Feature Extraction Method for Predicting Depression by Frequency Domain Analysis

Eun-Joo Seo [a], Kwang-Seok Hong [b]
College of Information and Communication Engineering, Sungkyunkwan University
Suwon, South Korea
[a]eunjoo1125@skku.edu, [b]kshong@skku.ac.kr

***Abstract*—In this paper, we propose a feature extraction method, which subdivides feature vectors into three frequency regions of 300-1000Hz, 1000-2000Hz, and 2000-3000Hz. The range and mean of intensity are extracted for each frequency region. By so doing, we can compensate the defect of increasing the intensity value, when a person intentionally increases his or her vocalization. Previous studies extracted the slope and correlation of the glottal flow spectrum from the 300-3000Hz region. But, we extract the slope and correlation from each separated frequency region. The overall experimental results show 92.85% for men, and 92.08% for women. The proposed method enhances the respective classification accuracy by 6.73% for men, and 8.09% for women.**

Keywords-depression, speech Analysis, Random Forest, frequency Domain

I. INTRODUCTION

Recently, contents that recognize a user's emotion state through his or her voice are being utilized, such as the intelligent personal assistant and knowledge navigator Siri of Apple's iOS. Also, researches to predict the user's illness and mental health state are actively being conducted. For example, the U.S. Massachusetts Lab implemented a computer program that can predict Parkinson's disease [7]. The research team used the fact that the sounds of breathing and voice tremor weaken in early Parkinson's disease. On the domestic front, Dong-Uk Cho's lab [8] conducted researches about speech features related to the five viscera (liver, lungs, heart, kidneys, and spleen).

In addition, studies of the relations between depression and voice are being actively conducted abroad. Previous studies compared depression groups and normal groups for speech features like intensity [1], pitch [1,2], and formant frequency and formant bandwidth [3]. But most previous studies used databases that consisted of voices received in quiet environments; and the databases were composed of voices uttering fixed sentences. Therefore, it is difficult to apply these studies to the real-world environment.

In this paper, we analyze the relationship between depression and the user's voice, by using voice that is naturally uttered. We design a system that can predict the degree of depression, by analyzing speech data in real time. We extract a total of 55 feature vectors, including feature vectors that are separated into 3 frequency regions from the user's speech data. We then use a random forest classifier to predict the degree of depression. Training models are created according to gender. When inputting the data, the voice is compared with the training model that matches the user's gender.

II. RELATED WORK

A. Emotion Recognition from Speech Analysis

Pao et al. [5] studied Mandarin emotional recognition. Their study obtained the best recognition of 84.2% with LPC, MFCC, LPCC, and LFPC features, and used SVM and NN classifiers. Pan et al. [6] observed that the feature combination of MFCC+ MEDC+ Energy has the highest accuracy rate on the Chinese emotional database.

B. Analysis of Depression patient's voice

The feature parameters that were used in previous studies on the acoustic characteristics of depressive patients vary. A study of formant frequency and formant bandwidth [3] reported that men in the normal group had a lower value of second formant frequency and third formant bandwidth than men in the depression group. Also, the study observed that for women, the normal group had a lower value of first formant frequency, third formant frequency, first formant bandwidth and second formant bandwidth, than did the depression group.

A study of fundamental frequency and intensity [1] showed strong negative correlation between two parameters and the BDI score (the BDI score represents the depression scale).

A related study [4] reported that the slope that is obtained from the glottal flow spectrum and jitter are the best features to separate the depression from the normal group.

But most previous studies utilized speech data that consisted of speaking fixed sentences. Also, the speech data was collected in a confined environment on the depressive patient and normal groups. If the subjects speak fixed sentences, they cannot express their emotion. So,

such studies are difficult to recognize emotion about speech data in mobiles.

III. System Design and Implementation

Figure 1 shows that the overall system consists of an Input module, Feature Extraction module, and Recognition module. Each module is described below.

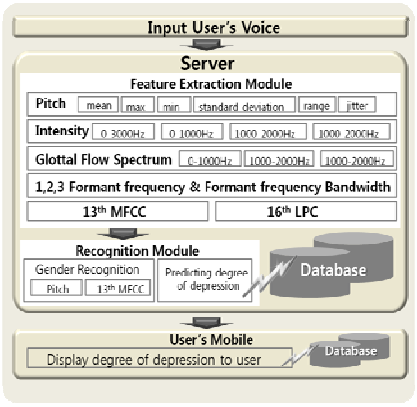

Figure 1. The System Architecture

A. Feature Extraction module

The speech data that is received from the user is subdivided into 256 samples per frame. Each frame was applied to the 256 length of the Hamming window. A total of 55 different features (mean of intensity, range of intensity, mean of pitch, range of pitch, max. and min. value of pitch, standard deviation of pitch, slope and correlation of glottal flow spectrum, formant frequency, formant bandwidth, jitter, 13-order MFCC, and 15-order LPC) were extracted from each segmented speech signal.

1) Intensity mean and range: Normal human vocalization frequency is 3000Hz. When analyzing a speech signal by spectrum in the time domain, we observe that the constituents are concentrated in the 300-3000Hz range. If a person intentionally increases his or her vocalization, the intensity values are higher than usual. So, in this paper, we extracted the intensity value in the 300-3000Hz range, which is the normal human vocalization frequency. We subdivided the 300-3000Hz range into three frequency regions, of 300-1000Hz, 1000-2000Hz and 2000-3000Hz, and extracted the mean and range of intensity in each frequency region.

2) Intensity mean and range: For analyzing the glottal flow spectrum, we extracted features based on the method of glottal flow spectrum proposed in a previous study [4]. While their study measured the slope and correlation in the 300-3000Hz region, we separate the frequency regions into 300-1000Hz, 1000-2000Hz, and 2000-3000Hz, and extract the slope and correlation in each region.

3) Formant frequency: Formant bandwidth, and 16-order LPC

Each frame was carried out by 16-order LPC analysis, and we determined peak points of the graph from LPC analysis. The position of the peak point became the formant frequency. We obtained the value of formant bandwidth, as well as the formant frequency. Also, we used a 16-order LPC coefficient, by performing LPC analysis.

4) Pitch: The fundamental frequency (f_0) was extracted by each frame. We calculated the mean, range, max., min., and standard deviation for the fundamental frequency (f_0) of each frame. Also, we used Eq. 1 to calculate the jitter, which was the frequency rate of change about the fundamental frequency of each frame.

$$\text{Jitter}(\%) = \frac{\frac{1}{N}\sum_{i=1}^{N-1}|F_i - F_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}F_i} \quad (1)$$

5) The 13-order MFCC: The MFCC coefficient is calculated by dividing the frequency based on the human auditory organ. After extracting from each frame a 13-order MFCC through the algorithm of Figure 2, we calculated the mean of each coefficient by frame.

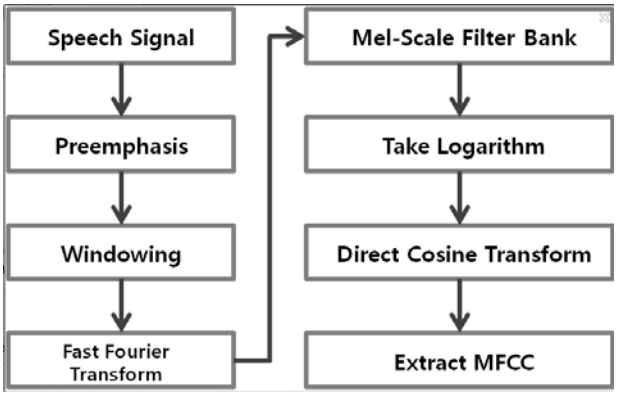

Figure 2. The MFCC

B. Recognition Module

We applied a total of 55 different features from the user speech signal to a random forest classifier to recognize gender, and predict the degree of depression. The proposed system was applied to the model with a number of different features for the recognition of the two.

For gender recognition, we used the13-order MFCC, mean of pitch, and range of pitch. The features were then applied to the random forest classifier.

In our study, when people were speaking, we observed a difference in pitch and intensity according to gender. So, we created both a men's and a women's training model. Through gender recognition, the input signal was compared with the training model that matched the user's gender. To predict the degree of depression, we applied a number of different speech parameters to a random forest. The value of predicting depression was calculated

depending on the number of random forest trees that voted on the depression class.

$$\text{Degree of depression}=\left(\frac{D}{N}\right)*100 \quad (2)$$

The value was calculated in percent value using Eq. 2, where N is the total number of trees, and D is the number of votes on the depression class.

After the extracted feature parameters were applied to the random forest, we obtained each speech recognition result and the number of votes for each rank, as Table 1 shows.

TABLE 1 DEGREE OF DEPRESSION

	Result	1st rank	2nd rank	Degree of depression
Normal	Normal	211	39	(39/250)*100=15.6
	Normal	164	86	(86/250)*100=34.4
	Normal	149	101	(101/250)*100=40.4
	Depression	126	124	(126/250)*100=50.4
	Depression	168	82	(168/250)*100=67.2
Depressed	Depression	191	59	(191/250)*100=76.4

IV. PERFORMANCE EVALUATION

A. Experimental Environment and Database

To evaluate the performance of the proposed system, C++ based on Windows 8 was used for feature extraction and recognition. We used the Yonsei University speech database, which was made by 15 amateur actors. The database is composed of 1350 women's speech data, and 2428 men's speech data.

B. Test and Result

In this paper, we divided the train data and test data in the ratio of 7:3 to evaluate the performance. Also, the recognition accuracy rate seems to vary, depending on whether or not some feature parameters are included. So we compared the recognition accuracy rate of a multiple combination model, by adding some parameters to the basic parameters (case 4).

Table 2 shows that we conducted the experiment by division into eight feature combination models. Basic_Intensity and Original_Glottal flow spectrum are features that were used in related works. Proposed_Intensity and Proposed_Glottal flow spectrum are segmented features that we propose.

TABLE 2 FEATURE COMBINATION

case 1	Basic_Intensity + Pitch + Original_Glottal flow spectrum + Jitter +Formant frequency + Formant bandwidth
case 2	Proposed_Intensity + Pitch + Original_Glottal flow spectrum + Jitter + Formant frequency + Formant bandwidth
case 3	Basic_Intensity + Pitch + Proposed_Glottal flow spectrum + Jitter + Formant frequency + Formant bandwidth
case 4	Proposed_Intensity + Pitch + Proposed _Glottal flow spectrum + Jitter + Formant frequency + Formant bandwidth
case 5	Case 4 + 13-order MFCC
case 6	Case 4 + 16-order LPC
case 7	Case 4 + f_0_min + f_0_max + f_0_std
case 8	Case 4 + 13-order MFCC + 16-order LPC + f_0_min + f_0_max + f_0_std

TABLE 3 RECOGNITION RATE

	case 1	case 2	case 3	case 4	case 5	case 6	case 7	case 8
Women	83.99%	83.99%	87.93%	90.14%	88.42%	89.99%	87.43%	92.08%
Men	86.12%	84.47%	88.32%	88.32%	92.85%	85.71%	89.14%	90.93%

Table 3 shows the recognition accuracy rate for the eight feature combination models.

Cases 1 and 4 in Table 3 show that we can obtain improved recognition rates, when we use the proposed feature parameters.

The results show that the best recognition rate for women is 92.08% in case 8, and the best recognition rate for men is 92.85% in case 5. So, by creating two recognition models according to gender, we implemented a comparison and recognition of the model by gender

V. CONCLUSION

We propose a Feature Extraction Method by frequency domain to predict the degree of depression. In this paper, we performed gender recognition for input speech data as a prior process. The data was then compared with a model according to the user's gender. For men, the recognition

accuracy rate increased by 6.73% using the model that is comprised of 36 feature parameters. For women, the recognition accuracy rate increased by 8.09% using the model that is comprised of 55 feature parameters.

In future, our proposed system will contribute to monitoring services for depressive patients. The system can make an early diagnosis, and manage the patient's mental health state. By using the system, depressive patients can receive hospital management in real time.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through NRF of Korea, funded by MOE(NRF-2010-0020210) and This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2014023704).

REFERENCES

[1] Baek Y. -s., Kim S. -j., Kim E. -y., Choi Y. -l.: *Vocal acoustic characteristics of speakers with depression,* edited by Journal of The Korean Society of Speech Science, vol.4, no.1, pp. 91-98 (2012).

[2] Cannizzaro M. et al.: *Voice acoustical measurement of the severity of major depression*, edited by Brain and cognition, vol.51, no.1, pp. 33-35 (2004).

[3] France D. J. et al.: *Acoustical properties of speech as indicator of depression and suicidal risk*, edited by Biomedical Engineering, IEEE Transactions on, vol.47, no.7, p 829-837 (2000).

[4] Ozdas A., et al.: *Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk*, edited by Biomedical Engineering, IEEE Transactions on, vol.51, no.9, pp. 1530-1540 (2004).

[5] Pao T.-L., Chen Y.-T., Yeh J.-H. and Li P.-J.: *Mandarin Emotional* Speech *Recognition Based on SVM and NN*, edited by Pattern Recognition 2006, ICPR 2006, 18th International Conference on, vol.1, pp. 1096-1100 (2006).

[6] Pan Y., Shen P. and Shen L.: *Speech Emotion Recognition Using Support Vector Machine*, edited by International Journal of Smart Home, vol.6, no.2, pp. 101-107 (2012).

[7] Information on http://health.chosun.com/site/data/html_dir/2012/10/08/20121008020006.html

[8] Kim B.-h., Cho D.-u.: *Analysis of Lung Function Influence by Stimulating Ear Reflex Point Using Voice Analysis*, edited by Journal of The Korea Institute of Communication Science, vol.37, no.6, pp. 520-526 (2012).