International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.9 (2014), pp.61-68
http://dx.doi.org/10.14257/ijmue.2014.9.9.07

Robust Speech Detection using SEM and SFN

In-Sung Han[1] and Chan-Shik Ahn[2]

[1] Dept. of The 2nd R&D Institute, Agency for Defense Development
460,Songpa-gu, Seoul, 138-600, South Korea
ishan78@add.re.kr
[2] Dept. of Computer Engineering, Kwangwoon University
20, Gwangun-ro, Nowon-gu, Seoul, South Korea
coolsahn@gmail.com

Abstract

Speech recognition, the problem of performance degradation is the difference between the model training and recognition environments. Silence features normalized using the method as a way to reduce the inconsistency of such an environment. Silence features normalized way of existing in the low signal-to-noise ratio. Increase the energy level of the silence interval for speech and non-speech classification accuracy due to the falling. There is a problem in the recognition performance is degraded. This paper proposed a robust speech detection method in noisy environments using a SFN (silence feature normalization) and SEM (speech energy maximize). In the high signal-to-noise ratio for the proposed method was used to maximize the characteristics receive less characterized the effects of noise by the speech energy. Cepstral feature distribution of speech and non-speech characteristics in the low signal-to-noise ratio and improves the recognition performance. Result of the recognition experiment, recognition performance improved compared to the conventional method.

Keywords: *Speech Recognition, Voice Detection, Noise Reduction, Speech Energy Maximization, Silence Feature Normalization*

1. Introduction

Voice detector is being applied to voice handling system such as speech recognition or noise reduction algorism in combination with various applications in mobile voice communication environment, and also is being developed as a core part that the major effect on systematic function. In noise reduction algorism that chases and gets rid of the noise signal through VAD(voice activity detection) algorism, voice detection function serves as an important element that generally affect noise reduction algorism [1].

In case of feature parameters used for voice detecting, in an environment that SNR is high, as the characteristics of feature parameter on speech and non-speech are relatively distinctive, their recognition function doesn't remarkably go down [2]. However, in an actual noise environment that there are various kinds of noise or in one that SNR is low, as feature parameter is sensitive against noise signal, its voice detecting function decline may arise as a problem.

Therefore, this paper suggests voice detecting method that has advantages in noise environment by using SEM (speech energy maximize) and SFN (silence feature normalization). Suggestion made use of characteristic of the silence feature's being less affected by maximizing speech energy in high SNR environment. On the other hand, in low SNR, recognition function got improved using the characteristics of cepstrum feature

ISSN: 1975-0080 IJMUE

distribution of speech and non-speech. Function improvement compared to existing methods was verified through recognition experiments.

This paper consists of 5 parts in total. Related studies will be mentioned in the second chapter and voice detecting method that has advantages against noise environment using speech energy maximization and silence feature normalization will be explained in detail in the third chapter. Systematic assessment is conducted in chapter 4 and this paper is concluded with chapter 5.

2. Related Work

2.1. Spectral Energy

In low SNR energy spectrum, speech band shows relatively high energy spectrum compared to non-speech band. Therefore, it can be assumed that speech energy spectrum has comparatively higher energy spectrum than non-speech energy spectrum. This kind of energy spectrum is similarly described with information entropy theory introduced by Shannon [3].

Shannon's entropy can be explained with below formula [4].

$$H(S) = -\sum_{i=1}^{N} P(s(i)) \bullet \log_2(P(s(i))) \tag{1}$$

N shows the overall symbol , ***s(i)*** shows ***i***'s simbol, and ***P(i)*** show posterior probability of ***i*** . Entropy can be explained with below formula in spectral energy band [5].

$$H(|Y(k,l)|^2) = -\sum_{k=1}^{N/2} \left\{ P(|Y(k,l)|^2) \bullet \log_2(P(|Y(k,l)|^2)) \right\} \tag{2}$$

To calculate entropy, dispersion spectral power is to be calculated using DTF(Discrete Fourier Transform). Spectral energy probability can be expressed in below formula

$$P(|Y(k,l)|^2) = \frac{|Y(k,l)|^2}{\sum_{k=1}^{N/2} |Y(k,l)|^2} \tag{3}$$

k shows Frequency bin Index, ***l*** shows frame index. Spectral energy probability against Frequency bin can be deduced from Frame ***l***. Deduced each frequency bin probability is calculated to entropy by formula [6].

2.2. Critical Band

In the side of frequency, audible range has the band between minimum audible limit and maximum audible limit and the band on sound intensity has the same with the audible range. The absolute minimum audible limit was defined as the minimum audible band that human can sense the sound in noise-free environment by ***H***. Fletcher and can be shown in the following formula. $\boldsymbol{T_q(f)}$ shows minimum audible band in the frequency band.[7].

$$T_q = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000-3.3}\right)} + 10^{-3}\left(\frac{f}{1000}\right)^4 \tag{4}$$

Masking effect is the phenomenon of weaker sound's being blocked by stronger sound and stronger sound's called masker, and the weaker sound is called maskee (the weaker sound). Maskee can be hear when it's over masking threshold calculated by the masker (the stronger sound) [8]

Frequency masking is that the maskee is being masked when the masker and the maskee simultaneously occur and the masking degree can be calculated through frequency analysis. The threshold of the sound being masked differ from frequency bandwidth, and frequency bandwidth with equivalent masking threshold is called 'critical band'. The width of critical band is constant as 100Hz when the frequency is below 1kHz, and when the frequency is over 1kHz, it proportionally increases according to the frequency [9].

The critical band width can be shown in the following formula.

$$BW_c(f) = 25 + 75(1 + 1.4(f/1000)^2)^{0.69} \tag{5}$$

The width of the critical band is called "Bark", the formula to convert frequency unit ***z(f)*** to bark unit is as below.

$$z(f) = 13\arctan(0.0076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \tag{6}$$

Bark Scale frequency can remarkably express the band with important elements of sounds [10].

3. Voice Detection that Excel in Noise Environment

3.1. Voice Energy Maximization

Voice has pitch frequency for vowels, and the pitch frequency is called 'Basic frequency'. Basic frequency has the characteristics of showing the biggest energy throughout the full-band of voice band, so it's expressed as the maximum energy and minimum energy is shown as noise signal unrelated to voice signal. Noise signal is marked as minimum energy in the given frame. By using the characteristics of voice energy, maximum energy can be described in the following formula in the given frame for maximizing voice energy

$$E_{\max}(l) = \min\{B(b_i, l), \ldots, B(b_{i-1}, l)\} \tag{7}$$

l shows relevant frame index, ***b_i*** and shows bark scale index. Voice energy has relatively bigger energy compared to noise and noise has smaller amount of energy, so calculation of the ratio of voice energy against noise energy and energy ratio on the output are shown in below formula.

$$PSR(b_i, l) = \frac{B(b_i, l) - E_{\max}(l)}{\mu(l)} \tag{8}$$

Express the ratio of voice energy against noise energy. In the Figure 1, Speech Energy Maximization at SNR10dB, (a) Input speech at SNR 10dB, (b) Frame energy for Speech Energy Maximization, (c) Result of Speech Energy Maximization.

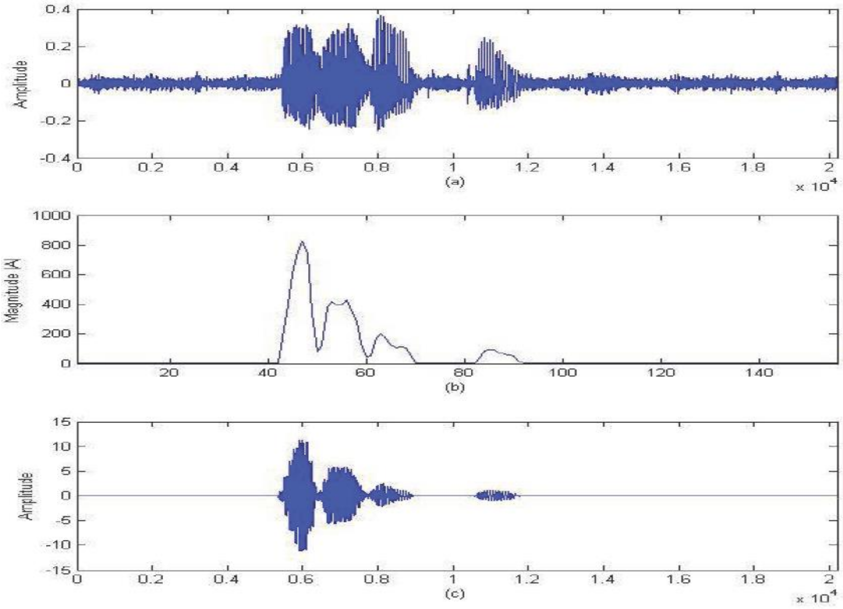

Figure 1. Speech Energy Maximization at SNR10dB, (a) Input speech at SNR 10dB, (b) Frame energy for Speech Energy Maximization, (c) Result of Speech Energy Maximization

3.2. Silence Feature Normalization

Silence feature normalization is the method of finding the silence interval with weak log energy and then cutting the value down under a certain value.

$$\log\overline{E}(n) = \frac{1}{2}\left(\log\overline{E}(n+1) - \log\overline{E}(n-1)\right) \tag{9}$$

log ***E(n)*** shows the log energy of ***n*** th frame and **log** ***~~E~~(n)*** shows the output value of filtering. The standard value ***T_0*** of classification on voice and silence is as below.

$$T_0 = \frac{1}{N}\sum_{n=1}^{N} \log\overline{E}(n) \tag{10}$$

N shows the number of frames for each voice data. Normalized log energy is shown in the below formula.

$$\log\hat{E}(n) = \begin{cases} \log E(n) & \text{if } \log\overline{E}(n) > T_0 \\ \log(\epsilon) + \delta & \text{if } \log\overline{E}(n) \le T_0 \end{cases} \tag{11}$$

If calculated value **log** ***~~E~~(n)*** in formula(9) is over the standard value, it's classified as voice and its original log energy value remains still, if it's below the standard value, it's classified as silence and normalized into tiny value ε . ε means the constant number 10^{-3}, δ shows tiny value with average value 0 and dispersion value 10^{-8}.

Voice interval contaminated by noise has broader frequency bandwidth compared to the band only with voice and the interval with big log energy is less affected by noise compared to the interval with small log energy, so used weighting function. The weighting function ω (***n***) can be shown like below formula.

By multiplying weighting function to output formula (9) increase the weighting to make voice interval have bigger value and decrease the weighting to make silence interval have smaller value. It is as below.

$$w(n)=\begin{cases}1/(1+\exp(-(\log\overline{E}(n)-T_0)/\beta\sigma_1)) \\ \qquad\qquad\qquad \text{if } \log\overline{E}(n)>T_0 \\ 1/(1+\exp(-(\log\overline{E}(n)-T_0)/\beta\sigma_2)) \\ \qquad\qquad\qquad \text{if } \log\overline{E}(n)\le T_0\end{cases}$$

(12)

$$\log\widetilde{E}(n)=w(n)\ \bullet\ \log\hat{E}(n)$$

(13)

Output result by formula can be seen with Silence feature normalization as in Figure 2.

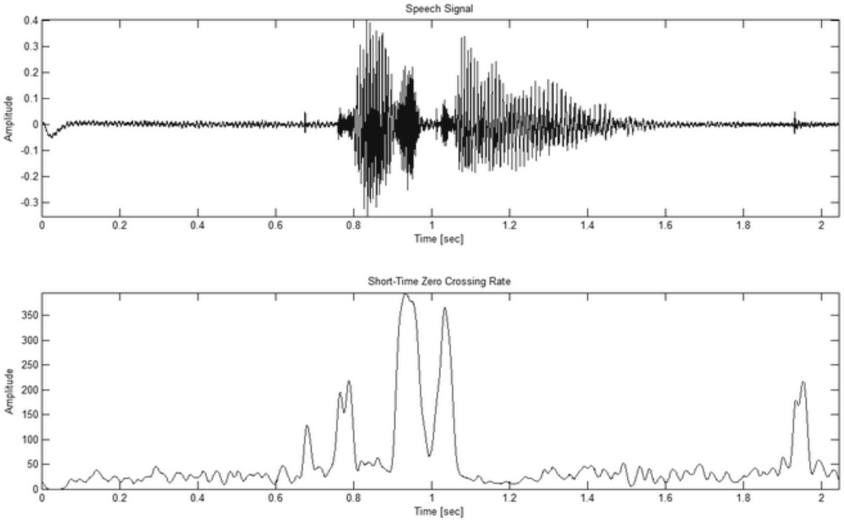

Figure 2. Output Result of Silence Feature Normalization

4. Experiment Result

This paper conducted the experiment on voice detection method that excel in noise environment by using voice energy maximization and silence feature normalization that this paper suggested

For assessment, Aurora 2.0 database was used. Aurora 2.0 consists of noise environment and each-noise level (including white gaussian noise, babble noise) and sorts each noise environment (street, airport, car noise etc.) so it's used to verify voice improvement algorism [11].

The experiment was conducted in each noise environment with 15dB, 10dB, 5dB and 0dB separately to verify its performance on SNR change, also to evaluate endpoint detection performance PHR(Pause Hit Ratio) and FAR (False Alarm Ratio) in sound area were used.

As sound source, 8kHz sampling rate, 16bits were used and FFT size is 256 samples and 1/2 overlapping interval was used. Hamming Window was also used [12].

As the Figure 3 (b) shows, the increased voice energy part in SNR 15dB is seen as the experiment result

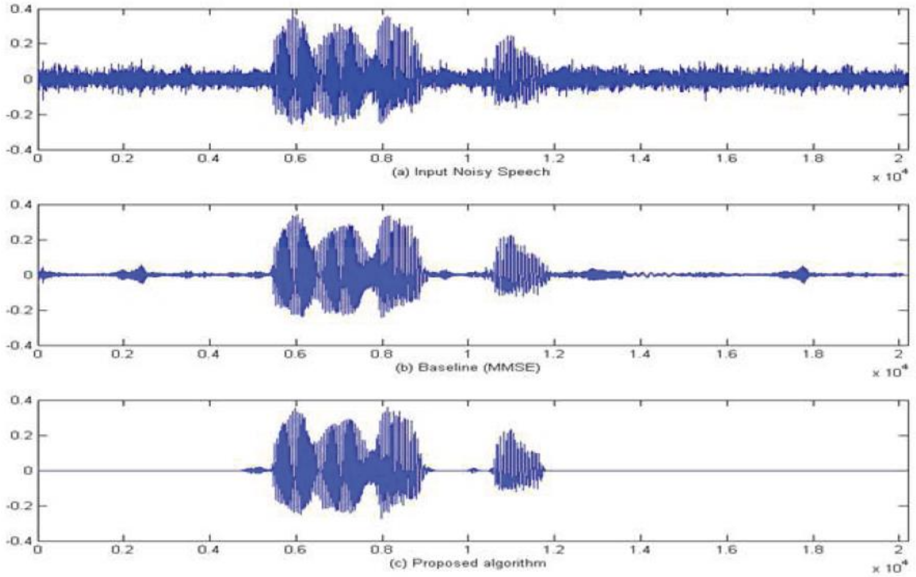

Figure 3. (a) Input Signal of SNR 15dB (b) Results of Speech Energy Maximization and Silence Feature Normalization (c) Results of Speech Detection

Table 1. PHR and FAR for the SNR

Noise	SNR	VAD Result (%)	
Noise	SNR	PHR	FAR
Car	0	98	2
Car	5	98	2
Car	10	100	0
Car	15	100	0
Airport	0	86	14
Airport	5	89	11
Airport	10	93	7
Airport	15	93	7
Street	0	89	11
Street	5	89	11
Street	10	95	5
Street	15	95	5

Regarding noise interval, the separation of speech interval and non-speech interval is seen by applying 0 value for its energy. Table 1 is the evaluation result on voice detection performance in each noise environment. PHR used as performance scale shows 98% accuracy in low SNR of noise environment and showed 100% accuracy in high SNR. FAR(False Alarm Ratio) in sound area was 0% by showing outstanding performance in SNR of 15dB and 10dB and 2% performance were marked for each in low SNR interval of 5dB and 0dB

5. Conclusion

This paper evaluated voice detection and recognition performance that excel in noise environment by using speech energy maximization and silence feature normalization. As feature parameter is sensitive against noise signal in actual noise environment with various kinds of environmental noise or for low SNR voice, so voice detection performance deteriorates. Therefore, voice detection method that excels in noise environment using voice energy maximization and silence feature normalization was suggested.

Suggested method made use of characteristics of silence feature's being less affected by noise by maximizing voice energy in high SNR and voice and non-voice cepstrum feature dispersion in low SNR.

In result, recognition performance has accuracy 99% in char noise environment. FAR performance is good (0%) in SNR 15dB and 10dB. However, in low level 5dB and 0dB it just improved 2% of its performance. Improved recognition performance compared to existing method was verified through experiment result.

References

[1] G. Shen and H.-Y. Chung, "Cepstral Distance and Log-Energy Based Silence Feature Normalization for Robust Speech Recognition", The Journal of the Acoustical Society of Korea, vol. 29, no. 4, (**2010),** pp. 278-285.

[2] Y.-S. Park and S. Lee, "Voice Activity Detection Using Global Speech Absence Probability Based on Teager Energy in Noisy Environments", Journal of the Institute of Electronics Engineers of Korea SP., vol. 49, no. 1, **(2012),** pp. 97-103.

[3] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection", IEEE Signal Processing Letters, vol. 6, no. 1, **(1999),** pp. 1-3.

[4] Y.-S. Park and J.-H. Chang, "A New Unified System of Acoustic Echo and Noise Suppression Incorporating a Novel Noise Power Estimation", The Journal of the Acoustical Society of Korea, vol. 28, no. 7, **(2009),** pp. 680-685.

[5] K. C. Wang and Y. H. Tsai, "Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy", Second International Symposium on Universal Communication, **(2008),** pp. 423-428.

[6] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs", in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., no. 2, **(2001),** pp. 749-752.

[7] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement", IEEE Trans. ASLP, no. 16, **(2008),** pp. 229–238.

[8] K. S. Yao, E. Visser, O. W. Kwon and T. W. Lee, "A Speech Processing Front-End with Eigenspace Normalization for Robust Speech Recognition in Noisy Automobile Environments", Proc. Eurospeech, **(2003),** pp. 9-12.

[9] C. F. Tai and J. W. Hung, "Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments", Proc. ICSLP, **(2006),** pp. 2558-2561.

[10] T. Zoltan, M. Peter, T. Zoltan and F. Tibor, "Robust voice activity detection based on the entropy of noise-suppressed spectrum", in Proc. of INTER-SPEECH, **(2005),** pp. 245-248.

[11] S. Rangachari and P. C. Loizou, "A noise-estimation algo-rithm for highly non-stationary environments", Speech Communication, vol. 48, no. 2, **(2006),** pp. 220-231.

[12] G.-K. Choi and S.-H. Kim, "Voice Activity Detection Method Using Psycho-Acoustic Model Based on Speech Energy Maximization in Noisy Environments", The Journal of the Acoustical Society of Korea, vol. 28, no. 5, **(2009),** pp. 447-453.

Authors

Insung Han, he received his B.S. degree in Computer Engineering from Paichai National University, Daejeon, Korea, in 2001, and his M.S. and Ph.D. degrees in Computer Science from Kwangwoon University, Seoul, Korea, in 2004 and 2009, respectively. From 2010 to 2012, he worked as a senior researcher at R&D Center of KICA (*Korea Information Certificate Authority, Inc*). He is now Senior Researcher in the Department of the 2nd R&D Institute Defense Cyber Warfare Technology Center, Agency for Defense Development, Seoul, Korea. His research interests include network security, ad hoc networks, wireless sensor networks and speech recognition

Chan-Shik Ahn, he received his B.S. and M.S. degrees in Computer Engineering from Kwangwoon University, Seoul, Korea, in 2000 and 2002, respectively. He is now a candidate for the Ph.D. degree in Computer Engineering from Kwangwoon University, Seoul, Korea.