Robust Voice Activity Detection Based on Discrete Wavelet Transform

Kun-Ching Wang

Department of Information Technology & Communication
Shin Chien University
kunching@mail.kh.usc.edu.tw

Abstract

This paper mainly addresses the problem of determining voice activity in presence of noise, especially in a dynamically varying background noise. The proposed voice activity detection algorithm is based on structure of three-layer wavelet decomposition. Appling auto-correlation function into each subband exploits the fact that intensity of periodicity is more significant in sub-band domain than that in full-band domain. In addition, Teager energy operator (TEO) is used to eliminate the noise components from the wavelet coefficients on each subband. Experimental results show that the proposed wavelet-based algorithm is prior to others and can work in a dynamically varying background noise.

Keywords: voice activity detection, auto-correlation function, wavelet transform, Teager energy operator

1. Introduction

Voice activity detection (VAD) refers to the ability of distinguishing 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. Although the existed VAD algorithms performed reliably, their feature parameters are almost depended on the energy level and sensitive to noisy environments [1-4]. So far, a wavelet-based VAD is rather less discussed although wavelet analysis is much suitable for speech property. S.H. Chen et al. [5] shown that the proposed VAD is based on wavelet transform and has an excellent performance. In fact, their approach is not suitable for practical application such as variable-level of noise conditions. Besides, a great computing time is needed for accomplishing wavelet reconstruction to decide whether is speech-active or not.

Compared with Chen's VAD approach, the proposed decision of VAD only depends on three-layer wavelet decomposition. This approach does not need any computing time to waste the wavelet reconstruction. In addition, the four non-uniform subbands are generated from the wavelet-based approach and the well-known "auto-correlaction function (ACF)" is adopted to detect the periodicity of subband. We refer the ACF defined in subband domain as subband auto-correlation function (SACF). Due to that periodic property is mainly focused on low frequency bands, so we let the low frequency bands have high resolution to enhance the periodic property by decomposing only low band on each layer. In addition to the SACF, enclosed herein the Teager energy operator (TEO) is regarded as a pre-processor for SACF. The TEO is a powerful nonlinear operator and has been successfully used in various speech processing applications [6-7]. F. Jabloun et al. [8] displayed that TEO can suppress the car engine noise and be easily implemented through time domain in Mel-scale subband. The later experimental result will prove that the TEO can further enhance the detection of subband periodicity.

To accurately count the intensity of periodicity from the envelope of the SACF, the Mean-Delta (MD) method [9] is utilized on each subband. The MD-based feature parameter has been presented for the robust development of VAD, but is not performed well in the non-stationary noise shown in the followings. Eventually, summing up the four values of MDSACF (Mean-Delta of Subband Auto-Correlation Function, a new feature parameter called "speech activity envelope (SAE)" is further proposed. Experimental results show that the envelope of the new SAE parameter can point out the boundary of speech activity under the poor SNR conditions and it is also insensitive to variable-level of noise.

This paper is organized as follows. Section 2 describes the concept of discrete wavelet transform (DWT) and shows the used structure of three-layer wavelet decomposition. Section 3 introductions the derivation of Teager energy operator (TEO) and displays the efficiency of subband noise suppression. Section 4 describes the proposed feature parameter, and the block diagram of proposed wavelet-based VAD algorithm is outlined in Section 5. Section 6 evaluates the performance of the algorithm and compare to other two wavelet-based VAD algorithm and ITU-T G.729B VAD. Finally, Section 7 discusses the conclusions of experimental results.

2. Wavelet transform

The wavelet transform (WT) is based on a time-frequency signal analysis. The wavelet analysis represents a windowing technique with variable-sized regions. It allows the use of long time intervals where 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 non-stationary property. Making use of the multi-resolution analysis (MRA) property of the WT, better time-resolution is needed a high frequency range to detect the rapid changing transient component of the signal, while better frequency resolution is needed at low frequency range to track the slowly time-varying formants more precisely [10]. Figure 1 displays the structure of three-layer wavelet decomposition utilized in this paper. We decompose an entire signal into four non-uniform subbands including three detailed scales such as D1, D2 and D3 and one appropriated scale such A3.

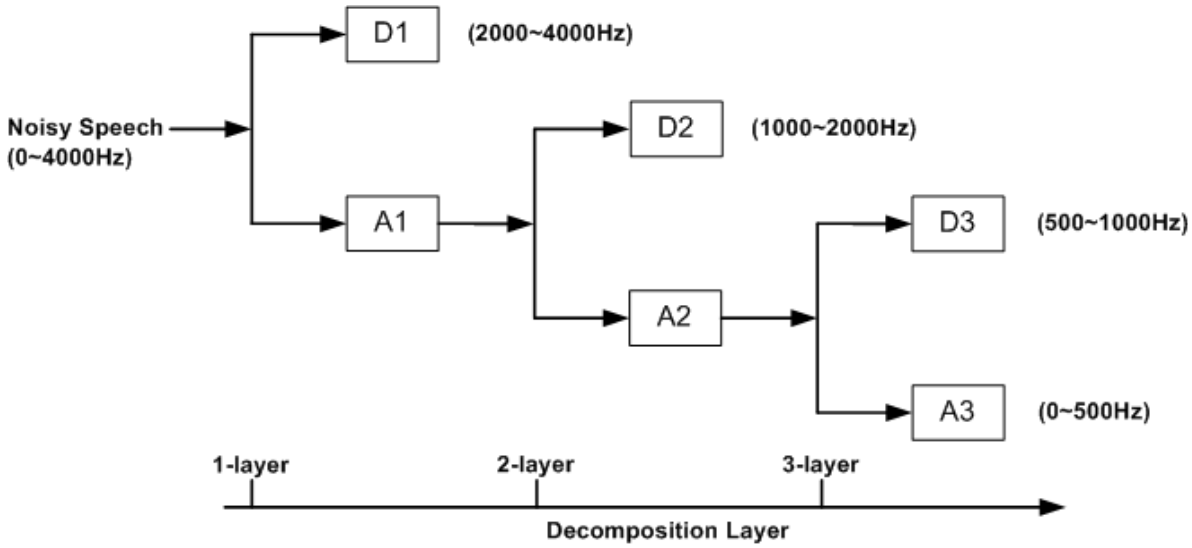

Figure 1. Structure of three-layer wavelet decomposition

3. Mean-delta method for subband auto-correlation function

The well-known definition of the term "Auto-Correlation Function (ACF)" is usually used for measuring the self-periodic intensity of signal sequences shown as below:

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

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

In order to increase the efficiency of ACF about making use of periodicity detection to detect speech, the ACF is defined in subband domain, which called "subband auto-correlation function (SACF)". Figure 2 clearly illustrates the normalized SACFs for each subband when input speech is contaminated by white noise. In addition, a normalization factor is applied to the computation of SACF. This major reason is to provide an offset for insensitivity on variable energy level. From this figure, it is observed that the SACF of voiced speech has more obviously peaks than that of unvoiced speech and white noise. Similarly, for unvoiced speech the ACF has greater periodic intensity than white noise especially in the approximation $A3$.

Furthermore, a Mean-Delta (MD) method [9] over the envelope of each SACF is utilized herein to evaluate the corresponding intensity of periodicity on each subband. First, a measure which similar to delta cepstrum evaluation is mimicked to estimate the periodic intensity of SACF, namely "Delta Subband Auto-Correlation Function (DSACF)", shown below:

$$\dot{R}_M(k) = \frac{\sum_{m=-M}^{M} m\left(\frac{R(k+m)}{R(0)}\right)}{\sum_{m=-M}^{M} m^2}, \tag{2}$$

where $\dot{R}_M$ is DSACF over an M-sample neighborhood ($M = 3$ in this study).

It is observed that the DSACF measure is almost like the local variation over the SACF. Second, averaging the delta of SACF over a M-sample neighborhood $\dot{R}_M$, a mean of the absolute values of the DSACF (MDSACF) is given by

$$\bar{R}_M = \frac{1}{N}\sum_{k=0}^{N-1}\left|\dot{R}_M(k)\right|. \tag{3}$$

Observing the above formulations, the Mean-Delta method can be used to value the number and amplitude of peak-to-valley from the envelope of SACF. So, we just only sum up the four values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale, a robust feature parameter called "speech activity envelope (SAE)" is further proposed.

Figure 3 displays that the MRA property is important to the development of SAE feature parameter. The proposed SAE feature parameter is respectively developed with/without band-decomposition. In Figure 3(b), the SAE without band-decomposition only provides obscure periodicity and confuses the word boundaries. Figure 3(c)~Figure 3(f) respectively show each value of MDSACF from D1 subband to A3 subband. It implies that the value of MDSACF can provide the corresponding periodic intensity for each subband. Summing up the four values of MDSACFs, we can form a robust SAE parameter. In Figure 3(g), the SAE with band-decomposition can point out the word boundaries accurately from its envelope.

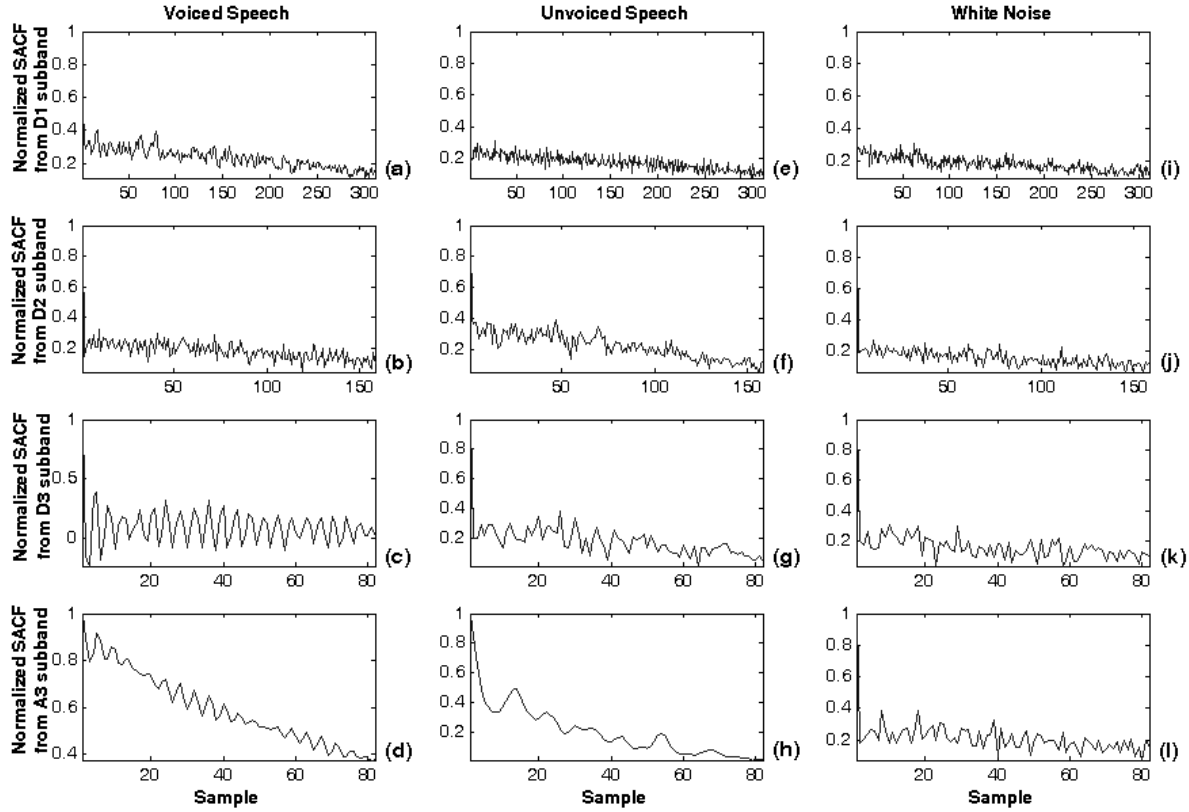

Figure 2. SACF on voiced, unvoiced signals and white noise

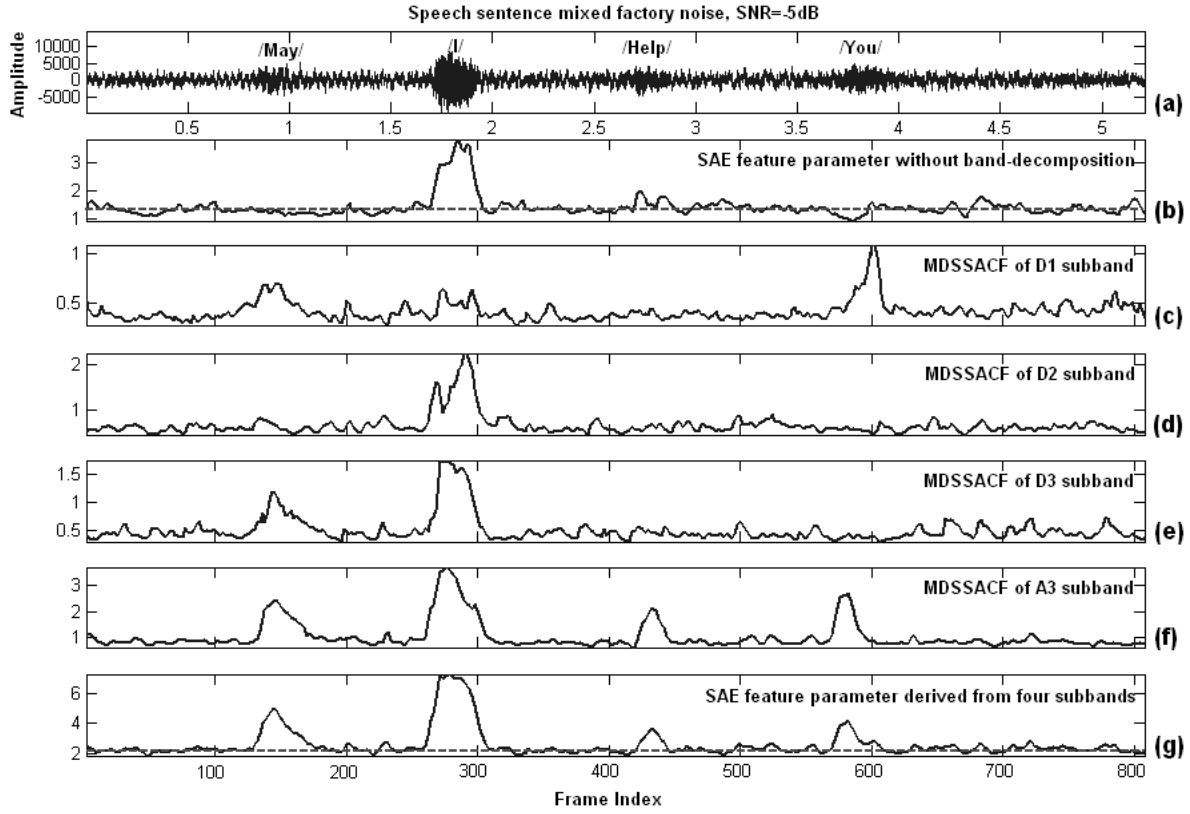

Figure 3. SAE with/without band-decomposition

4. Teager energy operator

The Teager energy operator (TEO) is a powerful nonlinear operator, and can track the modulation energy and identify the instantaneous amplitude and frequency [7-10].

In discrete-time, the TEO can be approximate by

$$\Psi_d[s(n)] = s(n)^2 - s(n+1)s(n-1), \tag{4}$$

where $\Psi_d[s(n)]$ is called the TEO coefficient of discrete-time signal $s(n)$.

Figure 4 indicates that the TEO coefficients not only suppress noise but also enhance the detection of subband periodicity. TEO coefficients are useful for SACF to discriminate the difference between speech and noise in detail.

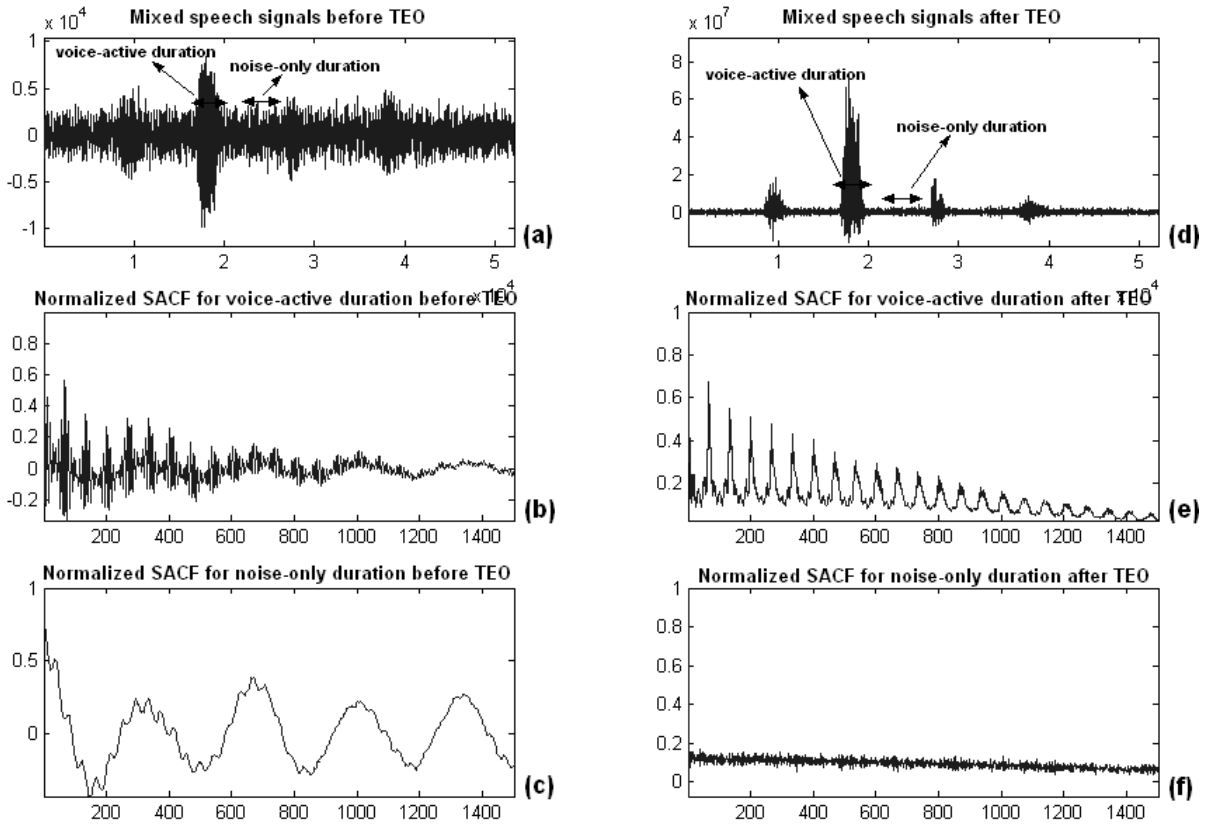

Figure 4. Illustration of TEO processing for the discrimination between speech and noise by using periodicity detection

5. Proposed voice activity detection algorithm

In this section, the proposed VAD algorithm based on DWT and TEO is presented. Fig. 8 displays the block diagram of the proposed wavelet-based VAD algorithm in detail. For a given layer j, the wavelet transform decomposed the noisy speech signal into $j+1$ subbands corresponding to wavelet coefficients sets $w_{k,n}^{j}$. In this case, three-layer wavelet decomposition is used to decompose noisy speech signal into four non-uniform subbands including three detailed scales and one appropriated scale. Let layer $j=3$,

$$w_{k,m}^{3} = DWT\{s(n),3\}, \quad n=1....N,\ k=1....4, \tag{5}$$

where $w_{k,m}^{3}$ defines the m^{th} coefficient of the k^{th} subband. N denotes as window length. The decomposed length of each subband is $N/2^{k}$ in turn.

For each subband signal, the TEO processing [8] is then used to suppress the noise

component, and also enhance the periodicity detection. In TEO processing,

$$t_{k,m}^3 = \psi_d[w_{k,m}^3], \quad k = 1...4. \tag{6}$$

Next, the SACF measures the ACF defined in subband domain, and it can sufficiently discriminate the dissimilarity among of voiced, unvoiced speech sounds and background noises from wavelet coefficients. The SACF derived from the Teager energy of noisy speech is given by

$$R_{k,m}^3 = R[t_{k,m}^3], \quad k = 1...4. \tag{7}$$

To count the intensity of periodicity from the envelope of the SACF accurately, the Mean-Delta (MD) method [9] is utilized on each subband.

The DSACF is given by

$$\dot{R}_{k,m}^3 = \Delta[R_{k,m}^3], \quad k = 1...4. \tag{8}$$

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

Then, the MDSACF is obtained by

$$\overline{R}_k^3 = E[\dot{R}_{k,m}^3]. \tag{9}$$

where $E[\cdot]$ denotes the operator of mean.

Finally, we sum up the values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale and denote as SAE feature parameter given by

$$SAE = \sum_{k=1}^{4} \overline{R}_k^3 . \tag{10}$$

6. Experimental results

In our first experiment, the results of speech activity detection are tested in three kinds of background noise under various values of the SNR. In the second experiment, we adjust the variable noise-level of background noise and mix it into the testing speech signal.

6.1. Test environment and noisy speech database

The proposed wavelet-based VAD algorithm is based on frame-by-frame basis (frame size = 1024 samples/frame, overlapping size = 256 samples). Three noise types, including white noise, car noise and factory noise, are taken from the Noisex-92 database in turn [11]. The speech database contains 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. To vary the testing conditions, noise is added to the clean speech signal to create noisy signals at specific SNR of 30, 10, -5 dB.

6.2. Evaluation in stationary noise

In this experiment we only consider stationary noise environment. The proposed wavelet-based VAD is tested under three types of noise sources and three specific SNR values mentioned above. Table 1 shows the comparison between the proposed wavelet-based VAD and other two wavelet-based VAD proposed by Chen et al. [5] and J. Stegmann [12] and ITU standard VAD such as G.729B VAD [4], respectively. The results from all the cases involving various noise types and SNR levels are averaged and summarized in the bottom row of this table. We can find that the proposed wavelet-based VAD and Chen's VAD algorithms are all superior to Stegmann's VAD and G.729B over all SNRs under various types of noise. In terms of the average correct and false speech detection probabilities, the proposed wavelet-based VAD is comparable to Chen's VAD algorithm. Both the algorithms are based on the DWT and TEO processing. However, Chen et al. decomposed the input speech signal into 17 critical-subbands by using perceptual wavelet packet transform (PWPT). To obtain a robust feature parameter, called as "VAS" parameter, each critical subband after their processing is synthesized individually while other 16 subband signals are set to zero values. Next, the VAS parameter is developed by merging the values of 17 synthesized bands. Compare to the analysis/synthesis of wavelet from S. H. Chen et al., we only consider analysis of wavelet. The structure of three-layer decomposition leads into four non-uniform bands as front-end processing. For the development of feature parameter, we do not again waste extra computing power to synthesize each band. Besides, Chen's VAD algorithm must be performed in entire speech signal. The algorithm is not appropriate for real-time issue since it does not work on frame-based processing. Conversely, in our method the decisions of voice activity can be accomplished by frame-by-frame processing. Table 2 indicates that the computing time for the listed VAD algorithms running Matlab programming in Celeron 2.0G CPU for processing 118 frames of an entire recording. It is found that the computing time of Chen's VAD is nearly four times greater than that of other three VADs. Besides, the

computing time of Chen's VAD is closely relative to the entire length of recording.

Table 1. Comparison performance.

Noise Conditions		The probability of *correctly* detecting speech frames (%)				The probability of *falsely* detecting speech frames (%)			
Type	SNR(dB)	Proposed VAD	Chen's VAD	Stegmann's VAD	G.729B VAD	Proposed VAD	Chen's VAD	Stegmann's VAD	G.729B VAD
Car Noise	*30*	99.3	97.3	90.2	94.5	6.1	6.9	8.2	6.3
	10	97.8	96.1	85.3	90.3	8.4	9.3	13.5	12.3
	-5	92.9	93.5	79.1	82.7	10.5	10.9	16.6	17.5
Factory Noise	*30*	97.4	97.2	94.5	97.2	7.3	10.3	11.2	7.1
	10	93.2	94.1	83.1	88.4	8.8	13.2	14.6	13.4
	-5	87.8	85.6	75.3	80.7	10.7	15.4	17.3	19.2
White Noise	*30*	99.4	97.2	95.3	98.3	1.2	1.9	4.5	2.3
	10	98.6	98.1	90.1	86.3	1.3	1.8	6.7	2.9
	-5	93.4	92.9	85.8	84.3	1.5	2.3	10.1	3.7
Average		**95.5**	**94.7**	**86.5**	**89.2**	**6.2**	**8**	**11.4**	**9.4**

Table 2. Illustrations of subjective listening evaluation and the computing time

VAD types	Computing time (sec)
Proposed VAD	0.089
Chen's VAD [5]	0.436
Stegmann's VAD [12]	0.077
G.729B VAD [4]	0.091

6.3. Evaluation in non-stationary noise

In practice, the additive noise is non-stationary in real-world, since its statistical property change over time. We add the decreasing and increasing level of background noise on a clean speech sentence in English and the SNR is set 0 dB. Figure 6 exhibits the comparisons among proposed wavelet-based VAD, other one wavelet-based VAD respectively proposed by S. H. Chen et al. [5] and MD-based VAD proposed by A. Ouzounov [9]. Regarding to this figure, the mixed noisy sentence "May I help you?" is shown in Fig. 9(a). The increasing noise-level and decreasing noise-level are added into the front and the back of clean speech signal. Additionally, an abrupt change of noise is also added in the middle of clean sentence. The three envelopes of VAS, MD and SAE feature parameters are showed in Figure 6(b)~Figure

6(d), respectively. It is found that the performance of Chen's VAD algorithm seems not good in this case. The envelope of VAS parameter closely depends on the variable level of noise. Similarly, the envelope of MD parameter fails in variable level of noise. Conversely, the envelope of proposed SAE parameter is insensitive to variable-level of noise. So, the proposed wavelet-based VAD algorithm is performed well in non-stationary noise.

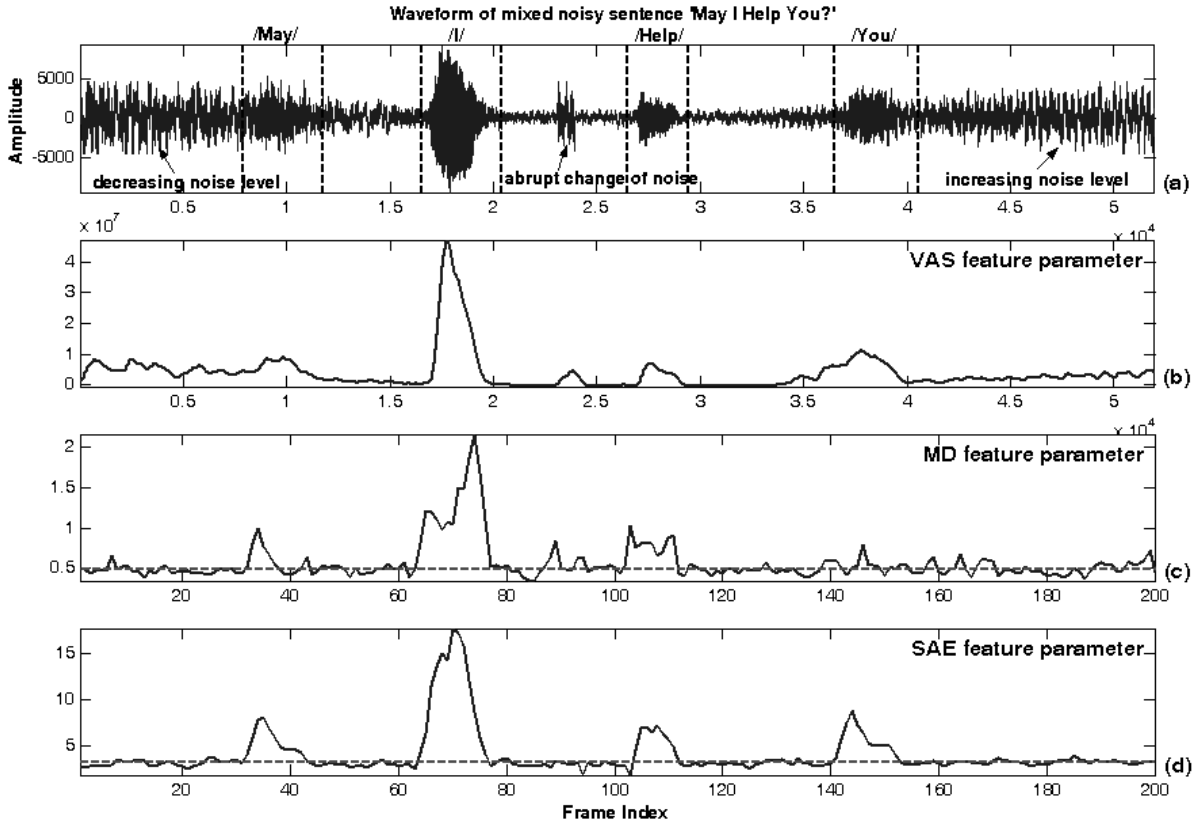

Figure 6. Comparisons among VAS, MD and proposed SAE feature parameters

7. Conclusions

The proposed VAD is an efficient and simple approach and mainly contains three-layer DWT (discrete wavelet transform) decomposition, Teager energy operation (TEO) and auto-correlation function (ACF). TEO and ACF are respectively used herein in each decomposed subband. In this approach, a new feature parameter is based on the sum of the values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale, and it has been shown that the SAE parameter can point out the boundary of speech activity and its envelope is insensitive to variable noise-level environment. By means of the MRA property of DWT, the ACF defined in subband domain sufficiently discriminates the dissimilarity among of voiced, unvoiced speech sounds and background

noises from wavelet coefficients. For the problem about noise suppression on wavelet coefficients, a nonlinear TEO is then utilized into each subband signals to enhance discrimination among speech and noise. Experimental results have been shown that the SACF with TEO processing can provide robust classification of speech due to that TEO can provide a better representation of formants resulting distinct periodicity.

References

[1] Cho, Y. D. and Kondoz, A., "Analysis and improvement of a statistical model-based voice activity detector", *IEEE Signal Processing Lett.*, Vol 8, 276-278, 2001.

[2] Beritelli, F., Casale, S. and Cavallaro, A., "A robust voice activity detector for wireless communications using soft computing", *IEEE J. Select. Areas Comm.*, Vol 16, 1818-1829, 1998.

[3] Nemer, E., Goubran, R. and Mahmoud, S., "Robust voice activity detection using higher-order statistics in the LPC residual domain", *IEEE Trans. Speech and Audio Processing*, Vol. 9, 217-231, 2001.

[4] Benyassine, A., Shlomot, E., Su, H. Y., Massaloux, D., Lamblin, C. and Petit, J. P., "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*, Vol. 35, 64-73, 1997.

[5] Chen, S. H. and Wang, J. F., "A Wavelet-based Voice Activity Detection Algorithm in Noisy Environments", *2002 IEEE International Conference on Electronics, Circuits and Systems (ICECS2002)*, 995-998, 2002.

[6] Kaiser, J. F., "On a simple algorithm to calculate the 'energy' of a signal", *in Proc. ICASSP'90*, 381-384, 1990.

[7] Maragos, P., Quatieri, T., and Kaiser, J. F., "On amplitude and frequency demodulation using energy operators", *IEEE Trans. Signal Processing*, Vol. 41, 1532-1550, 1993.

[8] Jabloun, F., Cetin, A. E., and Erzin, E., "Teager energy based feature parameters for speech recognition in car noise", *IEEE Signal Processing Lett.*, Vol. 6, 259-261, 1999.

[9] Ouzounov, A., "A Robust Feature for Speech Detection", *Cybernetics and Information*

Technologies, Vol. 4, No 2, 3-14, 2004.

[10] Stegmann, J., Schroder, G., and Fischer, K. A., "Robust classification of speech based on the dyadic wavelet transform with application to CELP coding", *Proc. ICASSP*, Vol. 1, 546 - 549, 1996.

[11] Varga, A. and Steeneken, H. J. M., "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 Commun.*, Vol. 12, 247-251, 1993.

[12] Stegmann, J. and Schroder, G., "Robust voice-activity detection based on the wavelet transform", *IEEE Workshop on Speech Coding for Telecommunications Proceeding*, 99 - 100, 1997.