# A Wavelet-Based Voice Activity Detection Algorithm in Variable-Level Noise Environment

Kun-Ching Wang
Department of Information Technology & Communication
Shin Chien University
No. 200, University Rd, Neimen Shiang, Kaohsiung 845
Taiwan
kunching@mail.kh.usc.edu.tw

*Abstract:* In this paper, a novel entropy-based voice activity detection (VAD) algorithm is presented in variable-level noise environment. Since the frequency energy of different types of noise focuses on different frequency subband, the effect of corrupted noise on each frequency subband is different. It is found that the seriously obscured frequency subbands have little word signal information left, and are harmful for detecting voice activity segment (VAS). First, we use bark-scale wavelet decomposition (BSWD) to split the input speech into 24 critical subbands. In order to discard the seriously corrupted frequency subband, a method of adaptive frequency subband extraction (AFSE) is then applied to only use the frequency subband. Next, we propose a measure of entropy defined on the spectrum domain of selected frequency subband to form a robust voice feature parameter. In addition, unvoiced is usually eliminated. An unvoiced detection is also integrated into the system to improve the intelligibility of voice. Experimental results show that the performance of this algorithm is superior to the G729B and other entropy-based VAD especially for variable-level background noise.

*Key-Words:* Voice activity detection, Bark-scale wavelet decomposition, Adaptive frequency subband extraction

## 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, hands-free telephony, audio conferencing and echo cancellation [1]. In the GSM-based wireless system, for instance, a VAD module [2] is used for discontinuous transmission to save battery power. Similarly, a VAD device is used in any variable bit rate codec [3] to control the average bit rate and the overall coding quality of speech. In wireless systems based on code division multiple access, this scheme is important for enhancing the system capacity by minimizing interference. Common VAD algorithms use short-term energy, zero-crossing rate and LPC coefficients [4] as feature parameters for detecting voice activity segment (VAS). Cepstral features [5], formant shape [6], and least-square periodicity measure [7] are some of the more recent metrics used in VAD designs. In the recently proposed G.729B VAD [8], a set of metrics including line spectral frequencies (LSF), low band energy, zero-crossing rate and full-band energy is used along with heuristically determined regions and boundaries to make a VAD decision for each 10 ms frame.

In this paper we present a robust VAD algorithm for the detection of speech segment, which is based on the entropy of the spectrum domain of selected critical subband. First, the bark-scale wavelet decomposition (BSWD) is utilized to decompose the input speech signal into 24 critical subband signals. In contrast to the conventional wavelet packet decomposition, the BSWPD is designed to match the auditory critical bands as close as possible and has been applied into various speech processing systems [9]-[10]. The entropy, on the other hand, a measure of amount of expected information, is broadly used in the field of coding theory. Shen *et al.* [11] first used it on speech detection and revealed that voiced spectral entropy is quite different from non-voiced one. Based on this character, the entropy-based approach is more reliable than pure energy-based methods in some cases, particularly when noise-level varies with time.

Since the frequency energy of different types of noise focus on different frequency subbands, the effect of corrupted noise on each frequency subband is different [12]. The seriously obscured frequency

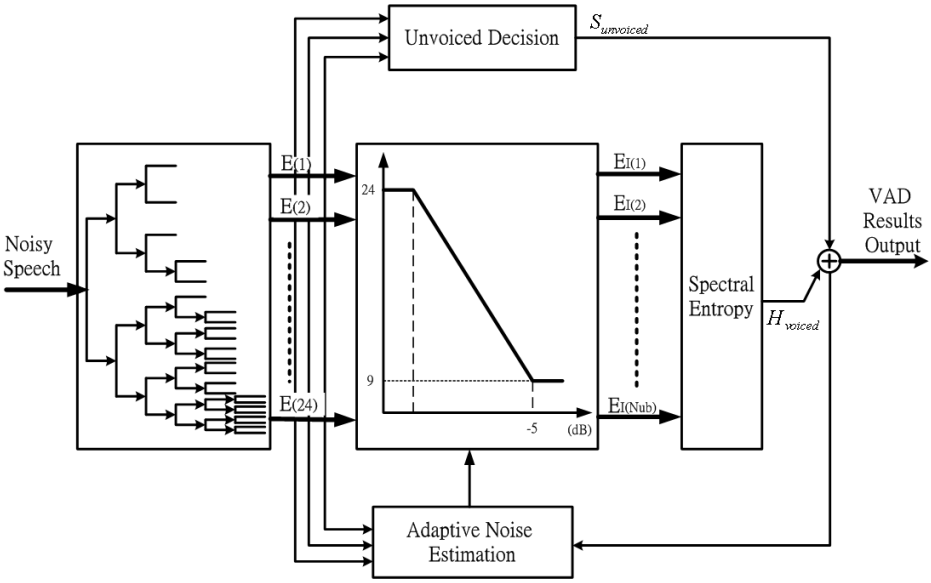


Figure 1. The block diagram of proposed VAD algorithm

subbands have little word signal information left, and are harmful for detecting VAS. Based on the finds, we adopt the theory of adaptive frequency subband extraction (AFSE) to only uses the frequency subband which are slightest corrupted and discard the seriously obscured ones. The frequency subband energies are sorted and only the first several frequency subband with the highest energy are selected. Experiment results show that when more frequency subbands are corrupted by noise, the number of the selected frequency subbands decreases with the decrease of the SNR. A measure of entropy defined on the spectrum domain of selected frequency subband by the AFSE approach is proposed to refine the classical entropy-based VAD [12]. Finally, an unvoiced detection is integrated into entropy-based VAD system to improve the intelligibility of voice.

# 2 Implementation of the proposed VAD algorithm

In the block diagram shown in Fig. 1, the proposed VAD algorithm consists of five main parts: bark-scale wavelet decomposition, adaptive frequency subband extraction, calculation of spectral entropy, adaptive noise estimation, and unvoiced decision. In this section, the five main parts are described in turn.

## 2.1 Bark-scale wavelet decomposition

Critical subband is widely used in perceptual auditory modeling [13]. In this section, we propose the wavelet tree structure of BSWD to mimic the time-frequency analysis of the critical subbands according to the hearing characteristics of human cochlea. A BSWD is used to decompose the speech signal into 24 critical wavelet subband signals, and it is implemented with an efficient five-level tree structure. The corresponding BSWD decomposition tree can be constructed as shown in Fig. 2. Observing the Fig.2, the input speech signal is obtained by using the high-pass filter and low-pass filter [14], implemented with the Daubechies family wavelet, where the symbol $\downarrow 2$ denotes an operator of downsampling by 2.

## 2.2 Adaptive frequency subband extraction

In fact, the frequency energies of difference types of noise are concentrated on different frequency subbands. This observation demonstrates that not all the frequency subbands have harmful word signal information. In our algorithm, we must use only the useful frequency subbands or discard the harmful subbands for detecting VAS. Since our goal is to select some useful frequency subbands having the maximum word signal information, we need a parameter to stand for the amount of word signal information of each frequency subband. According to Wu *et al.* [12], the estimated pure speech signal is a good indicator. The frequency subbands energy of pure speech signal is accomplished by removing the frequency energy of background noise from the frequency energy of input noisy speech.

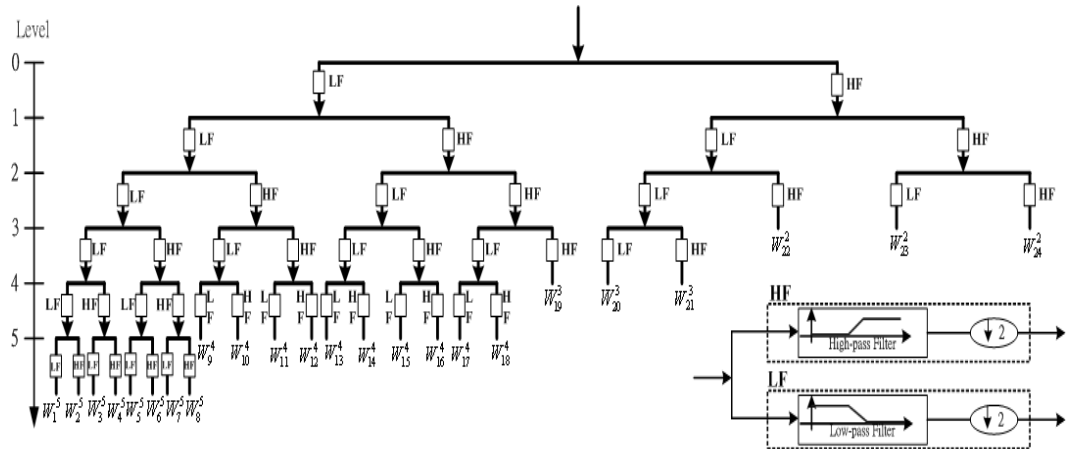


Figure 2. The Bark-scale wavelet decomposition (BSWD) tree

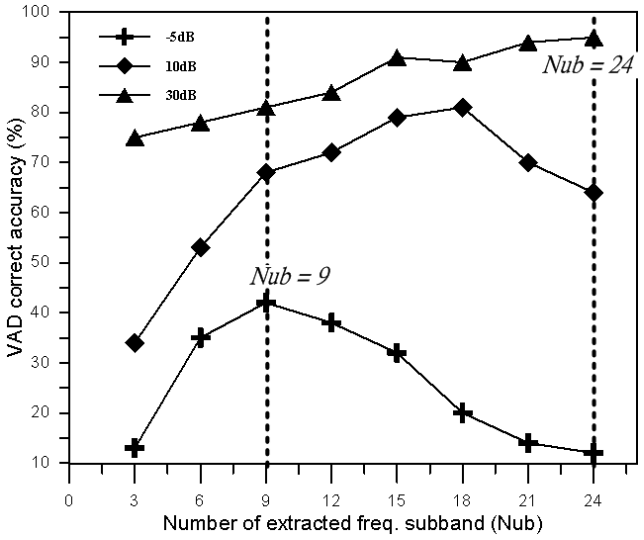


Figure 3. The results of correct detection accuracy with number of different frequency subband at –5dB, 10 dB and 30 dB under three types of noise

For the $m$th frame, the spectral energy of the $\xi$th subband is evaluated by the sum of squares:

$$E(\xi,m)=\sum_{\omega_{\xi,l}}^{\omega_{\xi,h}}\left|X(\omega,m)\right|^2, \quad (1)$$

where $X(\omega,m)$ means the $\omega$th wavelet coeffience. $\omega_{\xi,l}$ and $\omega_{\xi,h}$ denote the lower boundaries and the upper boundaries of the $\xi$th subband, respectively. The $\xi$th frequency subbands energy of pure speech signal of the $m$th frame $\tilde{E}(\xi,m)$ is estimated:

$$\tilde{E}(\xi,m)=E(\xi,m)-\tilde{N}(\xi,m), \quad (2)$$

where $\tilde{N}(\xi,m)$ is the noise power of the $\xi$th frequency subband.

During the initialization period, the noisy signal is assumed to be noise-only and the noise spectrum is estimated by averaging the initial 10 frames. To recursively estimate the noise power spectrum, the subband noise power, $\tilde{N}(\xi,m)$, can be adaptively estimated by smoothing filtering and be discussed later.

It is found that the more the frequency subband covered by noise would result in the smaller the $\tilde{E}(\xi,m)$. Since the frequency subband with higher $\tilde{E}(\xi,m)$ contains more pure speech information, we should sort the frequency subband according to their $\tilde{E}(\xi,m)$ value.

That is,

$$\tilde{E}(I_1,m)\geq\tilde{E}(I_2,m)\geq\cdots\geq\tilde{E}(I_N,m), \quad (3)$$

where $I_i$ is the index of the frequency subband with the $i$th max energy.

It means that the index of the frequency subband with higher energy is the more useful index of one. Moreover, we should only select the useful frequency subbands for VAD results output. That is, the first $N$ frequency subbands $I_1,I_2,\ldots,I_N$ are selected and denoted as the useful number of frequency subband, $N_{ub}$, for the succeeding calculation of spectral entropy.

According to the relation between the number of useful frequency subbands $N_{ub}$ and $SNR$ (shown as Fig. 3), we can see that the number of useful frequency subband increases with the increase of $SNR$ under three types noises including white noise, factory noise and vehicle noise. $N_{ub}=9$ and $N_{ub}=24$ denote the boundary of among the range from -5dB to 30dB, respectively. Based on the above finds, a linear function can be used to

simulate the relationship between $N_{ub}$ and $SNR$, and shown as Fig. 4.

$$N_{ub}(m) = \begin{cases} 9 & ,SNR(m) < -5dB \\ [(24-9)\times\dfrac{(SNR(m)-(-5))}{30-(-5)}+9] & ,\text{-}5dB \le SNR(m) \le 30dB \\ 24 & ,SNR(m) > 30dB. \end{cases} \quad (4)$$

where $[\ ]$ is the round off operator and $SNR(m)$ denotes a frame-based posterior SNR for the $m$th frame.

In addition, $SNR(m)$ is depended on the all summation of subbnad-based posterior SNR $snr(\xi,m)$ on the $\xi$th useful subband and defined as:

$$SNR(m) = 10\log_{10}\sum_{\xi\in N_{ub}} snr(\xi,m), \quad \text{where} \quad snr(\xi,m) = \frac{|X(\xi,m)|^2}{\tilde{N}(\xi,m)}. \quad (5)$$

## 2.3 Calculation of spectral entropy

To calculate the spectral entropy, the probability density function (pdf) and the entropy calculation are both necessary steps.

The pdf for the spectrum can be estimated by normalized the frequency componemts:

$$P(\xi,m) = E(\xi,m)\Big/\sum_{\omega=1}^{N} E(\omega,m) \quad (6)$$

where $P(\xi,m)$ is the corresponding probability density, and $N$ denotes the total number of critical subbnad divided by BSWD ( $N=24$ in this paper).

Some frequency subbands, however, are corrupted seriously by additive noise, and those harmful subbands may result in low performance of entropy-based VAD if those are extracted. Moreover, we use only the useful frequency subbands to calculate a measure of entropy defined on the spectrum domain of selected frequency subbands. The probability associated with subband energy modified from (6) is described as follows:

$$P(\xi,m) = E(\xi,m)\Big/\sum_{\omega=1}^{N_{ub}} E(\omega,m), \quad (7)$$

where $N_{ub}$ is the number of useful frequency subbands.

Having finishing applying the above constraints, the spectral entropy $H(m)$ of frame $m$ can be defined below.

$$H(m) = -\sum_{\xi=1}^{N_{ub}} P(\xi,m)\cdot\log[P(\xi,m)]. \quad (8)$$

The foregoing calculation of the spectral entropy parameter implies that the spectral entropy depends only on the variation of the spectral energy but not on the amount of spectral energy. Consequently, the spectral entropy parameter is robust against changing level of noise.

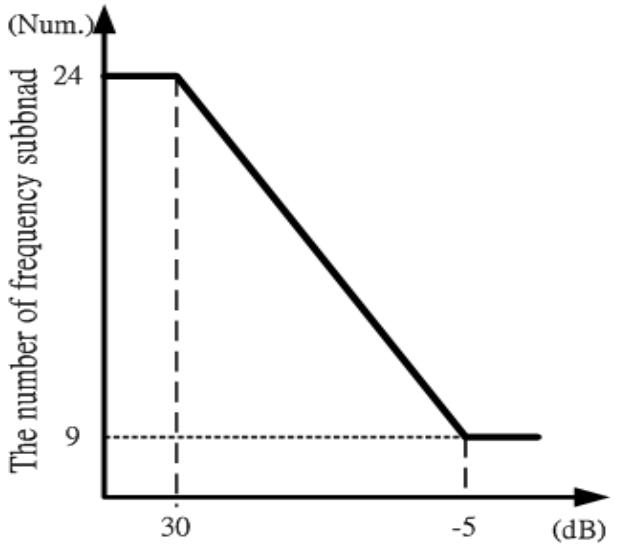


Figure 4. A linear function of the relationship between $N_{ub}$ and $SNR$

## 2.4 Adaptive noise estimation

To recursively estimate the noise power spectrum, the spectral power of subband noise can be estimated by averaging past spectral power values using a time and frequency dependent smoothing parameter as following:

$$\tilde{N}(\xi,m) = \alpha(\xi,m)\cdot\tilde{N}(\xi,m-1) + (1-\alpha(\xi,m))\cdot E(\xi,m) \quad (9)$$

where $\alpha(\xi,m)$ means the smoothing parameter and be defined as

$$\alpha(\xi,m) = \begin{cases} 1, & \text{if VAD(m-1)=1,} \\ \dfrac{1}{1+e^{-k\cdot(snr(\xi,m)-T)}}, & \text{otherwise.} \end{cases} \quad (10)$$

where $T$ is used for center-offset of the transition curve in Sigmoid.

Observing (10), it is found that the smoothing parameter set one when previous speech-dominated frame, the spectral power of subband noise keep until noise-dominated frame. Otherwise, the smoothing parameter may be chosen as a Sigmoid functions when noise-dominated frame.

## 2.5 Unvoiced decision

More unvoiced information is eliminated from conventional VAD algorithm. In order to overcome this drawback, a method of unvoiced decision is

proposed in this section. According to the structure of BSWD tree (shown as Fig. 2), the three sub-energies corresponding to the wavelet subband signals are defined as

$$E_{L0} = \sum_{j=1}^{8} W_j^5,$$
$$E_{L1} = \sum_{j=9}^{12} W_j^4, \quad (11)$$
$$E_{L2} = \sum_{j=13}^{18} W_j^4 + W_{19}^3.$$

The unvoiced segments are determined as:

$$S_{unvoiced} = \begin{cases} 1 & \text{, if } E_{L2} > E_{L1} > E_{L0} \text{ and } E_{L0}/E_{L2} < 0.99 \\ 0 & \text{, otherwise.} \end{cases} \quad (12)$$

### 2.6 Voice activity segment detection

Finally, the voice activity segment (VAS) is derived as:

$$VAS = \{H_{voiced} = 1\} \cup \{S_{unvoiced} = 1\}. \quad (13)$$

where $H_{voiced}$ is the pre-defined $H$ .

## 3 Experimental Results

The speech database contained 60 speech phrases (in Mandarin and in English) spoken by 35 native speakers (20 males and 15 females), sampled at 4 KHz with 16-bit resolution. To set up the noisy signal for test, we add the prepared noise signals to the recorded speech signal with different SNRs range from - 5dB to 30 dB. The noise signals are all taken from the noise database NOISEX-92 [15]. Of the various noises available on the NOISEX database, white noise, factory noise and vehicle noise are selected as speech containment.

Fig. 5 shows the VAD result of the proposed algorithm on the noisy speech signal "May-I-Help-you" under variable-level of noise. It is founded that the VAS of the proposed algorithm can correctly extract speech segments especially for unvoiced segment /H/ occurred at /Help/ sentence in Fig. 5(b). Conversely, in Fig. 5(c) the VAS of standard G729B performs fail during high variable-level of noise segment and unvoiced segment. In order to compare with other VADs specified in the ITU standard G.729B, we introduce three criteria: 1) the probability of correctly detecting speech frames $P_{cS}$ is the ratio of the correct speech decision to the total number of hand-labeled speech frames. 2) the probability of correctly detecting noise frames $P_{cN}$ is the ratio of the correct noise decision to the total number of hand-labeled noise frames. 3) the false-alarm $P_f$ is the ratio of the false speech decision or false noise decision to the total hand-labeled frames. Under a variety of SNR's, the $P_{cS}$ , $P_{cN}$ and $P_f$ of the proposed algorithm are compared with those of the VAD specified in the ITU standard G.729B [8] and other entropy-based VAD [11]. The experimental results are summarized in Table I. It is shown that. In high SNR, the result of Shen's VAD is comparable to proposed VAD. But, the proposed VAD has superior performance to the Shen's VAD and G.729B particularly in low SNR.

## 4 Conclusion

In this paper, a novel entropy-based VAD algorithm has been presented in non-stationary environment. The algorithm is based on bark-scale wavelet decomposition to decompose the input speech signal into critical sub-band signals. Motivated by the concept of adaptive frequency subband extraction, we use the frequency subband that are slightest corrupted and discard the seriously obscured ones. It is found that the proposed algorithm improves the classic entropy-based approach.

Experimental results show that the performance of this algorithm is superior to the G.729B and other entropy-based approach in low SNR. The proposed algorithm has excellent presentation especially for variable-level background noise.

## 5 Acknowledgment

This work was supported by National Science Council of Taiwan under grant no. NSC97-2218-E-158-003.

*References:*

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] 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. Int. Conf. Acoustics, Speech, Signal Processing*, 1989, pp. 369-372.

[3] *Enhanced variable rate codec*, speech service option 3 for wideband spread spectrum digital systems, TIA doc. PN-3292, Jan. 1996.

[4] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in *Proc. Int. Conf.*

*Acoustics, Speech, Signal Processing*, May 1977, pp. 323-326.

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

[6] J. D. Hoyt and H. Wechsler, Detection of human speech in structured noise, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1994, pp. 237-240.

[7] R. Tucker, Voice activity detection using a periodicity measure, *Proc. Inst. Elect. Eng.*, Vol. 139, No. 4, 1992, pp. 377-380.

[8] A. Benyassine, E. Shlomot, and H. Su, 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 spplications, *IEEE Commun. Mag.*, 1997, pp. 64-72.

[9] I. Pinter, Perceptual wavelet-representation of speech signals and its application to speech enhancement, *computer speech and language*, Vol.10, No.1, 1996, pp. 1-22.

[10] P. Srinivasan and L. H. Jamieson, High quality audio compression using an adaptive wavelet decomposition and psychoacoustic modeling, *IEEE Trans. Signal Processing*, Vol.46, No.4, 1998, pp. 1085-1093.

[11] J. L. Shen, J. W. Hung, and L. S. Lee, Robust entropy-based endpoint detection for speech recognition in noisy environments, *International Conference on Spoken Language Processing*, Sydney, 1998, pp. 232–235.

[12] G. D. Wu and C. T. Lin, Word boundary detection with mel-scale frequency bank in noise environment, *IEEE Trans. Speech Audio Process.*, Vol.8, No.3, 2000, pp. 541-554.

[13] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, New York, 1990.

[14] S. Mallat, Multifrequency channel decomposition of images and wavelet model, *IEEE Trans. Acoust. Speech Signal Process.* Vol.37, 1989, pp. 2091-2110.

[15] 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 Commun.*, Vol. 12, 1993, pp. 247-251.

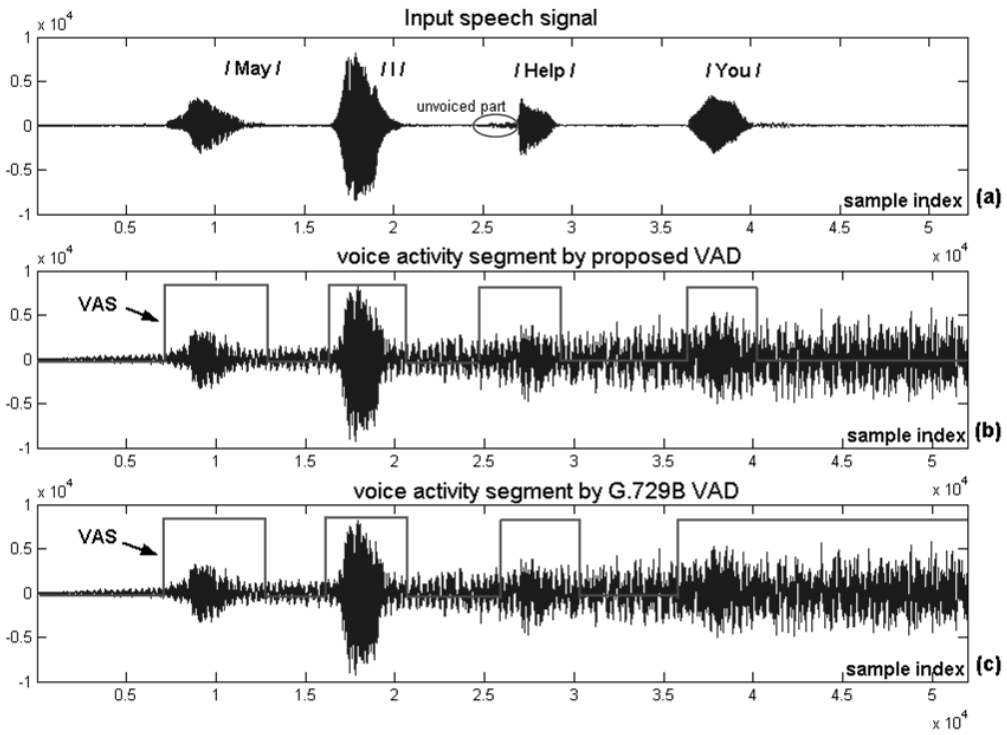


Figure 5. Comparison between the two VADs: (a) Waveform of a clean speech 'May I help you?' (b) The VAS of proposed VAD (c) The VAS of G.729B

Table I.
Performance comparisons for three noise types and levels

| **Noise Conditions** | | $P_{cS}$ **(%)** | | | $P_{cN}$ **(%)** | | | $P_f$ **(%)** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Type** | **SNR(dB)** | **Proposed VAD** | **G.729B** | **Shen *et al.* [11]** | **Proposed VAD** | **G.729B** | **Shen *et al.* [11]** | **Proposed VAD** | **G.729B** | **Shen *et al.* [11]** |
| **White Noise** | ***30*** | 99.8 | 93.1 | 99.1 | 99.2 | 84.6 | 99.8 | 1.5 | 12.9 | 1.6 |
| | ***10*** | 95.6 | 85.2 | 94.6 | 98.7 | 81.5 | 95.4 | 4.6 | 17.3 | 4.9 |
| | ***-5*** | 92.4 | 78.1 | 85.2 | 92.1 | 72.7 | 82.3 | 8.4 | 25.5 | 10.2 |
| **Factory Noise** | ***30*** | 94.6 | 92.9 | 94.3 | 93.1 | 88.9 | 93.0 | 10.2 | 13.6 | 10.8 |
| | ***10*** | 89.7 | 84.3 | 85.1 | 89.7 | 83.3 | 85.1 | 13.2 | 18.4 | 15.7 |
| | ***-5*** | 80.5 | 74.6 | 74.8 | 85.3 | 73.6 | 76.5 | 16.2 | 24.2 | 20.1 |
| **Vehicle Noise** | ***30*** | 96.8 | 95.3 | 96.5 | 94.2 | 92.3 | 93.1 | 6.3 | 14.3 | 6.5 |
| | ***10*** | 92.5 | 90.1 | 91.1 | 89.6 | 84.1 | 85.3 | 9.5 | 17.4 | 12.4 |
| | ***-5*** | 88.4 | 81.4 | 82.7 | 84.1 | 79.4 | 82.4 | 14.7 | 21.5 | 19.6 |