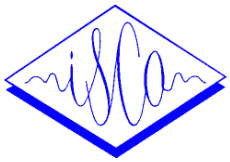

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

A VOICE ACTIVITY DETECTION ALGORITHM BASED ON PERCEPTUAL WAVELET PACKET TRANSFORM AND TEAGER ENERGY OPERATOR

Jhing-Fa WANG

Department of Electrical Engineering,
National Cheng Kung University, Tainan
wangjf@server2.iie.ncku.edu.tw

Shi-Huang CHEN

Department of Electrical Engineering,
National Cheng Kung University, Tainan
shchen@cad.ee.ncku.edu.tw

ABSTRACT

This paper presents a new voice activity detection (VAD) algorithm based on the perceptual wavelet packet transform (PWPT) and the Teager energy operator (TEO). The basic procedure of the proposed VAD algorithm is to make use of the PWPT to decompose the input speech into critical subband signals. Then a parameter called voice activity shape (VAS) can be derived from the TEO of these critical subband signals. It is shown in this paper that the VAS can be used as a robust feature for VAD. The advantage of this new algorithm is that the preset threshold values or a priori knowledge of the SNR usually needed in conventional VAD methods can be completely avoided. Various experimental results show that the proposed VAD algorithm is capable of outperforming to the ITU-T G.729B VAD and can operate reliably in real noisy environments.

1. INTRODUCTION

The voice activity detection (VAD) is used to distinguish speech from noise and is required in a variety of speech processing systems. For example, in the GSM-based communication system, a VAD module [1] can save battery power by discontinuing transmission when no voice activity is detected. Also, a VAD algorithm can be used in a variable bit rate speech coding system [2] in order to control the average bit rate and the overall coding quality of speech. Various types of different approaches to VAD have been proposed. Most of the conventional VAD algorithms are accomplished by applying several parameters extracted from the input speech signal to compare with predetermined thresholds. If the measured parameters exceed the thresholds, then a voice-active decision is made. In the earlier VAD algorithms [3-4], the parameters used for speech detection are based on autocorrelation coefficients, LPC distance measure, short-time energy levels, zero-crossing rates, and pitch period. Cepstral features and the line spectral frequencies (LSF) are some of the more recent developments used in VAD algorithms [5-6].

These current VAD algorithms provide a satisfactory performance in quite situations, however, the correct detection rate is of little value under noisy environments. The decision parameters used in these conventional VAD algorithms are based on averages over windows of fixed length or on methods using a uniform time-frequency resolution [3-6]. It is well known that speech signals are non-stationary and contain many transient components. For this reason, using a uniform time-frequency resolution method to extract parameters for VAD is not suitable and is not appropriate for noisy environments. In addition, a consistent accuracy cannot be achieved since most VAD algorithms rely on fixed threshold values for comparison. In conventional VAD algorithms, the threshold values are calculated in the silent conditions and are improper for noisy conditions. Therefore, a robust VAD should utilize time-varying threshold values to obtain a better performance.

In order to overcome the above two problems, this paper presents a new VAD algorithm using the perceptual wavelet packet transform (PWPT) and the Teager energy operator (TEO). The PWPT and the TEO are applied to provide a variable time-frequency resolution analysis and a time-varying threshold value, respectively. The PWPT is designed to match the human psychoacoustic model and can improve the performance of various wavelet-based speech processing systems, such as speech enhancement [7] and speech coding [8]. Furthermore, the TEO is a powerful nonlinear operator and is capable of extracting the signal energy based on mechanical and physical considerations. It is shown that the TEO has promising results in a number of speech applications [9-10]. By the use of these two improved techniques, i.e. PWPT and TEO, the proposed algorithm can develop a robust parameter called voice activity shape (VAS) for VAD. Since the magnitude of the VAS is time adapted in function of speech components, a robust VAD algorithm can be constructed by tracing the magnitude of VAS. Furthermore, the preset threshold values or *a priori* knowledge of the SNR usually needed in conventional VAD methods can be completely avoided in this new algorithm. Using speech signals corrupted by additive and real noises, experimental results show that the proposed VAD algorithm obtains a better performance than that of G.729B [6].

2. PWPT AND TEO FOR VAD ALGORITHM

2.1 Perceptual Wavelet Packet Transform

As mentioned in [7-8], the decomposition tree structure of the PWPT is designed to approximate the critical bands as close as possible. It has been shown that frequency components of sound can be integrated into critical bands that refer to bandwidths at

which subjective response become significantly different [11]. One class of critical band scales is called the Bark scale. Based on the measurements by Zwicker *et al*. [12], the Bark scale z can be approximately expressed in terms of the linear frequency by

$$z(f) = 13\arctan(7.6\times10^{-4} f) + 3.5\arctan(1.33\times10^{-4} f)^2 \quad \text{[Bark]} \tag{1}$$

where f is the linear frequency in Hertz. The corresponding critical bandwidth (CBW) of the center frequencies can be expressed by

$$\text{CBW}(f_c) = 25 + 75(1 + 1.4\times10^{-6} f_c^{\,2})^{0.69} \text{ [Hz]} \tag{2}$$

where f_c is the center frequency (unit: Hertz). Theoretically, the human auditory frequency range spreads from 20 to 20000 Hz and covers approximately 25 Barks.

Table 1. The characteristics of critical bands under 4 kHz

No.	f_c (Hz)	CBW	No.	f_c (Hz)	CBW
1	50	-	10	1170	190
2	150	100	11	1370	210
3	250	100	12	1600	240
4	350	100	13	1850	280
5	450	110	14	2150	320
6	570	120	15	2500	380
7	700	140	16	2900	450
8	840	150	17	3400	550
9	1000	160			

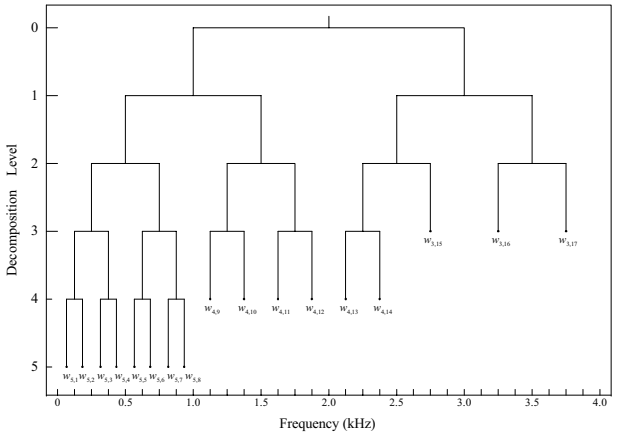

Fig. 1. The tree structure of the PWPT.

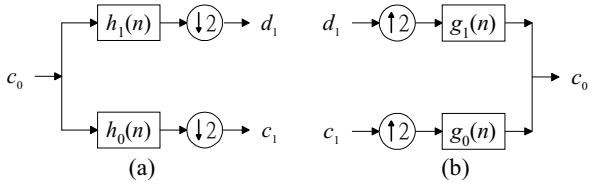

Fig. 2. (a) The basic wavelet decomposition cell and (b) reconstruction cell.

In this paper, the underlying sampling rate is set to be 8 kHz, yielding a bandwidth of 4 kHz. Within this bandwidth, there are approximately 17 critical bands as listed in Table 1 [11]. Fig. 1 shows the corresponding PWPT decomposition tree which contains 16 basic wavelet decomposition cell and five decomposition levels. Each decomposition cell can be implemented via the filter band approach proposed by Mallat [13]. Figs. 2(a) and 2(b) show the basic wavelet decomposition and reconstruction cells. In Fig. 2, $h_0(n)$ and $h_1(n)$ are the analysis low-pass scaling filter and the high-pass wavelet filter, respectively, whereas $g_0(n)$ and $g_1(n)$ are the synthesis low-pass scaling filter and the high-pass wavelet filter, respectively. Also, the symbols $\downarrow 2$ and $\uparrow 2$ shown in Fig. 2 denote the operation of downsampling by 2 and upsampling by 2, respectively. In Fig. 2(a), $\{c_0(n)\}_{n\in Z}$ denotes the input to the basic decomposition cell and the outputs of this cell [14] are given by

$$c_1(k) = \sum_n h_0(n-2k)c_0(n), \tag{3}$$

$$d_1(k) = \sum_n h_1(n-2k)c_0(n) \tag{4}$$

where $c_1(k)$ and $d_1(k)$ are called the approximation coefficients and the detail coefficients of the first level wavelet decomposition of $c_0(n)$, respectively. And its corresponding wavelet reconstruction cell as shown in Fig. 2(b) can be operated as

$$c_0(m) = \sum_k \left[g_0(2k-m)c_1(k) + g_1(2k-m)d_1(k)\right]. \tag{5}$$

By the use of this PWPT, the input speech signal can be transformed into 17 critical wavelet subband signals. The choice of wavelet filter is important for the frequency selectively as well as the time domain resolution of the wavelet filter bank. The filters proposed by Daubechies are the ones that best preserve frequency selectivity as the decomposition level of the PWPT increases. This is due to their regularity property [17]. In addition, computational complexity of a wavelet filter bank is directly depended on the length of wavelet filter. Hence, a 10-point Daubechies wavelet filter [15] is chosen in the proposed VAD algorithm to preserve a sufficient frequency selectivity for each critical wavelet subband signal. Simultaneously, it makes the PWPT quite practical and efficient for the proposed VAD algorithm. Figs. 3 (a) and 3(b) plot the resulting 17-band PWPT of the Bark scale and the CBW, respectively.

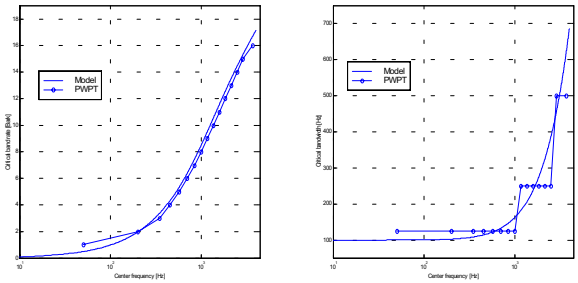

Fig. 3. (a) Bark scale as a function of center frequency, and (b) critical bandwidth as a function of center frequency.

2.2 Teager Energy Operator

It has been shown [9-10] that the TEO is a powerful nonlinear operator which has been used successfully in various speech applications. For a given bandlimited discrete speech signal $y(n)$, the discrete form of the TEO introduced by Kaiser [9] is given by

$$\Psi[y(n)] = y^2(n) - y(n+1)y(n-1) \tag{6}$$

where $\psi[y(n)]$ is called the TEO coefficient of $y(n)$. Note that the TEO can enhance the discriminability of speech components among those of noise [10].

3. PROPOSED VAD ALGORITHM

In the flow chart of the proposed VAD algorithm shown in Fig. 4, the proposed VAD algorithm first computes the PWPT of the input speech signal $x(n)$ and results in 17 critical subband signals, namely $w_{j,m}(k)$ where $3 \le j \le 5$ is the level of the PWPT, $1 \le m \le 17$, and $1 \le k \le 2^j$. Then, from (6), a set of $t_{j,m}(k) = \psi[w_{j,m}(k)]$ can be derived from the TEO of $w_{j,m}(k)$.

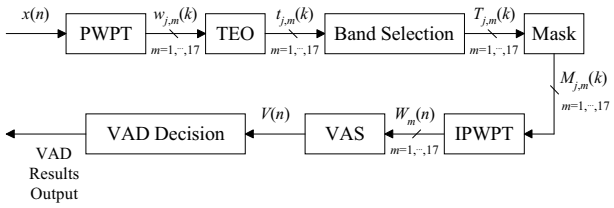

Fig. 4. The flow chart of the proposed VAD algorithm.

3.1 Band Selection

The level-dependent threshold, i.e. $\lambda_j = \sigma_j \sqrt{(2\log(N)}$, proposed by Johnstone and Silverman [16] is embedded for the band selection. That is,

$$T_{j,m}(k) = \begin{cases} t_{j,m}(k), & \text{if } \mathrm{var}\{t_{j,m}(k)\} \ge \lambda_j \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where var{ $t_{j,m}(k)$} denotes the variance of $t_{j,m}(k)$. Here the band selection is used to reject the subband which occurs noise only for VAD.

3.2 Masks Construction

For each band-selected $T_{j,m}(k)$, a mask is obtained by

$$M_{j,m}(k) = T_{j,m}(k) * H_j(k) \tag{8}$$

where $*$ denotes the convolution operation and $H_j(k)$ is a $256/2^j$-point level-dependent Hamming window.

3.3 Calculation of voice activity shape (VAS)

The voice activity shape $V(n)$ is calculated by

$$V(n) = \sum_{m=1}^{17} W_m(n) \tag{9}$$

where $W_m(n)$ is the directly inverse PWPT (IPWPT) of each $M_{j,m}(k)$ given by (8).

3.4 VAD Decision

As mentioned previously, the VAS is time-adapted in function of speech components such that it can be applied to the proposed VAD algorithm. It is observed that the magnitudes of $V(n)$ in voice-active regions are always greater than those in voice-inactive regions. In noiseless conditions, the $V(n)$ of voice-inactive regions is zero and voice-active regions can be easily detected by checking whether $V(n)>0$. However, $V(n)$ exists an offset value β in the voice-inactive regions while the speech is contaminated by background noises. Under this condition, the voice-active regions are detected when $V(n)> \beta$. To determine this offset value, an iteration algorithm proposed by this paper is composed of the following three steps:

1. Initially set $k = 1$ and define $V^{(1)}(n) = V(n)$.
2. Let $V^{(k+1)}(n)$ be defined as
$$V^{(k+1)}(n) = \begin{cases} V^{(k)}(n), & \text{if } V^{(k)}(n) < E\left[V^{(k)}(n)\right] \\ E[V^{(k)}(n)], & \text{otherwise} \end{cases} \tag{10}$$
where $E\left[V^{(k)}(n)\right]$ is the expected value of $V^{(k)}(n)$.
3. If $E\left[V^{(k)}(n)\right] = E\left[V^{(k+1)}(n)\right]$, compute the offset value $\beta = 1.5 \times E\left[V^{(k)}(n)\right]$. Otherwise, set $k = k+1$ and return to step 2.

4. EXPERIMENTAL RESILTS

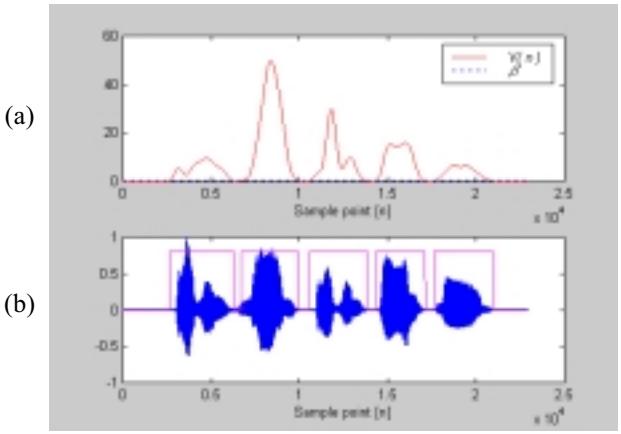

Fig. 5. (a) The VAS (solid line) and the offset value (dashed line) of the clean speech signal "seven-nine-seven-o-nine" shown in (b), (b) the waveform of the clean speech signal and its VAD result.

In this paper, the probabilities of speech correct P_d and false-alarm P_f for a number of noisy speech signals are utilized to evaluate the performance of the proposed VAD algorithm. All of these tested speech signals are selected from the "Aurora 2" database corrupted by additive white Gaussian noises and real noises. An illustrative example of the proposed VAD algorithm with a clean speech signal is shown in Fig. 5. And Fig. 6 shows

the VAD result of the proposed algorithm on the noisy speech signal (SNR = 0dB) of Fig. 5.

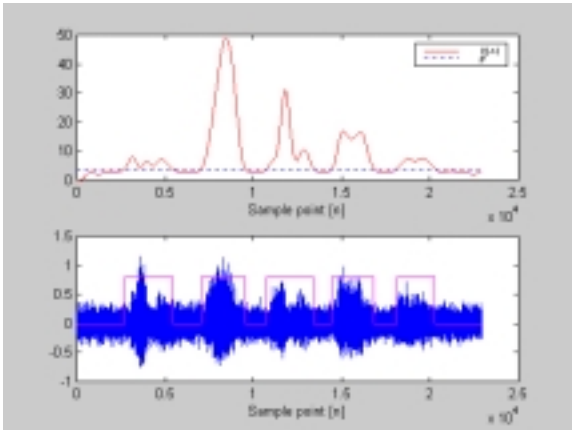

Fig. 6. (a) the VAS (solid line) and the offset value (dashed line) of the noisy speech signal "seven-nine-seven-o-nine" (white noise, SNR = 0dB) shown in (b), and (b) the waveform of the noisy speech signal and its VAD result.

Table 2. P_d's (%) and P_f's (%) of the proposed, wavelet-based*, and G.729B VAD for noisy environments

Method / Noisy Environments		Proposed VAD		Wavelet-based VAD*		G.729B VAD	
		P_d	P_f	P_d	P_f	P_d	P_f
Noise	SNR	Speech	Noise	Speech	Noise	Speech	Noise
White	25 dB	99.8	3.2	99.8	3.2	97.7	2.9
	15 dB	97.3	1.6	96.4	1.8	86.4	1.8
	10 dB	88.5	1.3	86.4	1.3	75.8	1.6
	5 dB	86.4	1.1	82.1	1.2	63.4	1.3
	0 dB	83.6	1.0	78.3	1.2	48.9	1.0
Street	25 dB	99.6	11.8	98.9	12.5	99.8	14.3
	15 dB	98.3	12.5	96.8	12.7	95.2	19.1
	10 dB	95.5	12.7	91.4	12.8	81.9	21.3
	5 dB	92.6	13.2	85.6	14.1	77.1	23.5
	0 dB	90.8	14.2	81.1	14.4	70.8	28.7
Car	25 dB	99.8	6.9	99.3	7.8	98.6	6.4
	15 dB	99.4	8.3	94.6	8.4	92.3	11.3
	10 dB	98.1	9.8	88.4	10.1	84.5	12.6
	5 dB	97.2	9.9	82.9	11.2	82.4	13.4
	0 dB	92.4	10.2	79.1	12.1	78.2	15.9

∗ The algorithm of this wavelet-based VAD is the same as the proposed VAD except the use of conventional 4-stage wavelet packet transform.

To obtain P_d and P_f, the active and inactive regions to the clean speech signals are first marked manually. Then, P_d is defined as the ratio of the correct speech decisions to the hand-marked speech regions, while P_f is defined as the ratio of the false speech decisions to the hand-marked noise regions. Under a variety of noise sources and levels, the P_d and the P_f of the proposed algorithm are compared with those of the VAD specified in the ITU standard G.729B and of the VAD using the conventional wavelet packet transform. The above experimental results are summarized in Table 2. From Table 2, one observes that the proposed VAD has superior performance to the G.729B and the conventional wavelet-based algorithm.

5. CONCLUSIONS

A new approach to the problem of VAD using the PWPT and the TEO has been presented in this paper. By the use of the PWPT and the TEO, a robust parameter called VAS which is time-adapted to the speech waveform is developed. Then, such a VAS can be utilized to achieve a robust VAD. The advantage of the proposed algorithm over the conventional VAD is that it does not require the preset threshold values or *a priori* knowledge of the SNR. Various experimental results show that the proposed VAD algorithm obtains a better performance than that of the G.729B. In the future work, the proposed VAD algorithm will be used with speech coding systems to increase their encoding performance under various types of background noises.

6. REFERENCES

[1] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan European digital cellular mobile telephone service," *in Proc. ICASSP'89*, May 1989, pp. 369-372.

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

[3] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," *in Proc. ICASSP'77*, May 1977, pp. 323-326.

[4] J. C. Junqua, B. Reaves, and B. Mak, " A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," *in Proc. Eurospeech'91*, 1991, pp. 1371-1374.

[5] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," *in Proc. IEEE TENCON*, 1993, pp. 321-324.

[6] ITU-T Rec. G.729, Annex B, A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70.

[7] I. Pinter, "Perceptual wavelet-representation of speech signals and its application to speech enhancement," *Computer Speech and Language*, 10(1), pp. 1–22, 1996.

[8] P. Srinivasan and L. H. Jamieson, "High quality audio compression using an adaptive wavelet decomposition and psychoacoustic modeling," *IEEE Trans. Signal Processing*, 46(4), pp. 1085 –1093, April 1998.

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

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

[11] L. Rabiner and B. H. Juang, *Fundamental of speech recognition*, Upper Saddle River, NJ: Prentice-Hall, 1993.

[12] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *JASA*, vol. 68, pp. 1523-1525, 1980.

[13] S. Mallat, "Multifrequency channel decomposition of images and wavelet model," I*EEE Trans. Acoustic, Speech and Signal Processing*, vol. 37, pp. 2091-2110, 1989.

[14] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms, A primer*, Upper Saddle River, NJ: Prentice-Hall, 1998.

[15] I. Daubechies, *Ten lectures on wavelets*, CBMS, SIAM publ., 1992.

[16] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Roy. Statist.* Soc. B, vol. 59, pp. 319–351, 1997.

[17] D. Sinha and A. Tewfit, "Low bit rate transparent audio compression using adapted wavelet", IEEE Trans. On Signal Processing, vol. 41, pp. 1170-1183, Dec. 1993.