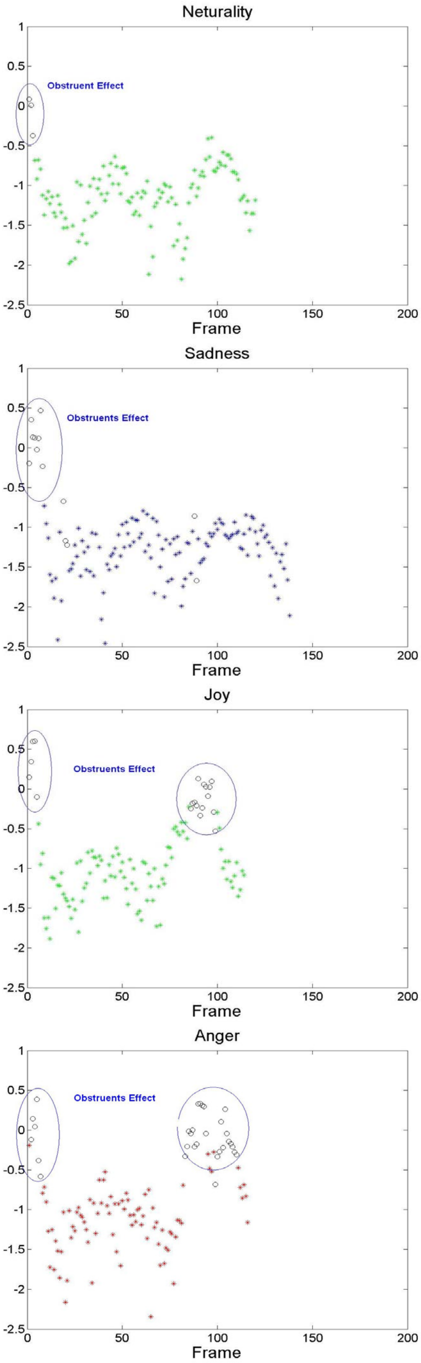

Fig. 2. Obstruents effect on 1st order LPC for four emotions

From table I and II, we can see that the variation between obstruents and non-obstruents is higher than the variation between two emotions, not only in terms of spectral center but also in SFM and LPC. For example, the variation of SFM between joy and anger is about 0.0288 whereas the variation of SFM between obstruents and non-obstruents is about 0.18, or about six times larger than variation of two emotions. And the variation of LPC between neutrality and sadness is about

TABLE I
VARIATION OF SFM, LPC AND SPECTRAL CENTER WITH EACH EMOTIONS

SFM				
	Neutrality	Joy	Sadness	Anger
Mean	0.3305	0.3326	0.3320	0.3614
Std.	0.0145	0.0149	0.0163	0.0169
LPC				
	Neutrality	Joy	Sadness	Anger
Mean	-1.2587	-1.1504	-1.4405	-1.1875
Std.	0.3983	0.3997	0.3242	0.4106
Spectral Center (Hz)				
	Neutrality	Joy	Sadness	Anger
Mean	492.9	534.1	360.8	554.2
Std.	201.0	219.4	160.5	227.2

TABLE II
VARIATION OF SFM, LPC AND SPECTRAL CENTER BETWEEN OBSTRUENTS AND NON- OBSTRUENTS

	Not obstruents	Obstruents
SFM	0.34	0.52
LPC	-1.2701	-0.4319
Spectral center (Hz)	461.4	1810.1

0.2 whereas the variation of LPC between obstruents and non-obstruents is about 0.84, four times larger. This means that SFM, LPC, and spectral center are influenced by the obstruents effect.

IV. FEATURE EXTRACTION WITH SEPARATION OF OBSTRUENTS

A. Separation of obstruents

Thus far in emotion recognition from voice, researchers have extracted emotion features from whole sentences except silence. They do not consider the characteristics of phonemes, especially obstruents. However, as outlined above, obstruents within the sentence decrease the emotion recognition rate. Hence, in this paper, we attempt to improve the emotion recognition rate by separating obstruents from the sample sentences.

Before extracting emotion features, we remove the obstruents effect via separation of obstruents. From fig. 1, obstruents can be separated by removing a frame having a spectral center higher than 1 kHz.

B. Feature extraction

All features are computed over a rectangular window of 20msec duration. The algorithms are described in detail in the following.

The SFM used for determining the noise-like or tone-like nature of the speech signal. The SFM is numerically calculated by equation (1). And LPC is extracted using the Matlab toolbox [11].

TABLE III
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			

$$SFM_j = \frac{\left(\prod_{k=1}^{N} X_j(k) \right)^{\frac{1}{N}}}{\frac{1}{N} \sum_{k=1}^{N} X_j(k)} \tag{1}$$

The spectral center is calculated by dividing $\int f \bullet E(f)\,df$ by $\int E(f)\,df$ or numerically calculated by equation (2), where f and f_i are the frequency in Hertz, and $E(f)$ and E_i are the spectral power as a function of the frequency.

$$Spectral\ Center = \frac{\sum f_i \bullet E_i}{\sum E_i} \tag{2}$$

V. EXPERIMENTAL RESULTS

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; its principles are as follows [12]:

- 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 a pulse code modulation filter.

To verify how accurately the database reflects a speaker's emotions, experiments were conducted at Yonsei University utilizing the subjective emotion recognition of human subjects [12]. Table III shows the results of these experiments. The recognition rate was unequal for each emotion; for the recognition characteristics, sadness was well recognized but joy was not.

B. Experiment method

The databases consist of 5400 sentences (45 dialogic sentences times three repetitions times four emotions times ten speakers comprised of 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, recognition experiments were conducted on the remaining 80 percent of the data.

Static features were extracted from the dynamic features (or primary features). In the SFM and the spectral center, we extracted 12 static features (mean, maximum, median and standard deviation of dynamic features, delta dynamic features and del-delta dynamic features) from dynamic features. The dimensions of the static features, i.e., 12, were reduced to 3 using PCA and LDA [13]. LPC static features having 12 dimensions were extracted from the mean of LPC 12th order dynamic features. We extracted a 3-dimension feature vector which is used for emotion recognition using LDA from the static features. For the classification, we used the bayes classifier (that is, a normal distribution based quadric classifier) [14].

C. Experimental results

Table IV shows the experimental results of SFM, LPC, and spectral center separating the obstruents comparison with whole sentence and shows the improvement in the emotion recognition rate attained using separation of obstruents.

In the case of SFM, after separating the obstruents, the emotion recognition rate for all speakers except male 1, female 1 and 2 was improved. The average improvement is about 5.3%.

In the case of LPC, after removing the obstruents effect, the emotion recognition rate for all speakers except male 3, 4 and female 1 was improved. The average improvement is about 2.0%. When humans utter obstruents, the vocal tract is changed more than when uttering non-obstruents. Hence, the LPC value of obstruents is significantly different from non-obstruents, as shown in Table II because the LPC is the model of vocal tract. This difference causes the obstruents effect, and thus the emotion recognition rate increases after separating the obstruents.

In the case of spectral center, when we separate the obstruents, the emotion recognition rate for all speakers except female 2 was increased. The average increase is about 5.7%. Obstruents have a high spectral center, as shown in fig. 1, because obstruents rapidly flow small area of constriction. Hence, the obstruents effect also occurs in the spectral center.

From the experiment with SFM, LPC, and spectral center we can clearly see the obstruents effect and usefulness of the separating of obstruents for emotion recognition from voice.

TABLE IV
IMPROVE MENTS OF EMOTION RECOGNTION RATE
BY SEPARATING OF OBSTRUENTS (%)

SFM			
	Before	After	Improvement
Male 1	55.0	51.9	-3.1
Male 2	50.0	62.4	12.4
Male 3	81.4	87.6	6.2
Male 4	70.0	73.3	3.3
Male 5	60.5	80.0	19.5
Female 1	68.8	56.2	-12.6
Female 2	74.0	71.2	-2.8
Female 3	53.8	62.1	8.3
Female 4	63.6	65.3	1.7
Female 5	45.2	65.2	20
Average	62.2	67.5	5.3
LPC			
	Before	After	Improvement
Male 1	82.6	84.1	1.5
Male 2	76.2	80.0	3.8
Male 3	85.7	84.3	-1.4
Male 4	71.9	70.2	-1.7
Male 5	76.0	83.1	7.1
Female 1	68.1	67.1	-1.0
Female 2	79.3	80.1	0.8
Female 3	74.8	76.4	1.6
Female 4	82.6	86.7	4.1
Female 5	72.1	76.9	4.8
Average	76.9	78.9	2.0
Spectral Center			
	Before	After	Improvement
Male 1	61.4	66.7	5.3
Male 2	66.7	76.4	9.7
Male 3	65.5	72.4	6.9
Male 4	54.0	63..8	9.8
Male 5	74.8	78.6	3.8
Female 1	60.0	65.7	5.7
Female 2	61.4	61.0	-0.4
Female 3	66.0	67.6	1.6
Female 4	44.8	54.0	9.2
Female 5	66.2	71.4	5.2
Average	62.1	67.8	5.7

VI. CONCLUSIONS

In this paper, we analyzed the characteristics of phonemes in emotion recognition from voice. From the analysis we found that in the frequency related feature, obstruents among the sentence decrease the emotion recognition rate, because obstruents have a high feature value. We refer to this effect as the obstruents effect.

We propose an algorithm to separate the obstruents using the spectral center. Using this algorithm, we acquire an average improvement of about 4.3% for ten speakers with SFM, LPC, and spectral center.

In the future, we can identify other important phonemes from the study of characteristics of phonemes. Using the concept of separating phonemes that have negative effects we can expect improvement in the emotion recognition from voice with existing emotion features.

ACKNOWLEDGMENT

This 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 Commerce, Industry and Energy of Korea

REFERENCES

[1] R. Plutchik, Emotion and Life: Perspectives from Psychology, Biology and Evolution, American Psychological Association Press, 2003.

[2] X. Lin, Y. Chen, S. Lin, and C. Lim, “Recognition of Emotional state from Spoken Sentences,” *IEEE Multimedia signal processing* pp. 469-473, 1999.

[3] V. Kostov, S. Fukuda, “Emotion in User Interface, Voice Interaction System,” *System Man and Cybernetics, IEEE conf.* vol. 2 pp.798-803, 2000.

[4] D. Ververidis, C. Kotropolos, “Automatic Emotional Speech Classification,” *IEEE int. conf. Acoustics, Speech and Signal Processing,* 2004.

[5] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, S. Stroeve, “Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark,*” in ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[6] K. Scherer, “Vocal Communication of Emotion: *A Review of Research Paradigms,” in Speech Communication*, 40(2003), 227-256, Elsevier 2003.

[7] V.A. Petrushin, “Emotion in Speech: Recognition and Application to Call Centers”, *Proceeding of the 1999 Conference on Artificial Neural Networks in Engineering*, 1999.

[8] A. Austermann, N. Esau, L. Kleinjohann, B. Kleinjohann, “Fuzzy Emotion Recognition in Natural Speech Dialogue”, *Robot and Human Interactive Communication IEEE 14th workshop*, 2005.

[9] R.J.J.H. van Son, Louis C.W. Pols, “an Acoustic Description of Consonant Reduction,” *Speech Communication*, pp.125-140, 1999.

[10]M. Kienast, Walter F. Sendlmeier, “Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech,” *proc. of the ISCA Workshop on Emotion and Speech,* 2000.

[11]http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[12]B. S. Kang, “Text Independent Emotion Recognition Using Speech Signals,” *Yonsei Univ.* 2000.

[13] J. Kitter, “A method for determining class subspace,” *Information Processing Letters,* vol. 6, issue 3, pp. 77-79, 1977.

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