프리엠퍼시스 FIR 필터링의 음성 검출 및 음소 분할에의 응용

이창영*

Application of Preemphasis FIR Filtering To Speech Detection and Phoneme Segmentation

Chang-Young Lee*

요 약

이 논문에서 우리는 음성 검출 및 음소 분할에 대한 새로운 방법을 제안한다. 배경 잡음으로부터 신호를 구분하기 위해 에너지를 활용하게 되는데, 그 이전에 프리엠퍼시스 FIR 필터링을 적용하는 효과에 대해 조사한다. 이 방법에 의해, 에너지 프로필에서 진폭과 주파수의 곱이 동시에 작은 부분이 두드러지게 나타나게 된다. 이 처방에 의해, 묵음 / 음성 경계가 종전의 방법에 비해 더 선명해짐을 실험적으로 확인하였다. 또한 이 방법을 적용함으로써, 음소 분할 또한 더 수월해짐을 밝혔다.

ABSTRACT

In this paper, we propose a new method of speech detection and phoneme segmentation. We investigate the effect of applying preemphasis FIR filtering on the speech signal before the usual speech detection that utilizes the energy profile for discriminating signals from background noise. By this procedure, only the speech section of low energy and frequency becomes distinct in energy profile. It is verified experimentally that the silence / speech boundary becomes sharper by applying the filtering compared to the conventional method. By applications of this procedure, phoneme segmentation is also found to be much facilitated.

키워드

Speech Detection, Speech Processing, Preemphasis Filtering, FIR Filtering, Phoneme Segmentation
음성 검출, 음성 처리, 프리엠퍼시스 필터링, FIR 필터링, 음소 분할

I. Introduction

As a method of communication between man and machine, speech recognition provides a very effective interface. Speech input to a machine is about twice as fast as information entry by a skilled typist [1]. The technique of speech recognition is now familiar that lots of applications are taking use of the state of the art recognition technology [2-5].

As a part of the front-end processing of the speech signals, speech detection or isolation of the silence portion from the speech section is invaluable with regard to cost reduction in memory storage

* 동서대학교 산업경영공학과(seewhy@dongseo.ac.kr)
접수일자 : 2013. 04. 02 심사(수정)일자 : 2013. 04. 25 게재확정일자 : 2013. 05. 20

and processing time. In case of TASI (Time Assignment Speech Interpolation), for example, the system should detect the idle time of talkers effectively to accommodate more users with less communication channels [6]. In speech recognition, the accurate detection of speech cannot be overestimated. An experiment reported that the recognition error rate has increased by 3% when the speech boundaries were displaced by 60 ms from the ideal ones [7].

There are two categories of factors that hinder the accurate speech detection: one is the speech production style of speaker and the other is environmental conditions in which the speech is produced. As examples of the former case, the talker often produces sound artifacts including lip smacks, heavy breathing, mouth clicks, pops and so on [8].

In speech detection employing the signal energy, the magnitude of such noises is often of the order that cannot be ignored. The noise comes inevitably from speech production environment, too. Though there have been lots of studies that tackle the background noises [9], cooperative and noise-free environment is preferred during speech production. Such demands are too stringent to be realized in practice. Besides, even if the environment is clean, there come various distortions in microphone, recorder, telephone line, and transmission lines.

In this paper, we study a method of speech detection that employs finite impulse response (FIR) filter. This method is especially effective in removing the noise of low-frequency regime. Generally, the onset of speech production is found to become sharper by the processing proposed in this paper, and thus the speech detection becomes easier and more accurate.

The organization of this paper is as follows. Section II describes the limits and drawbacks of the previous speech detection method based on the energy for the raw signal. After providing the theoretical background and experimental results of the applications of preemphasis FIR filtering to speech detection in Section III, concluding remarks will be given in section IV.

II. Speech Detection by Energy

Figure 1 shows a partial waveform of a speech pronounced by a male speaker. The speech signal was sampled by 16 kHz with 16 bits of quantization. Along with the waveform, shown is a profile of log energy defined by

$$E(j) = \log\left[\sum_{i=j-N/2}^{j+N/2-1} x^2(i)\right]. \quad (1)$$

The frame length is $N=512$ corresponding to 32 ms of time duration. Before estimating Eq. (1), DC bias should be properly removed in order for the energy to be effective for speech detection.

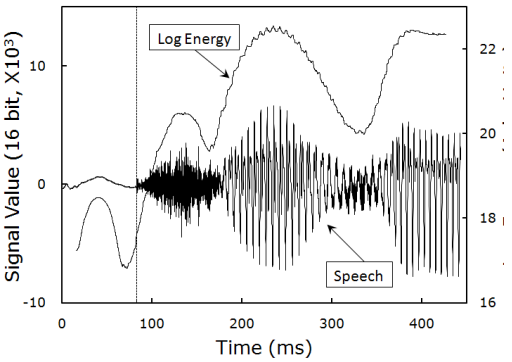

Fig. 1 A partial waveform(left scale) of a speech pronounced by a male speaker and its log energy profile (right scale) defined by Eq.(1).

We note the wiggly-shaped noise pattern before the onset of the speech production marked by a dotted vertical line on the graph. This background noise has so low frequency (~10Hz) that it is not heard by human ear. However, as can be seen from the graph, its energy is larger than that of

the beginning section (phoneme /ch/) of the first speech production. Therefore, the strategy of speech detection by comparing the energy profile with a specified threshold value should fail.

The case of Figure 1 is just an example among many instances that hinder accurate location of the silence-speech boundary. There have been lots of investigations to remove and tackle various sources of noises [10]. Our study in this paper is to help the speech detection method based on examining the energy, which is prone to fail in itself.

III. Speech Detection by FIR Filtering

An intuitive method of eliminating the background noise is to subtract the noise itself from the signal. This naive treatment, however, removes the signal, too. To circumvent this problem, we shift the signal by an amount, multiply the delayed signal by a factor, and subtract the resultant signal from the original one. This prescription can be expressed by a finite impulse response (FIR) filter of the form

$$y(i) = x(i) - \mu x(i-s) \tag{2}$$

where μ and s are adjustable parameters. This equation constitutes the backbone of the theory in this paper.

The prescription of Eq. (2) might be interpreted as a special case of a more general FIR filtering of order M:

$$y(i) = \sum_{j=0}^{M} a_j x(i-j)$$

Figure 2 shows the corresponding network.

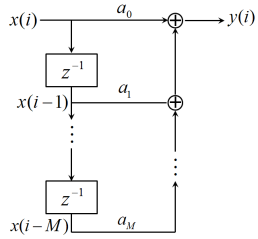

Fig. 2 Network of finite impulse response filtering of order M.

This sort of filtering is utilized frequently in signal processing. For example, the FIR filtering of order 1 is applied for spectral flattening [11] before the extraction of feature vectors such as linear predictive coding (LPC) and mel-frequency cepstral coefficients (MFCC) [12]. In this sense, our prescription of Eq. (2) for speech detection might be phrased in terms of interchange of speech detection and preemphasis FIR filtering, as is shown by Figure 3.

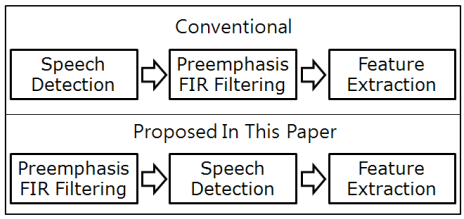

Fig. 3 Comparison of conventional and proposed methods for speech detection, preemphasis FIR filtering, and feature extraction.

Eq. (2) might be interpreted from another viewpoint. For a sinusoidal wave of amplitude A and angular frequency ω

$$x(t) = A\sin(\omega t) \tag{3}$$

if we delay (shift in time) it by δt and subtract the result from (3), we obtain

$$y(t) = A\left[\sin(\omega t) - \sin(\omega(t-\delta t))\right] \tag{4}$$

which can be approximated by

$$y(t) \approx (A\omega\delta t)\cos(\omega t) \quad (5)$$

for $\omega\delta t \ll 1$.

For the sinusoidal wave of Eq. (3), the magnitude is determined solely by the amplitude A. Compared to this, the magnitude of Eq. (5) (transformed signal) is proportional to amplitude times frequency. Consequently, the resultant signal has insignificant portion when both the amplitude and the frequency have small values at the same time. This means in turn that the background noise with low frequency might be well separated from the low energy speech signal of relatively high frequency which is characteristic of most consonants [13].

Though the two parameters of μ and s in Eq. (2) are to be adjusted for optimal speech detection, the efficacy of the prescription is clear even by a simple choice of $s=1$ and $\mu=0.95$, which is usually adopted in preemphasis filtering for spectral flattening [14].

Figure 4 shows the result of transformation of Figure 1 by Eq. (2) with this specification of parameters. Along with the transformed waveform, is shown the energy profile. We see that the wiggly-shaped low frequency noise of Figure 1 has been removed in Figure 5. It might be said that the problem of speech detection is remarkably facilitated.

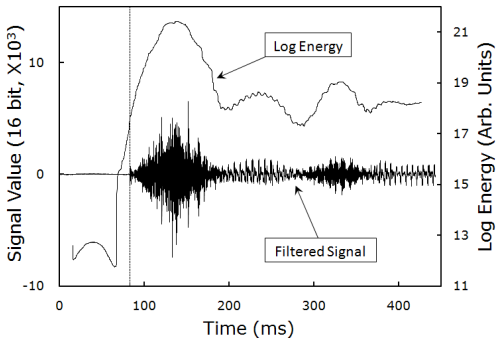

Fig. 4 Waveform (left scale) transformed from figure 1 and the profile of log energy(right scale).

The proposed method could also be utilized in speech segmentation [15]. Figure 5 shows a partial waveform of a sentence pronounced by a female speaker. Along with the signal waveform, log energy profile is displayed. We see that there are depressed valleys at the two locations of syllable boundaries. Though the syllables might be segmented roughly, it would be desirable if the energy profile becomes sharper at the locations of phoneme changes.

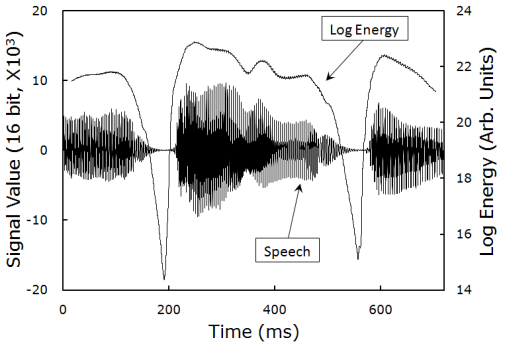

Fig. 5 A partial waveform(left scale) of a sentence pronounced by a female speaker and the profile of log energy(right scale).

For this signal, we apply the preemphasis FIR filtering. To enhance the efficacy of the prescription of the filtering, we try applying the prescription several times successively. Figure 6 shows the network of two successive applications of preemphasis FIR filtering. It is actually equivalent to second order FIR filtering.

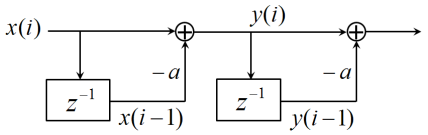

Fig. 6 Network of two successive applications of preemphasis FIR filtering.

Figure 7 shows the waveform that is transformed five times from Figure 5 according to

Eq. (2). Simple choice of the parameters, $\mu = 0.95$ and $s = 1$ was adopted, too. We notice that flat valleys appeared at the syllable boundaries. By comparing the energy profile of this graph with that of Figure 5, it might be said that the phoneme segmentation has become much facilitated.

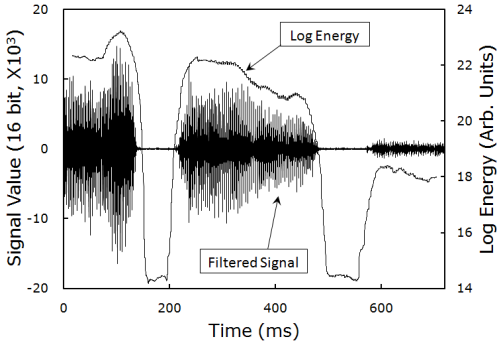

Fig. 7 Waveform(left scale) transformed from Figure 5 and its log energy profile(right scale).

IV. Conclusion

Conventional speech detection usually relies on examining energy profile of the waveform. The silence-speech boundary is determined whether the energy of a short-length frame exceeds a specified threshold or not. This method is prone to fail when there exists background noise of various kinds. It is not easy to specify the threshold to provide good criterion for separating the noise from the signal. In many cases, the energy of the noise is comparable to or even larger than that of the speech signal. If this happens, the speech detection by energy consideration should fail.

In this paper, we investigated the improvement of the speech detection method by applying preemphasis FIR filtering, sometimes successively. In this approach, only the portion of small value of amplitude times frequency is classified as non-speech section. Consequently, the background noise of small frequency is distinguished from the low-energy signal of high frequency.

This prescription of preemphasis FIR filtering could be applied for speech segmentation. The efficacy of the proposed method was found to be clear by comparing the energy profiles of the raw and transformed signals. While the energy profile for the raw signal changes smoothly near the region of phoneme transition, the energy profile of the transformed signal showed sharp change.

In conclusion, by the method proposed in this paper, the speech detection and phoneme segmentation could be facilitated.

References

[1] G. Kaplan, "Words Into Action I," IEEE Spectrum, Vol. 17, pp. 22-26, 1980.

[2] Myoung-ku Kang, “A Study on the Design of Multimedia Service Platform on Wireless Intelligent Technology,” The Journal of the Korea Institute of Electronic Communication Sciences, Vol. 4, No. 1, pp. 24-30, 2009.

[3] Jae-duck Yoo, Hong-tae Park, Hyun-sik Shin, & Yun-ho Shin, “A Study of the Communication Infrastructure Construction for u-City in Korea,” The Journal of the Korea Institute of Electronic Communication Sciences, Vol. 1, No. 2, pp. 127-135, 2006.

[4] Y. Chang, S. Hung, N. Wang, & B. Lin, "CSR: A Cloud-Assisted Speech Recognition Service for Personal Mobile Device," International Conference on Parallel Processing (ICPP), pp. 305-314. 2011.

[5] Beom-joon Kim, “Service Quality Criteria for Voice Services over a WiBro Network,” The Journal of the Korea Institute of Electronic Communication Sciences, Vol. 6, No. 6, pp. 823-829, 2011.

[6] J.E. Flood & D.I. Urquhart-Pullen, "Time-assignment speech interpolation in time-compression-multiplex transmission," Proceedings of the Institution of Electrical Engineers, Vol. 111, No. 4, pp. 675-683, 1964.

[7] J.G. Wilpon, L.R. Rabiner, & T.B. Martin, "An improved word-detection algorithm for telephone-quality speech incorporating both syn-

tactic and semantic constraints," AT&T Tech. J., Vol. 63, No. 3, pp. 479-498, 1984.
[8] L.R. Rabiner & B. Juang, "Fundamentals of Speech Recognition," Prentice Hall, pp. 143-149, 1993.
[9] T. Kristjansson, B. Frey, L. Deng, & A. Acero, "Towards non-stationary model-based noise adaptation for large vocabulary speech recognition," ICASSP '01, Vol. 1, pp. 337-340, 2001.
[10] J.R. Deller, J.G. Proakis, & J.H.L. Hansen, "Discrete-Time Processing of Speech Signals," Macmillan, New York, pp. 246-251, 1994.
[11] L.R. Rabiner & B. Juang, "Fundamentals of Speech Recognition," Prentice Hall, pp. 112-117, 1993.
[12] J.-C. Wang, J.-F. Wang, & Y. Weng, "Chip design of MFCC extraction for speech recognition," The VLSI Journal, Vol. 32, pp. 111-131, 2002.
[13] L.R. Rabiner & B. Juang, "Fundamentals of Speech Recognition," Prentice Hall, pp. 30-37, 1993.
[14] S. Kajita, K. Takeda, & F. Itakura, "Spectral weighting of SBCOR for noise robust speech recognition," ICASSP '98, Vol. 2, pp. 621-624, 1998.
[15] D.C. Costa, G.A.M. Lopes, C.A.B. Mello, & H.O. Viana, "Speech and phoneme segmentation under noisy environment through spectrogram image analysis," IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1017-1022, 2012.

저자 소개

이창영(Chang-Young Lee)

1982년 2월 서울대학교 물리교육학과 졸업(이학사)
1984년 2월 한국과학기술원 물리학과 졸업(이학석사)
1992년 8월 뉴욕주립대학교 (버펄로) 물리학과 졸업(이학박사)
1993년~현재 동서대학교 시스템경영공학과 교수
※ 관심분야 : 패턴인식, 신호처리