Bio-Medical Materials and Engineering 24 (2014) 3295–3301
DOI 10.3233/BME-141152
IOS Press

Voice activity detection algorithm using perceptual wavelet entropy neighbor slope

Gihyoun Lee[a], Sung Dae Na[a], Jin-Ho Cho[b] and Myoung Nam Kim[c,*]
[a] *Department of Medical & Biological Engineering, Graduate School, Kyungpook National University, 680, Gukchaebosang-ro, Jung-gu, Daegu 700-842, Korea*
[b] *School of Electronics Engineering, College of IT Engineering, Kyungpook National University, 680, Gukchaebosang-ro, Jung-gu, Daegu 700-842, Korea*
[c] *Department of Biomedical Engineering, School of Medicine, Kyungpook National University, 680, Gukchaebosang-ro, Jung-gu, Daegu 700-842, Korea*

Abstract. This paper presents a voice activity detection (VAD) approach using a perceptual wavelet entropy neighbor slope (PWENS) in a low signal-to-noise (SNR) environment and with a variety of noise types. The basis for our study is to use acoustic features that have large entropy variance for each wavelet critical band. The speech signal is decomposed by the proposed perceptual wavelet packet decomposition (PWPD), and the VAD function is extracted by PWENS. Finally, VAD is decided by the proposed VAD decision rule using two memory buffers. In order to evaluate the performance of the VAD decision, many speech samples and a variety of SNR conditions were used in the experiment. The performance of the VAD decision is confirmed using objective indexes such as a graph of the VAD decision and the relative error rate.

Keywords: Voice activity detection, wavelet transform, wavelet decomposition, neighbor slope, entropy

1. Introduction

Voice activity detection (VAD) algorithms are currently being employed by a variety of speech analysis systems such as speech recognition and noise cancellers, and have key characteristics that significantly affect the performance of different systems [1]. Recently, most VAD algorithms have included the feature extraction method because it has good performance in non-stationary noise environments. The signal energy and zero crossing rate (ZCR) methods, which are among the most widely used methods, have low computing power and high speech recognition rates. However, the ZCR method performs poorly in low signal-to-noise ratio (SNR) environments [2]. Statistical feature extraction algorithms such as the likelihood ratio (LR) and Entropy have good performance in low-SNR environments. However, these algorithms require extensive computing power and have poor performance in certain noise environments [3–11]. On the other hand, Asgari [12] proposed a VAD algorithm using the entropy of the frequency domain. This approach shows good performance for mono-

[*] Corresponding author: Myoung Nam Kim, Department of Biomedical Engineering, School of Medicine, Kyungpook National University, 680, Gukchaebosang-ro, Jung-gu, Daegu 700-842, Korea. Tel.: +82-53-200-5266; Fax: +82-53-427-5539; E-mail: kimmn@knu.ac.kr.

syllabic words, but the performance is not good for speech signals consisting of entire sentences. The G729B VAD of ITU-T [13] has been widely used in commercial products and shows good performance in quiet environments, but it remains error prone in low-SNR environments.

In this paper, a new VAD algorithm that is based on wavelet decomposition and signal entropy is proposed. First, the proposed algorithm includes signal decomposition using perceptual wavelet packet decomposition (PWPD), and the VAD feature function is extracted by the proposed perceptual wavelet entropy neighbor slope (PWENS). Finally, the VAD function is determined by the proposed VAD decision rule. The performance of the proposed algorithm is confirmed by performing experiments.

2. Theory and method

2.1. Perceptual wavelet packet decomposition

The structure of the critical bands in PWPD, which was modified from wavelet packet decomposition, is close to that of the psycho-acoustic model [14]. The primary reason for applying the psycho-acoustic model is that humans are able to perceive necessary sounds without prior knowledge. The frequency of sounds is composed of all 17 critical bands in the psycho-acoustic model; thus, the critical band of PWPD is also composed of 17 critical bands. The speech signal is decomposed to 17 sub-bands of the wavelet coefficient $w_{j,m}(k)$ using PWPD. In other words, $w_{j,m}(k)$ is the jth level, kth wavelet coefficient of the mth sub-band in PWPD, where j = 3,4,5, m = 1,…,17, and k = 1,…,$N/2j$. $w_{j,m}(k)$ can be modified in the time and critical band. The modified $w_{j,m}(k)$ can also be expressed in matrix form by Eq. (1).

$$\Psi_m(t) = \begin{bmatrix} \psi_1(t) \\ \psi_2(t) \\ \vdots \\ \psi_{17}(t) \end{bmatrix} = \begin{bmatrix} \psi_1(1) & \psi_1(2) & \cdots & \psi_1(t) \\ \psi_2(1) & \psi_2(2) & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{17}(1) & \psi_{17}(2) & \cdots & \psi_{17}(t) \end{bmatrix} \tag{1}$$

where $\Psi_m(t)$ is the signal composed of the mth sub-band at specific time t.

2.2. Signal entropy

The signal entropy is a statistical analysis method which is widely used to extract the envelope of a signal. The speech signal entropy has features that reduce the noise area, while enhancing the speech area. The signal entropy is computed using the following formula [15]:

$$E(t) = -\frac{1}{N}\sum_{i=1}^{N} x(i)\log|x(i)| \tag{2}$$

where x is the signal sample normalized to the positive maximum value of the signal and N is the number of samples within 20 *ms*. The frame size is decided by considering a maximum delay limit of the communication system [16] and computing delay. Then, the normalized average of the entire signal is computed as follows [11].

$$P(t) = \frac{E(t) - mean(E_{total})}{std(E_{total})} \tag{3}$$

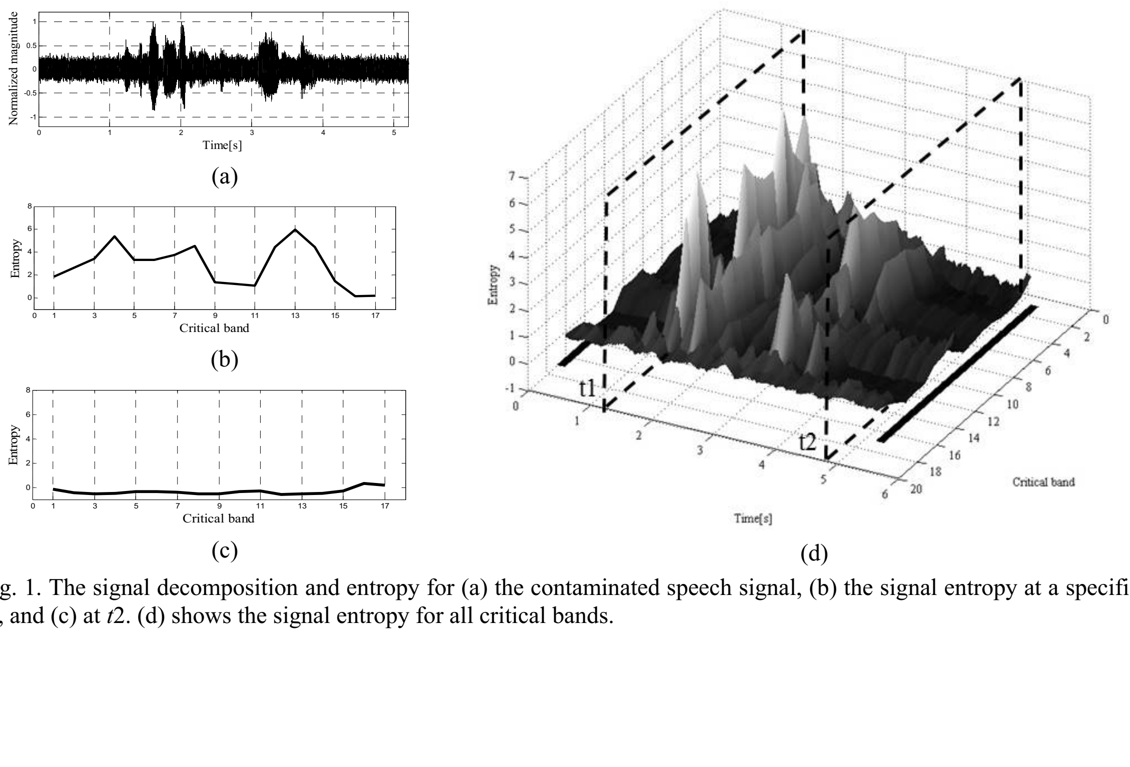
where *mean*(*E*) is the mean value of the entire *E* and *std*(*E*) is the standard deviation of the total E. The normalized average of the entire signal is used as the envelope of the speech signal. Then, each critical band envelopes are calculated by Eqs. (1) and (3). The signal entropy of each critical band can be expressed in matrix form by Eq. (4). The entropy of each critical band is obtained by Eq. (4),

$$P[\Psi_m(t)] = \begin{bmatrix} h_1(1) & h_1(2) & \cdots & h_1(t) \\ h_2(1) & h_2(2) & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ h_{17}(1) & h_{17}(2) & \cdots & h_{17}(t) \end{bmatrix} = \mathrm{H}_m(t) = [h_m(1) \quad h_m(2) \quad \cdots \quad h_m(t)] \tag{4}$$

where $\mathrm{H}_m(t)$ is the decomposed signal entropy of the mth sub-band at a specific time *t*. The decomposed signal entropy has a high value at the time or critical bands that largely include speech.

2.3. Proposed voice activity detection algorithm

In this section, the new VAD function is proposed based on PWENS. The entropy neighbor slope is calculated from the decomposed signal entropy. First, the contaminated speech signal used in the experiment and decomposed signal entropy are shown in Figure 1.



(a)

(b)

(c)

(d)

Fig. 1. The signal decomposition and entropy for (a) the contaminated speech signal, (b) the signal entropy at a specific time *t*1, and (c) at *t*2. (d) shows the signal entropy for all critical bands.

Figure 1(a) shows a speech signal mixed with white noise for an SNR of 0 dB, and (d) shows the decomposed signal entropy of all critical bands that are decomposed by PWPD. The regions of high entropy are shown brightly in (d). The regions with noise are represented by the darker shade, and the critical bands that include mainly speech are represented by brighter shades. Figures 1(b) and 1(c) illustrate the signal entropy at *t*1 and *t*2, respectively. *t*1 indicates a speech area and *t*2 indicates a noise area. Therefore, the entropy of (b) has large variances, but (c) has values that are similar. PWENS also uses this difference feature. The proposed PWENS is calculated by Eq. (5).

$$\Im(t) = \sum_{m=2}^{17} \left| h_m(t) - h_{m-1}(t) \right| + \sum_{m=1}^{17} h_m(t) \tag{5}$$

PWENS is more valuable in the speech regions compared to noise regions. Then, the voice activity function is determined by the next step. For increased accuracy, two memory buffers are used.

$$\Re(t) = -\frac{1}{L}\sum_{i=1}^{L} \Im(i) \tag{6}$$

$$B^1 = \begin{cases} 1, & \text{if } \Re(t-1) > 0 \\ 0, & \text{otherwise} \end{cases} \qquad B^2 = \begin{cases} 1, & \text{if } \Re(t-2) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where L is the number of samples within 30 *ms*, and B^1 and B^2 are decision buffers. Finally, the VAD function is determined by Eq. (8).

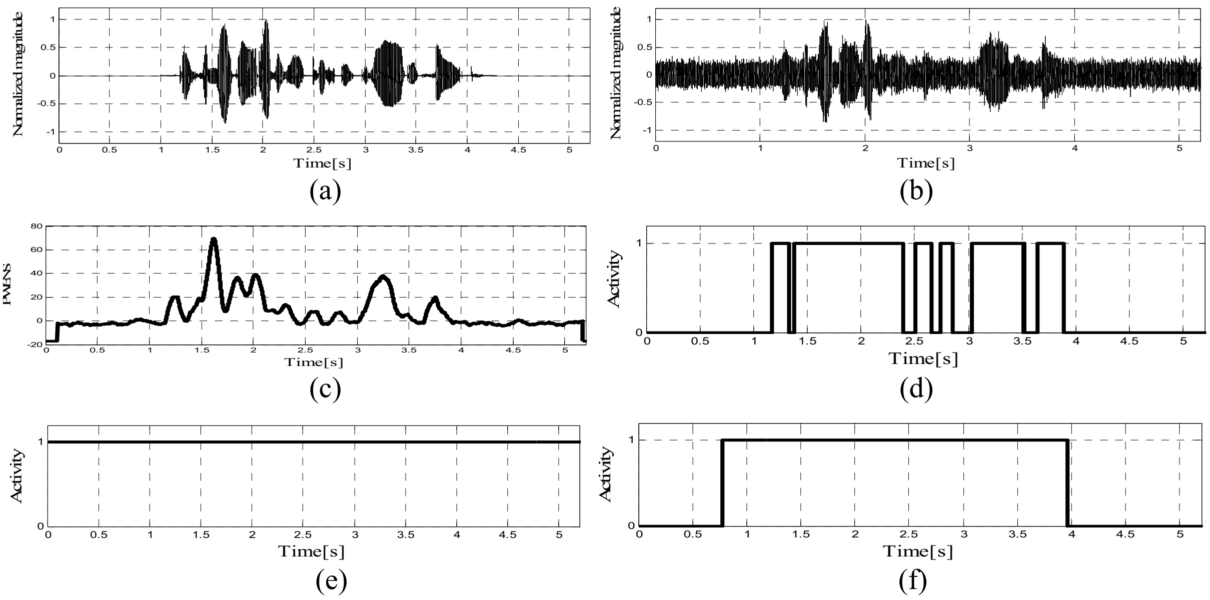
$$VAD(t) = \begin{cases} 1, & \Re(t) > 0 \,\&\, B^1 = 1 \,\&\, B^2 = 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The VAD function has the two values, 1 and 0, which represent the speech region and the noise region, respectively. In the next section, the experimental results are presented for a variety of experimental conditions.

3. Experiment and results

In this section, experimental results will be presented. For improved accuracy, the experiments were performed using a certified database and a variety of signal types. The experimental conditions are as follows. To test the performance of the proposed algorithm, the speech signal consisting of samples from the TIMIT database [17], and the noise signal comprised of samples from NOISEX-92 [18] that includes both stationary and non-stationary noises were used (of note, only stationary noises were used in this study). The data samples have a sampling rate of 16 kHz and a bit rate of 32 bps. The performance of the VAD decision is shown by comparing the basic entropy detection [12] with the results for G729B VAD [13] of ITU-T. The computed PWENS function obtained using Eq. (5) and the results of the VAD decision are shown in Figure 2. Figures 2(a) and 2(b) illustrate some of the speech signals used in our study. (c) is the result obtained for the calculated PWENS, which has a high value that corresponds to the speech area of (a). (d) shows the result of the VAD decision obtained using entropy detection, which performed well with the exception that the speech area was missing in some sections. (e) shows the result for G729B VAD, which indicates that the entire signal is comprised of speech. The G729B VAD therefore has poor performance in noisy environments. (f) shows the result for the proposed VAD algorithm, and in this case. All of the speech sections were accurately identified.

To objectively evaluate the performance of the VAD decision, more than 50 speech samples were used, we also repeated experiments using a variety of the stationary noise sources (white noise, car noise, babble noise, and pink noise) and in various SNR environments (0 dB, 5 dB, 10 dB, 15 dB, and 20 dB). Also, the relative error rate was used for the objective index. The relative error rate represents the error rate for the entire signal, and it is suitable for comparing a variety of algorithms. Figure 3 shows the result of the relative error rate in various noisy environments. As shown in Figure 3, the entropy detection has an error rate of 25~30%, and exhibits similar performance for all noise types. The G729B VAD performed poorly under most noise environments, while the proposed algorithm performed best compared to other algorithms. Our proposed algorithm was observed to have an error rate of less than 10%, and performed well under all noise environments.

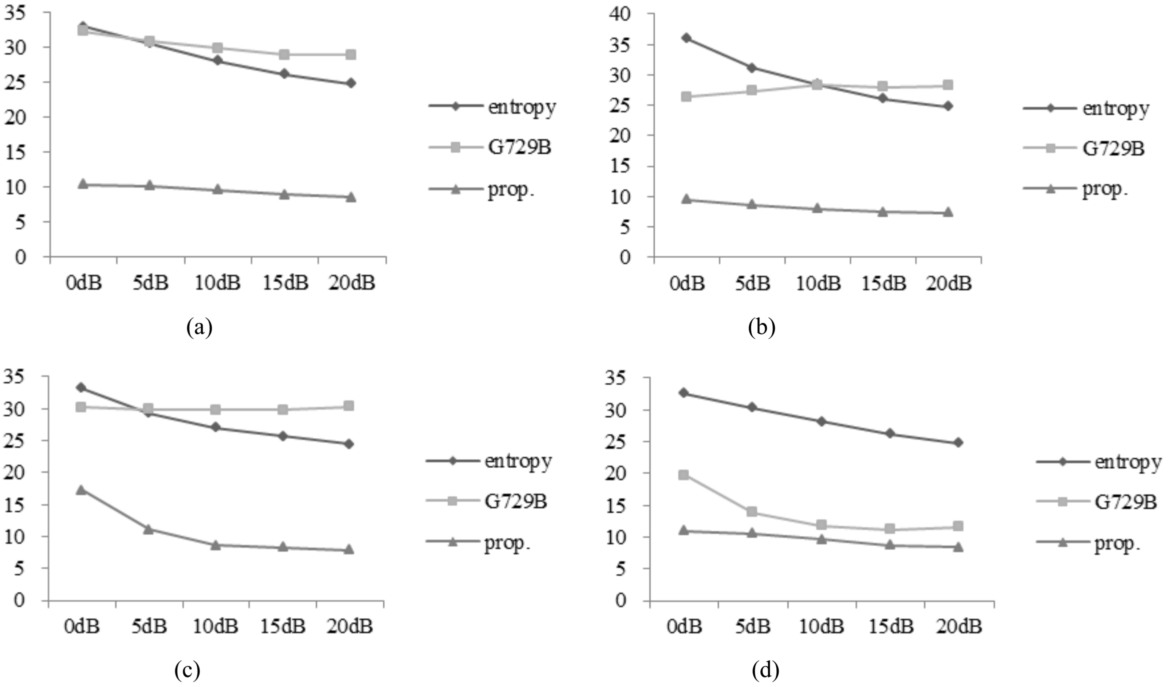


(a) (b)

(c) (d)

(e) (f)

Fig. 2. The result of the VAD decision and PWENS for (a) the clean speech signal (b) the contaminated speech signal, (c) the PWENS function, (d) the result for the entropy detection, (e) the result for G729B VAD, and (f) the result for the proposed algorithm



(a) (b)

(c) (d)

Fig. 3. The result of the relative error rate (a) in a white noise environment, (b) in a car noise environment, (c) in a babble noise environment, and (d) in a pink noise environment.

4. Conclusion

In this paper, we proposed a new VAD algorithm using the perceptual wavelet entropy neighbor slope. The proposed algorithm exhibits good performance in a variety of noisy environments due to suite psycho-acoustic model. The performance of the VAD decision was confirmed by performing experiments using many signal samples and in a variety of noisy environments. Currently, we are extending our research to enable us to successfully realize a usable system.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013R1A2A1A09015677) and a grant of the Korea Healthcare Technology R & D Project, Ministry of Health & Welfare (A092106).

References

[1] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993, pp. 69–129.

[2] G.K. Choi and S.H. Kim, Voice activity detection method using psycho-acoustic model based on speech energy maximization in noisy environments, Journal of the Acoustical Society of Korea **28** (2009), 447–453.

[3] D.G. Ha, S.J. Cho, G.G. Jin and O.K, Shin, Voice activity detection based on signal energy and entropy-difference in noisy environments, Journal of the Korean Society of Marine Engineering **32** (2008), 768–774.

[4] J. Ramíirez, J.C. Segura, C. Beníitez, A. de laTorre and A. Rubio, An effective subband OSF-based VAD with noise reduction for robust speech recognition, IEEE Trans. on Speech and Audio Processing **13** (2005), 1119–1129.

[5] R. Gemello, F. Mana and R.D. Mori, A modified ephraim-malah noise suppression rule for automatic speech recognition, Acoustics, Proceedings of Speech, and Signal Processing **1** (2004), 957–960.

[6] P. Teng and Y. Jia, Voice activity detection via noise reducing using non-negative sparse coding, IEEE Signal Processing Letters **20** (2013), 475–478.

[7] S.W. Deng and J.Q. Han, Statistical voice activity detection based on sparse representation over learned dictionary, Digital Signal Processing **23** (2013), 1228–1232.

[8] J. Han, S. Yook, K.W. Nam, S. Lee, D. Kim, S.H. Hong and I.Y. Kim, Erratum to: Comparative evaluation of voice activity detectors in single microphone noise reduction algorithms, Biomedical Engineering Letters **3** (2013), 58–58.

[9] D.W. Kim, J.C. Lee, Y.M. Park, I.Y. Kim and C.H. Im, Auditory brain-computer interfaces (bcis) and their practical applications, Biomedical Engineering Letters **2** (2012), 13–17.

[10] C.G Kim and B.S. Song, Proposal of a simultaneous ultrasound emission for efficient obstacle searching in autonomous wheelchairs, Biomedical Engineering Letters **3** (2013), 47–50.

[11] F. Wang and Z. Ji, Application of the dual-tree complex wavelet transform in biomedical signal denoising, Bio-Medical Materials and Engineering **24** (2014), 109–115.

[12] M. Asgari, A. Sayadian, M. Farhadloo and E.A. Mehrizi, Voice activity detection using entropy in spectrum domain, Proc. Telecommunication Networks and Applications Conference, 2008, 407–410.

[13] A. Benyassine, E. Shlomot, H.Y. Su, D. Massaloux, C. Lamblin and J.P. Petit, ITU recommendation G729 annex B: A silence compression scheme for use with G729 optimized for V.70 digital simultaneous voice and data applications, IEEE Commun. Mag. **35** (1997), 64–73.

[14] B. Mohammed and J. Rouat, Wavelet speech enhancement based on the teager energy operator, Signal Processing Letters IEEE **8** (2001), 10–12.

[15] C.E. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mobile Computing and Communications Review **5** (2001), 3–55.

[16] T. Kinnunen, E. Chermenko, M. Tuononen, P. Franti and H. Li, Voice activity detection using mfcc features and support vector machine, Proc. Speech and Computer **2** (2007), 556–561.

[17] V. Zue, S. Seneff and J. Glass, Speech database development at MIT: TIMIT and beyond, Speech Communication **9** (1990), 351–356.

[18] A. Varga and H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication **12** (1993), 247–251.