Computational Linguistics and Chinese Language Processing
Vol. 11, No. 1, March 2006, pp. 87-100

Voice Activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator

Bing-Fei Wu* and Kun-Ching Wang+

Abstract

In this paper, a new robust wavelet-based voice activity detection (VAD) algorithm derived from the discrete wavelet transform (DWT) and Teager energy operation (TEO) processing is presented. We decompose the speech signal into four subbands by using the DWT. By means of the multi-resolution analysis property of the DWT, the voiced, unvoiced, and transient components of speech can be distinctly discriminated. In order to develop a robust feature parameter called the speech activity envelope (SAE), the TEO is then applied to the DWT coefficients of each subband. The periodicity of speech signal is further exploited by using the subband signal auto-correlation function (SSACF) for. Experimental results show that the proposed SAE feature parameter can extract the speech activity under poor SNR conditions and that it is also insensitive to variable-level of noise.

Keywords: Voice Activity Detection, Auto-Correlation, Wavelet, Teager Energy

1. Introduction

Voice activity detection (VAD) refers to the ability to distinguish speech from noise and is an integral part of a variety of speech communication systems, such as speech coding, speech recognition, hand-free telephony, and echo cancellation. In the GSM-based communication system, a VAD scheme is used to lengthen the battery power through discontinuous transmission when speech-pause is detected [Freeman *et al.* 1989]. Moreover, a VAD algorithm can be used under a variable bit rate of the speech coding system in order to control the average bit rate and the overall quality of speech coding [Kondoz *et al*. 1994]. Perviously,

* Department of Electrical and Control Engineering, National Chiao-Tung University, HsinChu, Taiwan
E-mail: bwu@cssp.cn.nctu.edu.tw
+ Information & Communications Research Laboratories, Industrial Technology Research Institute, HsinChu, Taiwan
E-mail: kunching@itri.org.tw

[Received May 13, 2005; Revised Oct. 8, 2005; Accepted Oct. 17, 2005]

Sohn *et al.* [Sohn *et al.* 1998] presented a VAD algorithm that adopts a novel noise spectrum adaptation by applying soft decision techniques. The decision rule is drawn from the generalized likelihood ratio test by assuming that the noise statistics are known a priori. Cho *et al.* [Cho *et al.* 2001] presented an improved version of the algorithm designed by Sohn. Specifically, Cho presented a smoothed likelihood ratio test to reduce the detection errors. Furthermore, Beritelli *et al.* [Beritelli *et al.* 1998] developed a fuzzy VAD using a pattern matching block consisting of a set of six fuzzy rules. Additionally, Nemer *et al.* [Nemer *et al.* 2001] designed a robust algorithm based on higher order statistics (HOS) in the residual domain of the linear prediction coding coefficients (LPC). Meanwhile, the International Telecommunication Union-Telecommunications Sector (ITU-T) designed G. 729B VAD [Benyassine *et al.* 1997], which consists of a set of metrics, including line spectral frequencies (LSF), low band energy, zero-crossing rate (ZCR), and full-band energy. However, the common feature parameters mentioned above are based on averages over windows of fixed length or are derived through analysis based on a uniform time-frequency resolution. For example, it is well known that speech signals contain many transient components and exhibit the non-stationary property. The classical Fourier Transform (FT) works well for wide sense stationary signals but fails in the case of non-stationary signals since it applies only uniform-resolution analysis. Conversely, if the multi-resolution analysis (MRA) property of DWT [Strang *et al.* 1996] is used, the classification of speech into voiced, unvoiced or transient components can be accomplished.

The periodic property is an inherent characteristic of speech signals and is commonly used to characterize speech. In this paper, the periodic properties of subband signals are exploited to accurately extract speech activity. In fact, voiced or vowel speech sounds have a stronger periodic property than unvoiced sounds and noise signals, and this property is concentrated in low frequency bands. Thus, we let the low frequency bands have high resolution in order to enhance the periodic property by decomposing only the low band in each level. Three-level wavelet decomposition is further divided into four non-uniform subbands. Consequently, the well-known “Auto-Correlaction Function (ACF)” is defined in the subband domain to evaluate the periodic intensity of each subband, and is denoted as the “Subband Signal Auto-Correlaction Function (SSACF)”. Generally speaking, the existing methods for suppressing noise are almost all based on the frequency domain. However, these methods indeed waste too much computing power in on-line work. Considering computing complexity, the Teager energy operator (TEO), which is a powerful nonlinear operator and has been successfully used in various speech processing applications [Kaiser *et al.* 1990],[Bovik *et al.* 1993],[Jabloun *et al.* 1999] is applied to eliminate noise components from the wavelet coefficients in each subband priori to SSACF measurement. Consequently, to evaluate the periodic intensity of each subband signal, a Mean-Delta method [Ouzounov *et al.* 2004] is

applied in the envelope of each SSACF. First, the Delta SSACF, similar to the delta-cepstrum evaluation, is used to measure the local variation of each SSACF. Next, since the DSSACF is averaged over its length, the value of the Mean DSSACF (MDSSACF) can almost describe the amount of periodicity in each subband. Eventually, by only summing the values of the four MDSSACFs, we can apply a robust feature parameter, called the speech activity envelope (SAE) parameter. Experimental results show that the envelope of the SAE feature parameter can accurately indicate the boundary of speech activity under poor SNR conditions and that it is also insensitive to variable-level noise. In addition, the proposed wavelet-based VAD can be performed on-line.

This paper is organized as follows. Section 2 describes the proposed algorithm based on DWT and TEO. In addition, the proposed robust feature parameter is also discussed. Section 3 evaluates the performance of the proposed algorithm and compares it with that of other wavelet-based VAD algorithms and ITU-T G.729B VAD. Finally, Section 4 presents conclusions.

2. The Proposed Algorithm Based on DWT and TEO

In this section, each part for the proposed VAD algorithm is discussed in turn.

2.1 Discrete Wavelet Transform

The wavelet transform (WT) is based on time-frequency signal analysis. This wavelet analysis adopts a windowing technique with variable-sized regions. It allows the use of long time intervals when we want more precise low-frequency information, and shorter regions where we want high-frequency information. It is well known that speech signals contain many transient components and exhibit the non-stationary property. When we make use of the MRA property of the WT, better time-resolution is needed in the high frequency range to detect the rapid changing transient component of a signal, while better frequency resolution is needed in the low frequency range to track slowly time-varying formants more precisely. Through MRA analysis, the classification of speech into voiced, unvoiced or transient components can be accomplished. An efficient way to implement this DWT using filter banks was developed in 1988 by Mallat [Mallat 1989].

In Mallat's algorithm, the j-level approximations A_j and details D_j of the input signal are determined by using quadrature mirror filters (QMF). Figure 1 shows that the decomposed subband signals A and D are the approximation and detail parts of the input speech signal obtained by using the high-pass filter and low-pass filter, implemented with the Daubechies family wavelet, where the symbol $\downarrow 2$ denotes an operator of downsampling by 2. In fact, a voiced or vowel speech sound has more significant periodicity than an unvoiced sound on noise signal. Thus, the periodicity of a subband signal can be exploited to accurately

extract speech activity. In addition, the periodicity is almostly concentrated in low frequency bands, so we let the low frequency bands have high resolution in order to enhance the periodic property by decomposing only low bands in each level. Figure 2 employed the used structure of three-level wavelet decomposition. By using DWT, we can divide the speech signal into four non-uniform subbands. The wavelet decomposition structure can be used to obtain the most significant periodicity in the subband domain.

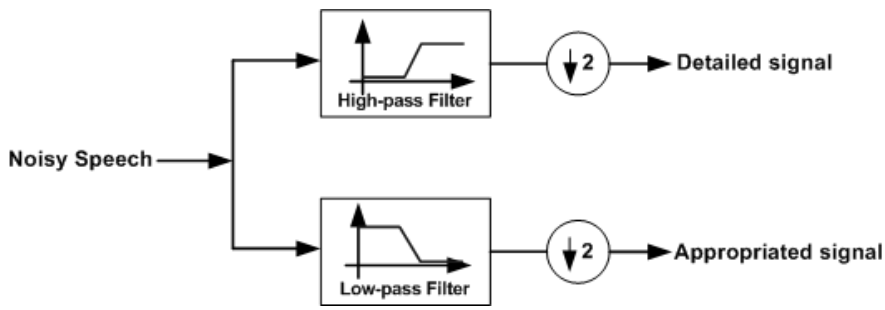

Figure 1. Discrete wavelet transform (DWT) using filter banks

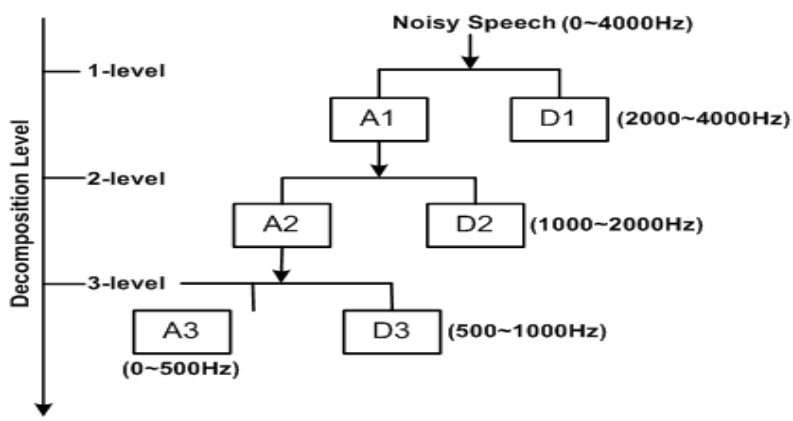

Figure 2. Structure of three-level wavelet decomposition

2.2 Teager Energy Operator

It has been observed that the TEO can enhance the discriminability between speech and noise and further suppress noise components from noisy speech signals [Jabloun *et al.* 1999]. Compared with the traditional noise suppression approach based on the frequency domain, the TEO based noise suppression can be more easily implemented through the time domain.

In continuous-time, the TEO is defined as

$$\psi_c[s(t)] = [\dot{s}(t)]^2 - s(t)\ddot{s}(t),$$

where $s(t)$ is a continuous-time signal and $\dot{s} = ds/dt$. In discrete-time, the TEO can be approximated by

$$\psi_d[s(n)] = s(n)^2 - s(n+1)s(n-1), \quad (1)$$

where $s(n)$ is a discrete-time signal.

Let us consider a speech signal $s(n)$ degraded by uncorrelated additive noise $u(n)$, the resulting signal is shown below:

$$y(n) = s(n) + u(n). \quad (2)$$

The Teager energy of the noisy speech signal $\psi_d[y(n)]$ is given by

$$\psi[y(n)] = \psi_d[s(n)] + \psi_d[u(n)] + 2\tilde{\psi}_d[s(n), u(n)], \quad (3)$$

where $\psi_d[s(n)]$ and $\psi_d[u(n)]$ are the Teager energy of the discrete speech signal and the additive noise, respectively. The subscript d means the "discrete." $\tilde{\psi}_d[s(n), u(n)]$ is the cross-ψ_d energy of $s(n)$ and $v(n)$, such that

$$\tilde{\psi}_d[s(n), u(n)] = s(n)u(n) - 0.5s(n-1)\cdot u(n+1) - 0.5s(n+1)\cdot u(n-1), \quad (4)$$

where the symbol $\cdot$ denotes the inner product. Since $s(n)$ and $u(n)$ are zero mean and independent, the expected value of the cross-ψ_d energy is zero. Thus, Eq.(5) can be derived from Eq.(3) as shown below:

$$E\{\psi[y(n)]\} = E\{\psi[s(n)]\} + E\{\psi[u(n)]\}. \quad (5)$$

Experimental results show that the Teager energy of the speech is much higher than that of the noise. Thus, compared with $E\{\psi_d[y(n)]\}$, $E\{\psi_d[u(n)]\}$ is negligible as shown by

$$E\{\psi_d[y(n)]\} \approx E\{\psi_d[s(n)]\}. \quad (6)$$

2.3 Subband Signal Auto-Correlation Function (SSACF)

The definition of the "Auto-Correlation Function (ACF)" used to measure the self-periodic intensity of subband signal sequences is shown below:

$$R(k) = \sum_{n=0}^{p-k} s(n)s(n+k), \quad k = 0,1,.....p, \quad (7)$$

where p is the length of ACF and k denotes the shift of the sample.

In this subsection, the ACF will be defined in the subband domain and called the "Subband Signal Auto-Correlation Function (SSACF)." It can be derived from the wavelet coefficients on each subband following TEO processing.

Figure 3 displays that the waveform of the normalized SSACFs ($R(0)=1$) of each subband, respectively. It is observed that the SSACF of voiced speech has more obvious peaks than that of unvoiced speech and white noise does. In addition, for unvoiced speech, the ACF has more intense periodicity than white noise does, especially in the $A3$ subband.

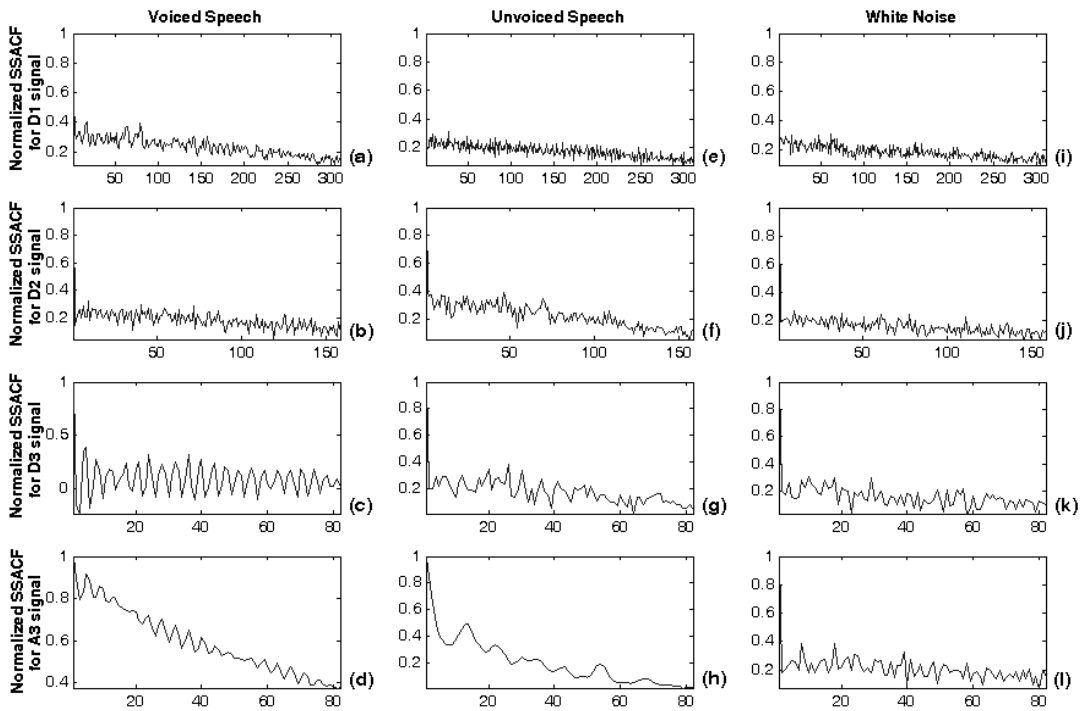

Figure 3. Examples of normalized SSACF for voiced speech, unvoiced speech and white noise

2.4 Mean of the absolute values of the DSSACF (MDSSACF)

To evaluate the periodic intensity of subband signals, a Mean-Delta method is applied here to each SSACF. First, a measure similar to delta cepstrum evaluation is used to estimate the periodic intensity of the SSACF, namely, the "Delta Subband Signal Auto-Correlation Function (DSSACF)," as shown below:

$$\dot{R}_M(k) = \frac{\sum_{m=-M}^{M} mR(k+m)}{\sum_{m=-M}^{M} m^2}, \tag{8}$$

where $\dot{R}_M$ is the DSSACF over an M -sample neighborhood.

For a particular frame, it is computed by using only the frame's SSACF (intra-frame processing), while the delta cepstrum is computed by using cepstrum coefficients from neighboring frames (inter-frame processing). It is observed that the DSSACF value is almost similar to the local variation over the SSACF.

Second, the delta of the SSACF is averaged over an M -sample neighborhood $\bar{R}_M$, where the mean of the absolute values of the DSSACF (MDSSACF) is given by

$$\bar{R}_M = \frac{1}{N_b} \sum_{k=0}^{N_b-1} \left| \dot{R}_M(k) \right|, \tag{9}$$

where N_b indicates the length of the subband signal.

Figure 4 shows that the SAE feature parameter is developed by summing the four MDSSACF values. Each subband can provide information for extracting voice activity precisely. It is found that the SAE feature parameter accurately indicates the boundary of speech activity under -5dB factory noise.

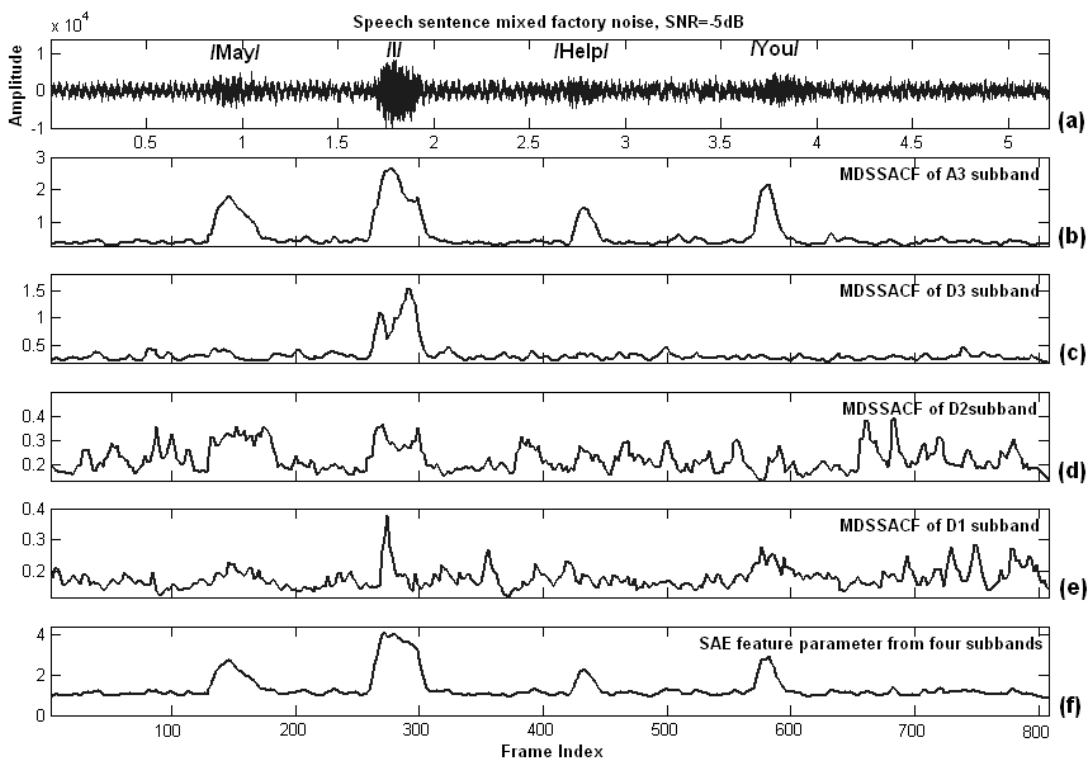

Figure 4. The development of the SAE feature parameter with and without band-decomposition

2.5 Block Diagram of the Proposed Wavelet-Based VAD

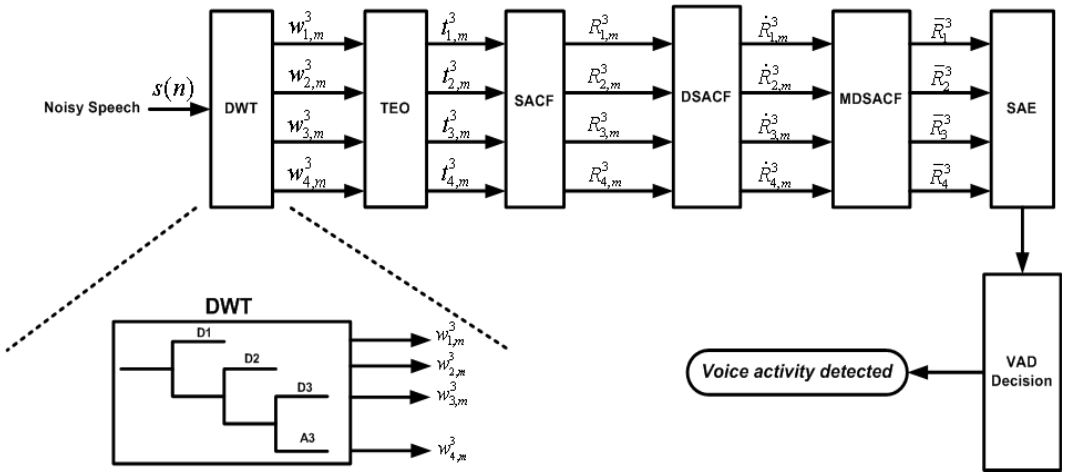

Figure 5. Block diagram of the proposed wavelet-based VAD

A block diagram of the proposed wavelet-based VAD algorithm is displayed in Figure 5. For a given level j, the wavelet transform decomposes the noisy speech signal into $j+1$ subbands corresponding to wavelet coefficients sets, $w_{k,n}^{j}$. In this case, for level $j=3$,

$$w_{k,m}^{3} = DWT\{s(n),3\}, \quad n=1....N,\ k=1....4, \tag{10}$$

where $w_{k,m}^{3}$ denotes the m^{th} coefficient of the k^{th} subband. N denotes the window length. The decomposed length of each subband is $N/2^{k}$. If $k=1$, $w_{1,m}^{3}$ corresponds to the $D1^{th}$ subband signal.

In TEO processing,

$$t_{k,m}^{3} = \psi_{d}[w_{k,m}^{3}],\ \text{k=1...4}. \tag{11}$$

The SSACF is derived from the Teager energy of noisy speech as follows:

$$R_{k,m}^{3} = R[t_{k,m}^{3}], \tag{12}$$

where $R[\cdot]$ denotes the auto-correlation operator.

Next, the DSSACF is given by

$$\dot{R}_{k,m}^{3} = \Delta[R_{k,m}^{3}], \tag{13}$$

where $\Delta[\cdot]$ denotes the Delta operator.

Then, the MDSSACF is obtained by

$$\bar{R}_k^3 = E[\dot{R}_{k,m}^3]. \tag{14}$$

where $E[\cdot]$ indicates the mean operator.

Finally, the SAE feature parameter is obtained by

$$SAE = \sum_{k=1}^{4} \bar{R}_k^3. \tag{15}$$

2.6 A VAD Decision Based on Adaptive Thresholding

In order to accurately determine the boundary of voice activity, the VAD decision is usually made through thresholding. To estimate the time-varying noise characteristics accurately, in this subsection, an adaptive threshold value is derived from the statistics of the SAE feature parameter during a noise-only frame, and the VAD decision process recursively updates the threshold by using the mean and variance of the values of the SAE parameters. We compute the initial noise mean and variance with the first five frames, assuming that the first five frames contain noise only. We then compute the thresholds for the speech and noise as follows [Gerven *et al.* 1997]:

$$T_s = \mu_n + \alpha_s \cdot \sigma_n, \tag{16}$$

$$T_n = \mu_n + \beta_n \cdot \sigma_n, \tag{17}$$

where T_s and T_n indicate the speech threshold and noise threshold, respectively. Similarly, μ_n and σ_n represent the mean and variance of the values of the SAE parameters, respectively.

The VAD decision rule is defined as follows:

$$\begin{aligned} &\text{if } (SAE(t) > T_s) \quad VAD(t)=1 \\ &\text{else if } (SAE(t) < T_n) \quad VAD(t)=0; \\ &\text{else } VAD(t)=VAD(t-1). \end{aligned} \tag{18}$$

If the detection result shows a noise period, the mean and variance of the values of the SAE are updated by as follows:

$$\mu_n(t) = \gamma \cdot \mu_n(t-1) + (1-\gamma) \cdot SAE(t), \tag{19}$$

$$\sigma_n(t) = \sqrt{[SAE_{buffer}{}^2]_{mean} - [\mu_n(t)]^2}, \tag{20}$$

$$[SAE_{buffer}{}^2]_{mean}(t) = \gamma \cdot [SAE_{buffer}{}^2]_{mean}(t-1) + (1-\gamma) \cdot SAE(t)^2 . \quad (21)$$

Here, $[SAE_{buffer}{}^2]_{mean}(t-1)$ is a mean of the buffer of the SAE value during a noise-only frame. We then update the thresholds by using the updated mean and variance of the values of the SAE parameters. Figure 6 displays the VAD decision, based on the adaptive threshold strategy. It is clearly seen that the boundary of voice activity has been accurately extracted. The two thresholds are updated during voice-inactivity but not during voice-activity.

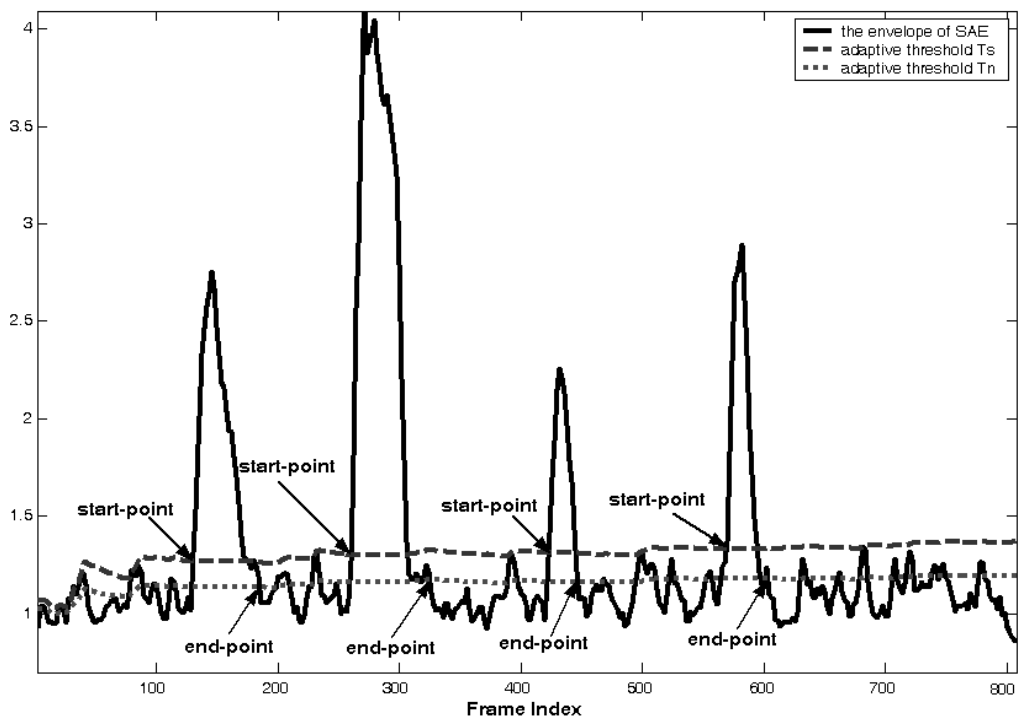

Figure 6. Adaptive thresholding strategy for extracting the boundary of voice activity

3. Simulation Results

The proposed wavelet-based VAD algorithm operates on a frame-by-frame basis (frame size = 256 samples/frame, overlapping size = 64 samples, $M = 8$, $\alpha_s = 5$, $\beta_n = -1$ and $\gamma = 0.95$). The results of speech activity detection were obtained under three kinds of background noise, which included white noise, car noise, and factory noise, taken from the Noisex-92 database [Varga *et al.* 1993]. The speech database contained 60 speech phrases (in

Mandarin and in English) spoken by 32 native speakers (22 males and 10 females), sampled at 8000 Hz and linearly quantized at 16 bits per sample. The two probabilities of correctly detecting speech frames, P_{cs}, and falsely detecting speech frames, P_f, were the ratio of the correct speech decision to the total number of hand-labeled speech frames and the ratio of the false speech decision or false noise decision to the total number of hand-labeled frames used to objectively measure performance of these three VADs.

Table 1 compares the performance of the proposed wavelet-based VAD, the wavelet-based VAD proposed by Chen *et al.* [Chen *et al*. 2002], and the ITU standard G.729B [Benyassine *et al*. 1997] under three types of noise and three specific SNR values: 30,10, and -5dB. From this table, it can be seen that in terms of the average correct and false speech detection probability, the proposed wavelet-based VAD is superior to Chen's VAD algorithm and G.729B VAD over all three SNRs under various types of noise. Table 2 shows the computing time of the three VAD algorithms, where Matlab was used on a Celeron 2.0G CPU PC to process 138 frames of a speech signal. It is found that the computing time consists of the time needed for feature extraction, and the voice activity decision process. The computing time of Chen's VAD was nearly twelve times longer than that of proposed VAD. We attribute the computing time of Chen's VAD to five-level wavelet decomposition. Its feature parameter is based on 17 critical-subbands, using the perceptual wavelet packet transform (PWPT). And after, wavelet reconstruction is required in Chen's approach. In our approach, however, we only divide four subbands using wavelet transform and do not waste extra computing time on wavelet reconstruction.

Table 1. Performance of the proposed wavelet-based approach, Chen's wavelet-based approach [9] and G.729B VAD

Noise Conditions		P_{cs} **(%)**			P_f **(%)**		
Type	**SNR(dB)**	**Proposed VAD**	**Chen's VAD**	**G.729B VAD**	**Proposed VAD**	**Chen's VAD**	**G.729B VAD**
Car Noise	***30***	99.1	97.3	92.1	6.2	6.9	7.3
	10	97.3	96.1	86.5	8.6	9.3	16.3
	-5	92.6	93.5	72.3	10.5	10.9	21.5
Factory Noise	***30***	96.9	97.2	96.9	7.6	10.3	9.1
	10	93.1	94.1	82.3	8.8	13.2	18.9
	-5	87.2	85.6	70.7	10.9	15.4	26.4
White Noise	***30***	99.1	97.2	98.4	1.3	1.9	2.0
	10	98.5	98.1	86.3	1.5	1.8	3.6
	-5	93.2	92.9	60.5	1.6	2.3	3.3
Average		**95.22**	**94.67**	**82.89**	**6.33**	**8**	**12.04**

Table 2. The computing time required by the three VAD algorithms

VAD type	Feature Extraction Processing	Voice Activity Decision
G.729B	0.048 s	0.023 s
Chen'VAD	4.126 s	0.098 s
Proposed VAD	0.23 s	0.12 s

Figure 7 shows the performance of the proposed VAD for an utterance produced continuously under variable-level noise. We decreased and increased the level of background noise and set the SNR value to 0 dB. Compared with the envelope of the VAS parameter, it is observed that the envelope of the SAE parameter was more robust against the variable noise-level and able to extract the exact boundary of the voice activity. This can be mainly attributed to the fact that the value of each MDSSACF depends on the amount of variation of the ACF, not on the energy level of the signal.

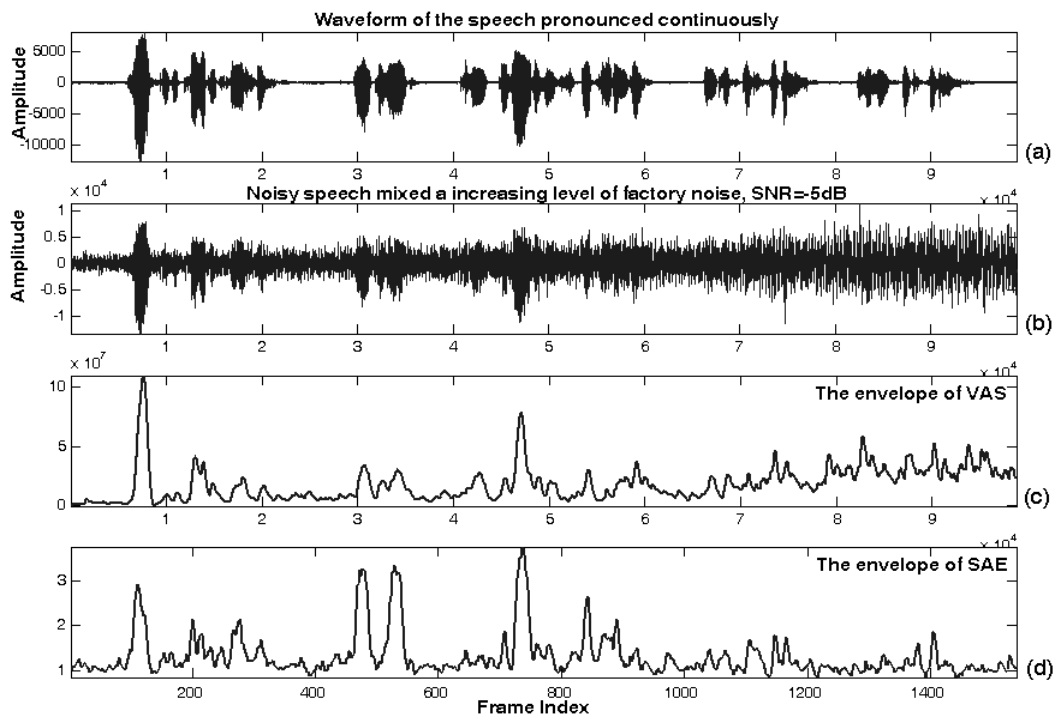

Figure 7. The effects of variable noise-level on the proposed SAE parameter and Chen's VAS parameter for a noisy speech sentence consisting of continuous words

4. Discussion

Compared with Chen's wavelet-based VAD, our experimental results shows that the proposed wavelet-based VAD algorithm is more suitable for on-line work. In terms of complexity, Chen's wavelet-based VAD algorithm [Chen *et al*. 2002] requires five-level wavelet decomposition to decompose the speech signal into 17 critical-subbands by using PWPT. In addition, it uses more extra computing time to complete wavelet reconstruction. In tests with non-stationary noise, it was found that each MDSSACF depends only on the amount of variation of the normalized ACF, not on the energy level of the signal, so the envelope of the proposed SAE feature parameter is insensitive to variable-level noise. Conversely, in Chen's wavelet-based method, the VAS feature parameter closely depends on the subband energy, so the achieved performance is poor under variable-level noise.

Acknowledgments

This work was supported by National Science Council of Taiwan under grant no. NSC 94-2213-E-009-066.

Reference

Benyassine, A., E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, 35(9), 1997, pp.64-73.

Beritelli, F., S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected Areas in Communications*, 16(9), 1998, pp.1818-1829.

Bovik, A. C., P. Maragos, and T. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Transactions on Signal Processing*, 41(12), 1993, pp.3245-3265.

Chen, S.H., and J.F. Wang, "A Wavelet-based Voice Activity Detection Algorithm in Noisy Environments," *International Conference on 9th Electronics, Circuits and Systems*, 2002, pp.995-998.

Cho, Y. D., and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, 8(10), 2001, pp.276-278.

Freeman ,D. K., G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan European digital cellular mobile telephone service," International Conference on *Acoustics, Speech, and Signal Processing*, 1989, pp.369-372.

Gerven , S. V., and F. Xie, "A comparative study of speech detection methods," In *Proceedings of Eurospeech*, 3, 1997, pp.1095-1098.

Jabloun, F., A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *IEEE Signal Processing Letters*, 6(10), 1999, pp.259-261.

Kaiser, J. F., "On a simple algorithm to calculate the 'energy' of a signal," *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp.381-384.

Kondoz, A. M., *Digital Speech Coding for Low Bit Rate Communications Systems*, John Wiley & Sons Ltd., 1994.

Mallat, S., "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 1989, pp.674-693.

Nemer, E., R. Goubran and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, 9(3), 2001, pp.217-231.

Ouzounov, A., "A Robust Feature for Speech Detection," *Cybernetics and Information Technologies*, 4(2), 2004, pp.3-14.

Sohn, J., and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," International Conference on *Acoustics, Speech, and Signal Processing*, 1998, pp.365-368.

Strang, G., and T. Nquyen, *Wavelet and Filter Banks*, Wellesley-Cambridge Press, 1996.

Varga, A., 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, pp.247-251.