2016 10th Workshop of Tohoku University, University of Electronic Science and Technology of China and Yeungnam University

End Point Detection on Time and Spectral Domain for Real-time Voice Detection

Jung-Seok Yoon
Dept. of Information Communication Engineering
Yeungnam University
Gyeongsan, Gyeongsangbuk-do, Korea
js920215@ynu.ac.kr

Jeong-Sik Park*
Dept. of Information Communication Engineering
Yeungnam University
Gyeongsan, Gyeongsangbuk-do, Korea
parkjs@yu.ac.kr

Abstract—**This paper proposes a voice detection technique using end point detection on time and spectral domain. In spectral domain, speech regions and non-speech regions represent quite a different energy density. Due to this property, spectral energy provides a good criterion for voice detection, thus reducing false alarm rate and false rejection rate. This paper also proposes a post-processing method for reliable end point detection. This method detects end points of speech on the basis of duration of non-speech intervals. To verify the efficiency of our proposed method, we performed real-time voice detection experiments with several participants. The proposed method showed better performance compared to the conventional method employing only a criterion like time-domain energy.**

Keywords-End Point Detection, Voic Detection, Spectral Domain, Time Domain

I. INTRODUCTION

Recently, people have tried to develop a variety of applications conveying voice user interface in order to seek convenience of human life. In particular, when using home electronic devices such as a television, people feel more satisfaction for controlling the devices with their voice [1].

Automatic speech recognition is a main technique for voice use interface. In the applications conveying speech recognition, it is very important to detect human voice correctly because the voice region enters into a recognition module. For this reason, techniques for voice detection, also known as Voice Activity Detection (VAD) have been steadily studied [2]. VAD means detecting speech regions and removing non-speech regions from input speech signals. The conventional VAD methods use various criterion like signal energy and zero-crossing rate.

A variety of studies have continuously introduced efficient voice activity detection ways. For instance, it is possible to obtain frame energy through simply processing a speech signal and use Fast Fourier Transform (FFT) methods simplifying the complex calculation of Discrete Fourier Transform (DFT) [3][4].

To determine the performance of the voice detection, two kinds of error rates are estimated: False Alarm Rate (FAR) and False Rejection Rate (FRR). FAR is an error rate in which a non-speech region is determined a speech region. On the other hand, FRR means an error rate when a speech region is ignored as a non-speech region [5].

Most VAD methods have concentrated on a single domain property of spectral domain or time domain. Thus, the methods showed poor results in terms of FAR and FRR. In this paper, we investigate the efficiency of each domain property and propose a more efficient way for real-time VAD. In particular, we search for a way of combining positive effects of spectral domain and time domain.

This paper is organized as follows. In Section II, we introduce the conventional methods for voice detection. In Section III, the proposed voice detection method is investigated in detail. Section IV explains experimental results and the performance. Finally, we conclude this study in Section IV.

II. RELATED WORKS

In this paper, we introduce several VAD methods that determine speech and non-speech regions.

A. *Cepstral distance*

In a voice detection method, cepstral distance was used based on a Euclidean distance [6]. To extract features of the cepstrum, speech signals are applied by the FFT as logarithmic scale and implemented by the Inverse Fast Fourier Transform (IFFT). Cepstral features are extracted by multiplying the cepstral window in the cepstrum domain.

B. *Zero-crossing rate*

This method detects voice regions on time domain [7]. Zero-crossing rate (ZCR) means the rate of how frequently the sign of signals is changed on the basis of zero point. This method is simple and meaningful, but it provides unreliable performance according to speech environments and speaking properties.

C. *Frame energy*

This method is the simplest way that only uses energy of each frame. Signal energy calculated in each frame is used for VAD, because of a property that speech regions indicate higher energy compared to non-speech regions. This method has been widely used, but it has weaknesses on noisy environments.

* Corresponding author

III. Voice Detection on Time and Spectral Domain

In this section, we introduce the voice detection algorithm operating on time and spectral domain. We also propose the post-processing algorithm.

A. Voice detection on time and spectral domain

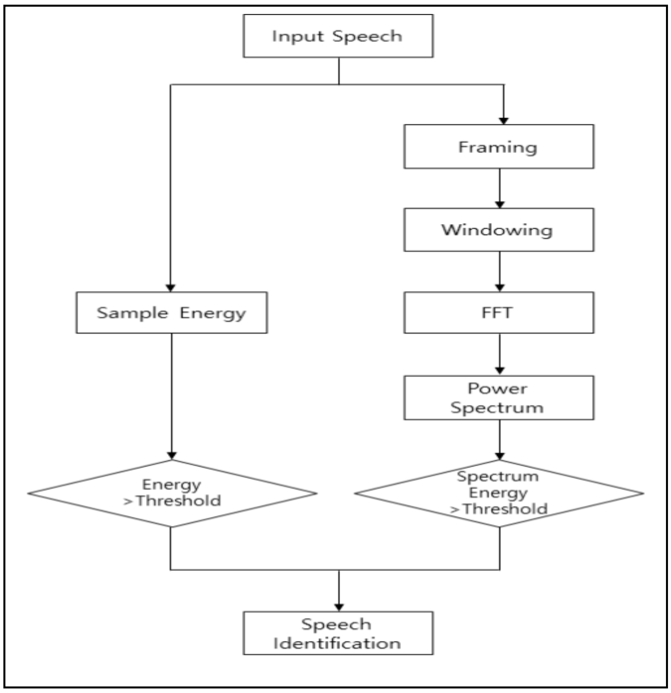

Figure 1. Voice detection procedure on time and spectral domain.

Left side of figure 1 represents the procedure of voice detection using the characteristics of the time domain signal energy. This method transforms the input speech signal into each frame to sample energy (1). Right side of this figure indicates the voice detection procedure using the power spectral energy in the spectral domain. After dividing the input speech signal as a fixed length, spectral energy shown in (3) is estimated for each frame and applied to VAD. And compared with a threshold value, the frame is determined to be either a speech region or a non-speech region, as shown in (2) and (4). If energy value of a given frame is greater than the threshold, the frame is determined as a speech region. Otherwise, it is determined as a non-speech region.

$$E_i = \sum_{m=0}^{N-1} \left(X_i[m] \right)^2 \tag{1}$$

$$\begin{aligned} &\mathrm{E_i} > \text{Threshold} \rightarrow \text{speech} \\ &\mathrm{E_i} < \text{Threshold} \rightarrow \text{non-speech} \end{aligned} \tag{2}$$

E_i is the sample energy of the ith frame in time domain. $x_i[m]$ means the mth sample data in a given frame. The threshold means the criterion to classify the speech and non-speech region. It is empirically determined as an optimal value.

$$\mathrm{P_i} = \sqrt{\sum_{m=0}^{N-1} \left(\mathrm{Re_i[m]}^2 + \mathrm{Im_i[m]}^2 \right)} \tag{3}$$

$$\begin{aligned} &\mathrm{P_i} > \text{Threshold} \rightarrow \text{speech} \\ &\mathrm{P_i} < \text{Threshold} \rightarrow \text{non-speech} \end{aligned} \tag{4}$$

P_i is the power spectral energy of the ith frame in spectral domain. $re_i[m]$ and $im_i[m]$ mean the real and imaginary numbers of mth sample data applied to the FFT process.

B. Post-processing algorithm

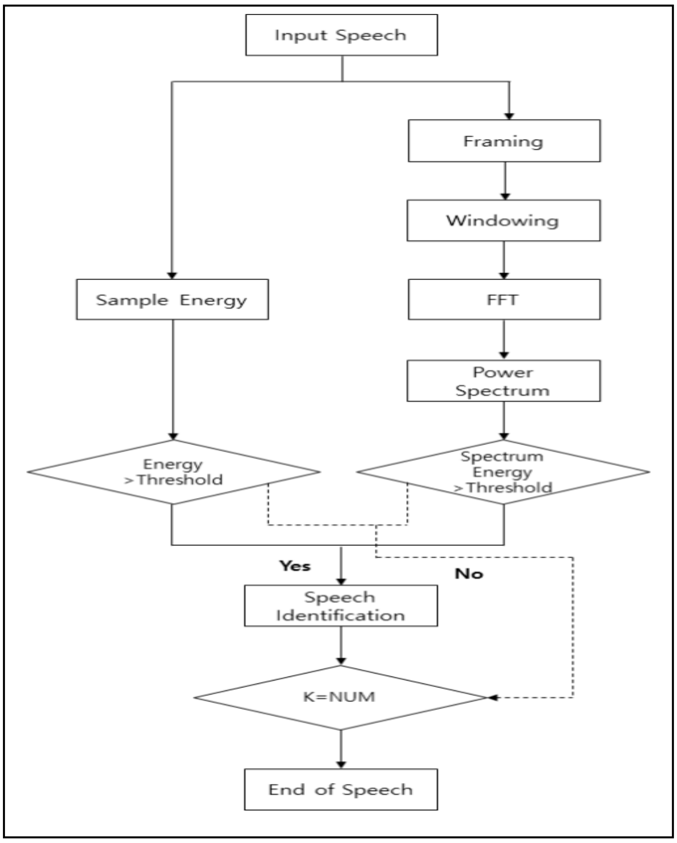

Figure 2. Post-processing procedure.

Figure 2 represents the procedure of the proposed post-processing. This process is similar to the voice detection procedure. This method also employs the signal energy. Proposed method determines the end-point of the detected speech region, by comparing the energy value of each domain with threshold. If the number of consecutive frames indicating lower values than threshold is equal to the predetermined value (NUM), the frame is determined as an end point of the speech region, as shown in (5).

$$\begin{aligned} &\mathrm{E_i} \text{ or } \mathrm{P_i} < \text{Threshold},\ \mathrm{k = NUM} \\ &\rightarrow \text{End of Speech} \end{aligned} \tag{5}$$

In the formula, E_i and P_i mean the sample energy and power spectral energy in the frame. K is the number of consecutive frames in which the energy values are lower than the threshold.

IV. Experimental Results

We performed real-time VAD experiments with several participants of men and women. We selected the speech utterances of short duration like “speech_start” and “mic_start”, and totally 8 words were used as test data by two times per word. A frame consisted of 320 samples and the sampling rate was 16 kHz. To form the spectral domain, the 512 points FFT process was set, applying the Hamming window.

We firstly investigated the voice detection on time and spectral domain. The performance was evaluated via two kinds of error rates: FAR (shortly denoted as ‘Fa’ in this section) and FRR (denoted as ‘Fr’). The threshold used in experiments is statistically estimated as values for inverse proportional graph. In order to fairly assess the performance, we conducted experiments by using same speech data in each trial.

TABLE I. Results of Voice Detection on Spectral Doamin

Threshold	Person1		Person2		Person3		Person4		Person5	
	Fr	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*
33000	3	48	2	39	2	59	1	42	2	73
83019	6	17	6	11	6	20	5	15	6	23
133240	6	16	6	12	6	18	5	12	6	21
229461	8	13	8	9	8	15	8	11	8	13

TABLE II. Results of Voice Detection on Time Domain

Threshold	Person1		Person2		Person3		Person4		Person5	
	Fr	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*
48000	2	55	1	48	1	97	1	39	3	57
152884	4	22	3	22	6	30	2	20	4	33
474918	6	20	4	14	6	25	4	13	6	28
616419	6	20	6	13	6	21	6	10	6	25

TABLE III. Result Comparison on Time and Spectral Domain

Threshold	Spectral		Threshold	Time	
	FR	*FA*		*FR*	*FA*
33000	10	261	48000	8	296
83019	29	86	152884	19	127
133240	29	79	474918	26	100
229461	40	61	616419	30	89

Table I and Table II explain the voice detection results on time and spectral domain. The spectral domain results indicated better performance compared to the time domain results in terms of FAR and FRR.

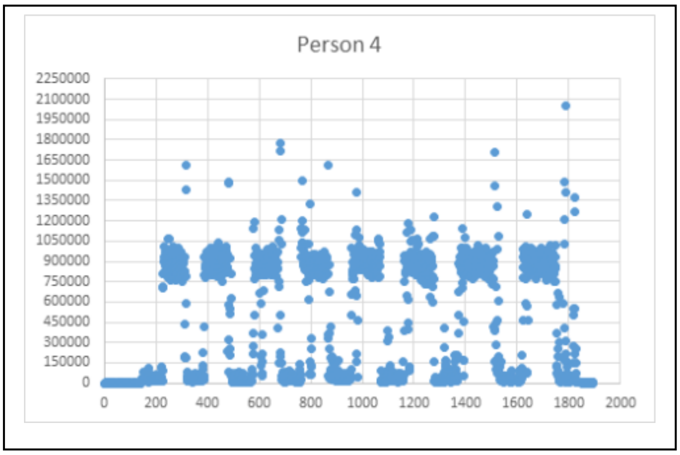

Figure 3. Time domain energy distribution of a participant.

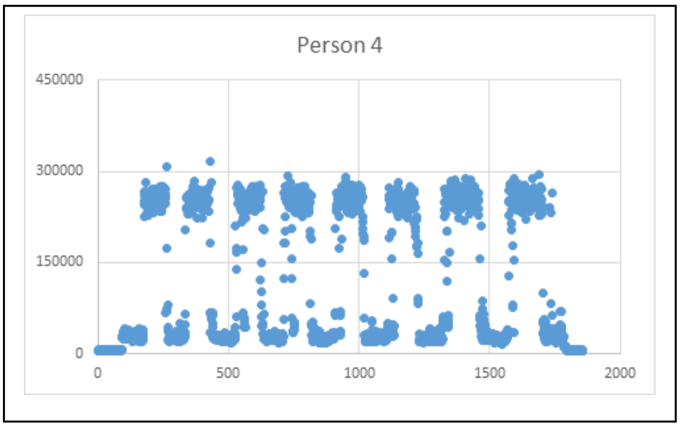

Figure 4. Spectral domain energy distribution of a participant.

Figure 3 and Figure 4 indicate distribution of time domain energy and spectral domain energy of a set of speech utterances recorded while a participant repeated to speak and stop speaking. As shown in this figure, the spectral energy indicated more intensive energy property compared to time domain energy. This property is highly related to the VAD performance in our experiment.

TABLE IV. Results of Post-processing (Power Spectrum Energy)

Threshold	Person1		Person2		Person3		Person4		Person5	
	Fr	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*	*Fr*	*Fa*
33000	1	4	0	10	0	11	3	5	0	15
83019	0	7	0	5	0	2	0	3	0	10
133240	0	6	0	4	0	2	0	2	0	7
229461	0	3	0	3	0	2	0	1	0	3

TABLE V. RESULTS OF POST-PROCESSING (TIME DOMAIN ENERGY)

Threshold	**Person1**		**Person2**		**Person3**		**Person4**		**Person5**	
	Fr	***Fa***	***Fr***	***Fa***	***Fr***	***Fa***	***Fr***	***Fa***	***Fr***	***Fa***
48000	2	1	0	11	0	7	2	1	0	13
152884	0	8	0	10	0	6	0	3	0	10
474918	0	6	0	7	0	3	0	1	0	8
616419	0	6	0	6	0	2	0	1	0	8

Table IV and Table V shows the results of the voice detection employing the proposed post-processing approach in spectral domain and time domain, respectively. The number of frames that we used for the end point detection is 15. It means that duration of 0.3 sec. was considered as a criterion for the end point determination in a speech region. As shown in the tables, our proposed post-processing approach significantly improved the performance of real-time voice detection.

V. CONCLUSION

In this paper, we introduced the voice detection method on time and spectral domain. The proposed method utilizes spectral energy as a criterion for determination of speech and non-speech regions. And time interval based post-processing is employed for more sophisticated end point detection.

In experimental results, spectral domain showed better VAD performance than time domain. In addition, the post-processing method improved the end point detection performance, significantly reducing two kinds of error rates (FAR, FRR). In future works, we will apply the proposed VAD for voice interface applications including speech recognition.

ACKNOWLEDGMENT

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-R2718-16-0035) supervised by the IITP(Institute for Information & communications Technology Promotion).

REFERENCES

[1] Yohan M., Ki-Joon K, and Dong-Hee S, “Voices of the Internet of Things: An exploration of multiple voice effects in smart homes,” Distributed, Ambient and Pervasive Interactions, vol. 9749, pp. 270-278, Jun. 2016.

[2] Yiming, S. and Rui, W., “Voice Activity Detection Based on the Improved Dual-Threshold Method”, In Intelligent Transportation, Big Data and Smart City (ICITBS), 2015 International Conference on. IEEE, pp. 996-999, Dec. 2015.

[3] Rabiner, L. R. and Sambur, M. R., “An algorithm for determining the endpoints of isolated utterances”, Bell System Technical Journal, Vol. 54, No. 2, pp. 297-315, 1975.

[4] Jae-Seung C., “Detection of Non-silent Sections using Threshold Value for Improvement of Speech Recognition Performance”, Proceedings of KIIT Summer Conference, pp. 300-302, Jun. 2015..

[5] Jungpyo H., Sangjun P., Sangbae J. and Minsoo H., “Robust Feature Extraction for Voice Activity Detection in Nonstationary Noisy Environments”, Journal of The Korean Society of Speech Sciences, Vol. 5, No. 1, pp. 11-16, Mar. 2013.

[6] Jae-Seung C., “A Detection method of Speech/Non-Speech Sections using scale of Cepstrum distance”, Proceedings of KIIT Summer Conference, pp. 489-492, Nov. 2012.

[7] Yang, X., Tan, B., Ding, J., Zhang, J. and Gong, J, “Comparative study on voice activity detection algorithm”, In Electrical and Control Engineering (ICECE), 2010 International Conference on. IEEE, pp. 599-602, Jun. 2010.