Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments

Bing-Fei Wu, *Senior Member, IEEE,* and Kun-Ching Wang, *Student Member, IEEE*

***Abstract*—In speech processing, endpoint detection in noisy environments is difficult, especially in the presence of nonstationary noise. Robust endpoint detection is one of the most important areas of speech processing. Generally, the feature parameters used for endpoint detection are highly sensitive to the environment. Endpoint detection is severely degraded at low signal-to-noise ratios (SNRs) since those feature parameters cannot adequately describe the characteristics of a speech signal. As a result, this study seeks the banded structure on speech spectrogram to distinguish a speech from a nonspeech, especially in adverse environments. First, this study proposes a feature parameter, called band-partitioning spectral entropy (BSE), which exploits the use of the banded structure on speech spectrogram. A refined adaptive band selection (RABS) method is extended from the adaptive band selection method proposed by Wu *et al.*, which adaptively selects useful bands not corrupted by noise. The successful RABS method is strongly depended on an on-line detection with minimal processing delay. In this paper, the RABS method is combined with the BSE parameter. Finally, a novel robust feature parameter, adaptive band-partitioning spectral entropy (ABSE), is presented to successfully detect endpoints in adverse environments. Experimental results indicate that the ABSE parameter is very effective under various noise conditions with several SNRs. Furthermore, the proposed algorithm outperforms other approaches and is reliable in a real car.**

***Index Terms*—Adaptive processing, endpoint detection, multiband analysis, spectral entropy.**

I. Introduction

ENDPOINT detection is used to distinguish speech from noise and is required in many speech applications, such as speech recognition, speech coding and communication, among others [1], [2]. In a speech recognition system, for example, accurate endpoint detection can improve the recognition ratio under various types of background noise and reduce the computing power waste induced by incorrect speech detection. Accurate endpoint detection is also used during discontinuous transmission to save battery power and to control the average bit rate and the overall coding quality of the speech in a digital communication system [3].

Manuscript received September 9, 2003; revised February 19, 2004. This work was supported by Promoting Academic Excellence of Universities under Contract 91x104, Ex-91-E-FA06-4-4. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Vary.

The authors are with the Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu, Taiwan 30050, R.O.C. (e-mail: bwu@cssp.cn.nctu.edu.tw; kunching@cssp.cn.nctu.edu.tw).

Digital Object Identifier 10.1109/TSA.2005.851909

A feature parameter that can sufficiently specify the characteristics of a speech and be robust in noisy environments is urgent. The current algorithms are based on short-time energy or spectral energy, zero-crossing rate (ZCR), and duration parameters [4]–[6]. All of these parameters, however, are rather sensitive to noise and cannot fully specify the characteristics of a speech signal. For example, the energy-based parameter and ZCR are not sufficient to distinguish a speech from a noise at low SNRs. In particular, the ZCR is very sensitive to various types of noise. Several other parameters have also been proposed, including linear prediction coefficients (LPCs), Cepstral coefficients, and pitch [7]–[9]. Although these parameters are quite effective in expressing the characteristics of a speech signal, the performance of endpoint detection using such parameters remains poor in adverse environments. The reliability of the LPCs has been observed to depend strongly on the noise in adverse environment. Pitch information can help to detect speech; even so, extracting the correct pitch in noisy environments is difficult. Additionally, some algorithms cannot be implemented for practical applications due to their high computational complexity, even though they perform well [10]. Among such approaches, however, Junqua *et al.* [11] proposed a time-frequency (TF) parameter to detect speech, which assumes that frequency information in the frequency ranges 250–3500 Hz is less contaminated by noise. The TF parameter is composed of both frequency energy in the fixed frequency bands and time energy. Based on the motivation that the frequency energies of various types of noise are concentrated in different frequency bands, Wu *et al.* [12] used the multiband technique to analyze noisy speech signals, and then proposed an adaptive band selection (ABS) method to cancel noise effectively by selecting useful bands. An adaptive time-frequency (ATF) parameter extended from TF parameter was proposed by them.

Although the ATF-based algorithm outperforms several algorithms commonly used for endpoint detection in the presence of various types of noise, it cannot be reliably implemented in practical environments. It is found that the selection of useful bands depends on the information of an entire recorded signal. Additionally, the ATF parameter is also energy based, and is, therefore, less reliable in the presence of nonstationary noise or a changing level of noise. Shen *et al.* [13] first used the entropy-based parameter to detect speech signals. Their study indicated that the spectral entropy of a speech segment differed significantly from that of a noise segment. Subsequently, Huang [14] integrated both the time energy and spectral entropy to form a new feature parameter (EE-feature), since the spectral entropy

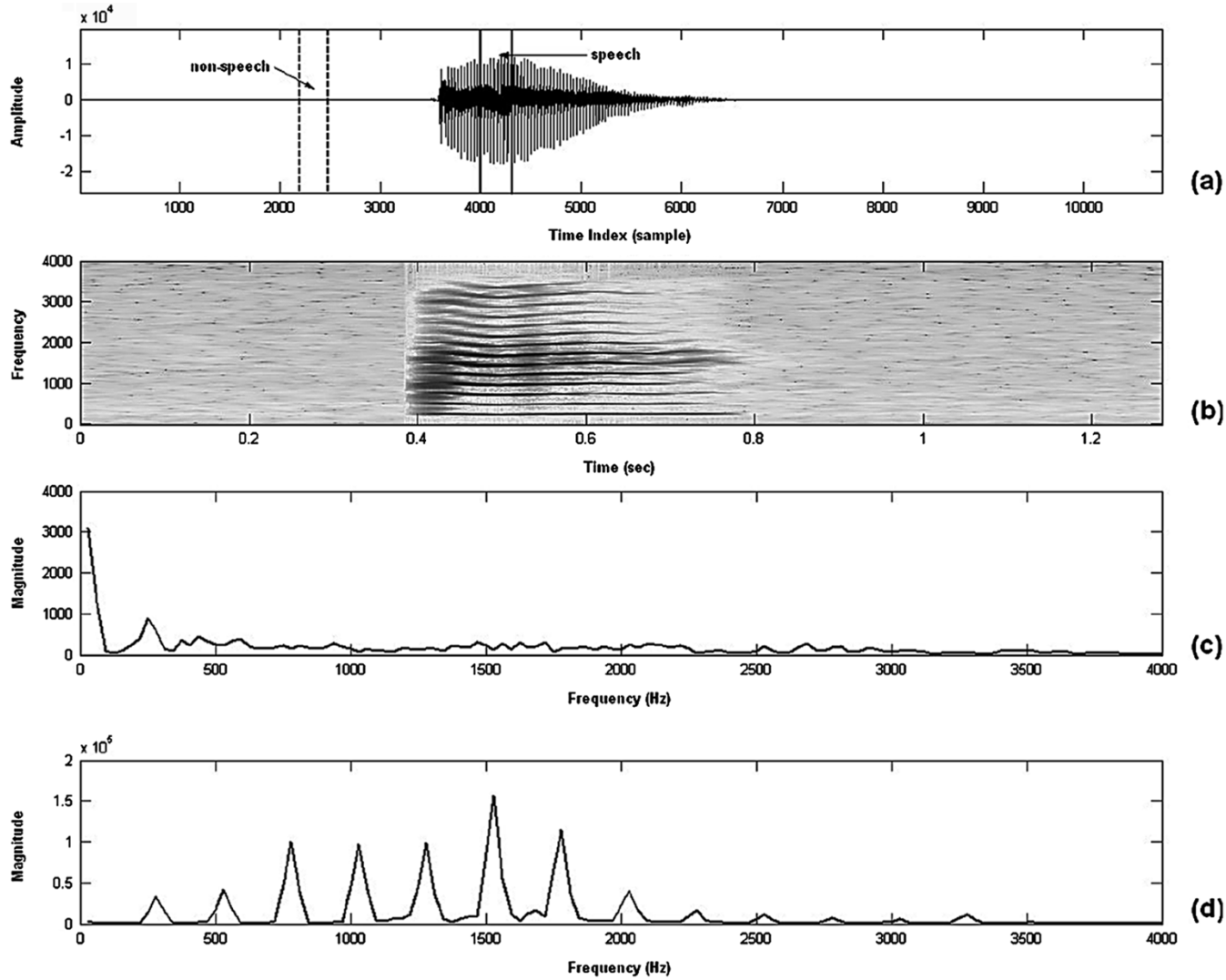

Fig. 1. Banded nature displayed on speech spectrogram: (a) A signal waveform of Mandarin digit "eight." (b) The continuous, striated lines only appearing on a speech spectrogram of the corresponding signal waveform. (c) Spectral magnitude of a speech segment. (d) Spectral magnitude of a nonspeech segment.

failed under multitalker babble and background music, but the energy performed well because of its additive property: The energy of the sum of speech plus noise always exceed the energy of noise. Although the EE-feature parameter proposed by Huang improved the endpoint detection under babble noise, it is unreliable when the noise level greatly exceeds the speech level. Additionally, the spectral entropy parameter relies on the variance of spectral magnitude to distinguish a speech signal from a noise signal, but the variance of spectral magnitude depends strongly on the noisy environments.

The inherent characteristic of the banded structure on speech spectrogram can be modeled by improving the spectral entropy. In this paper, a multiband analysis is combined with the spectral entropy parameter. A single band was divided into 32 uniform bands, since the frequency gap between each peak and valley of the speech spectrum is observed to be around 125 Hz, which the resolution is to be enough to distinguish speech from noise in the spectrum. Multiband analysis not only discards harmful bands corrupted by noise as noise cancellation, but can enhance the banded nature on speech spectrogram. The spectral entropy through multiband analysis is called the band-partitioning spectral entropy (BSE) feature parameter. The BSE feature parameter strengthens the boundary between speech and noise more clearly than the spectral entropy proposed by Shen at poor SNRs. To make the proposed algorithm perform well in real environments, a method of adaptive noise cancellation, which can adaptively select useful bands with time and called refined adaptive band selection (RABS), is presented in this study. A novel robust feature parameter, which combines the BSE parameter with the RABS method, is called the adaptive band-partitioning spectral entropy (ABSE) parameter and presented to detecting speech in adverse conditions.

This paper is organized as follows. Section II will introduce the theory of entropy and the motivation of using the entropy to detect speech, which describes the banded nature on speech spectrogram, and it also presents the deviation of the ABSE parameter. In Section III, an adaptive noise cancellation, called RABS method, which can adaptively select useful bands in on line, is derived in detail, and then the procedure for implementing the proposed ABSE-based endpoint detection algorithm is outlined. Section IV discusses the performance of the proposed algorithm under various noise conditions and

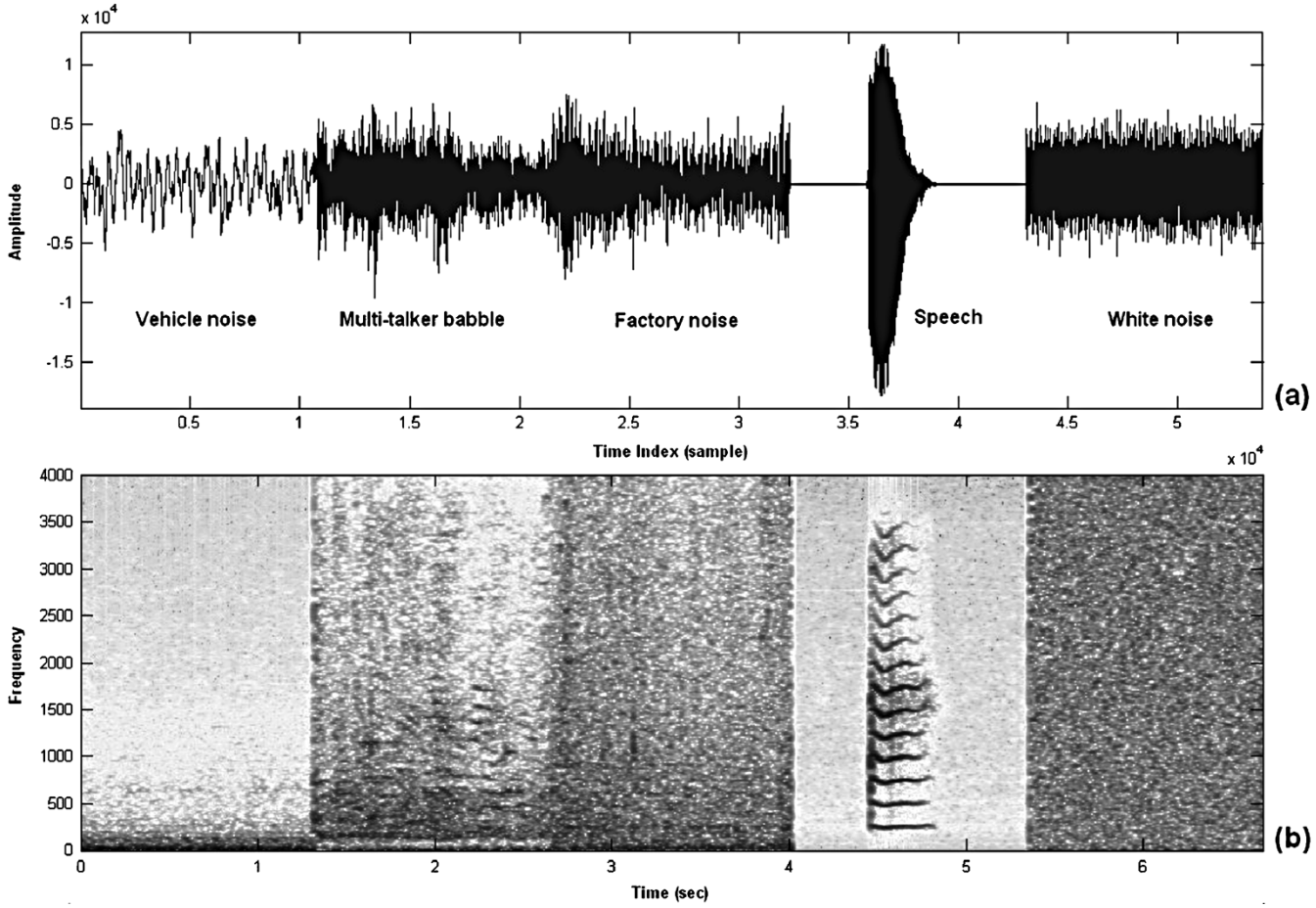

Fig. 2. Inherent characteristic of banded structures only appeared on speech spectrogram. (a) Mixed-signal waveform is composing vehicle noise, multitalker babble noise, factory noise, and speech signal and white noises in turn. (b) Spectrogram of the corresponding mixed signal.

compares its performance with that of other algorithms. Finally, Section V summarizes the findings and discusses possible directions for future work.

II. ABSE Feature Parameter

This section introduces the theory of entropy and then shows the motivation of using the entropy for detecting speech, which describes the banded nature on speech spectrogram. A deviation of the proposed ABSE parameter is also described in this section.

A. Motivation

Entropy, first used in information theory by Shannon [18], is regarded as the amount of information that must be provided about a random signal x in order to specify it uniquely. It measures the degree of organization (uncertainty) of the signal and is defined by

$$H(x) = \sum_{k} P(x_k) \cdot \log[1/P(x_k)] \tag{1}$$

where $x = \{x_k\}_{0 \le k \le N-1}$ and $P(x_k)$ is the probability of x_k.

Fig. 1 describes evidently the difference between the spectrum of a speech segment and that of a nonspeech segment. The waveform of a Mandarin digit eight uttered by native speaker is shown in Fig. 1(a). In speech production, the pitch varies continuously within a speech segment, so the striated lines, shown in Fig. 1(b), are also continuous. This results in a clear set of striated lines throughout speech spectrogram, which are called as a banded nature on speech spectrogram. When such lines exist in some frequency bands for a long enough time, the speech segment can be quite certainly presented [19]. So, it is observed that the banded nature only appears on speech spectrogram. Fig. 1(c) and (d) shows the spectrum magnitude of a speech segment obtained by the short-time Fourier transform (STFT) over a solid-line region in Fig. 1(a) and that of a nonspeech obtained segment by STFT over a dashed-line region in Fig. 1(a), respectively. The variance (uncertainty) of the spectral magnitude of a speech segment over all the frequency components is found to exceed that of a nonspeech segment. Based on the ability of measuring uncertainty, it is found that the entropy can be applied to speech spectrum for specifying a speech signal successfully. Fig. 2(a) displays the waveform of a mixed signal that consists of vehicle noise, multitalker babble noise, factory noise, speech, and white noise in turn. The corresponding spectrogram is shown in Fig. 2(b). The figure adequately illustrates that using spectral entropy to characterize the banded nature which only appear on speech spectrogram can distinguish a speech signal from a noise. Fig. 3 displays the spectrograms of clean speech and noisy speech with four kinds of noise (vehicle noise, factory

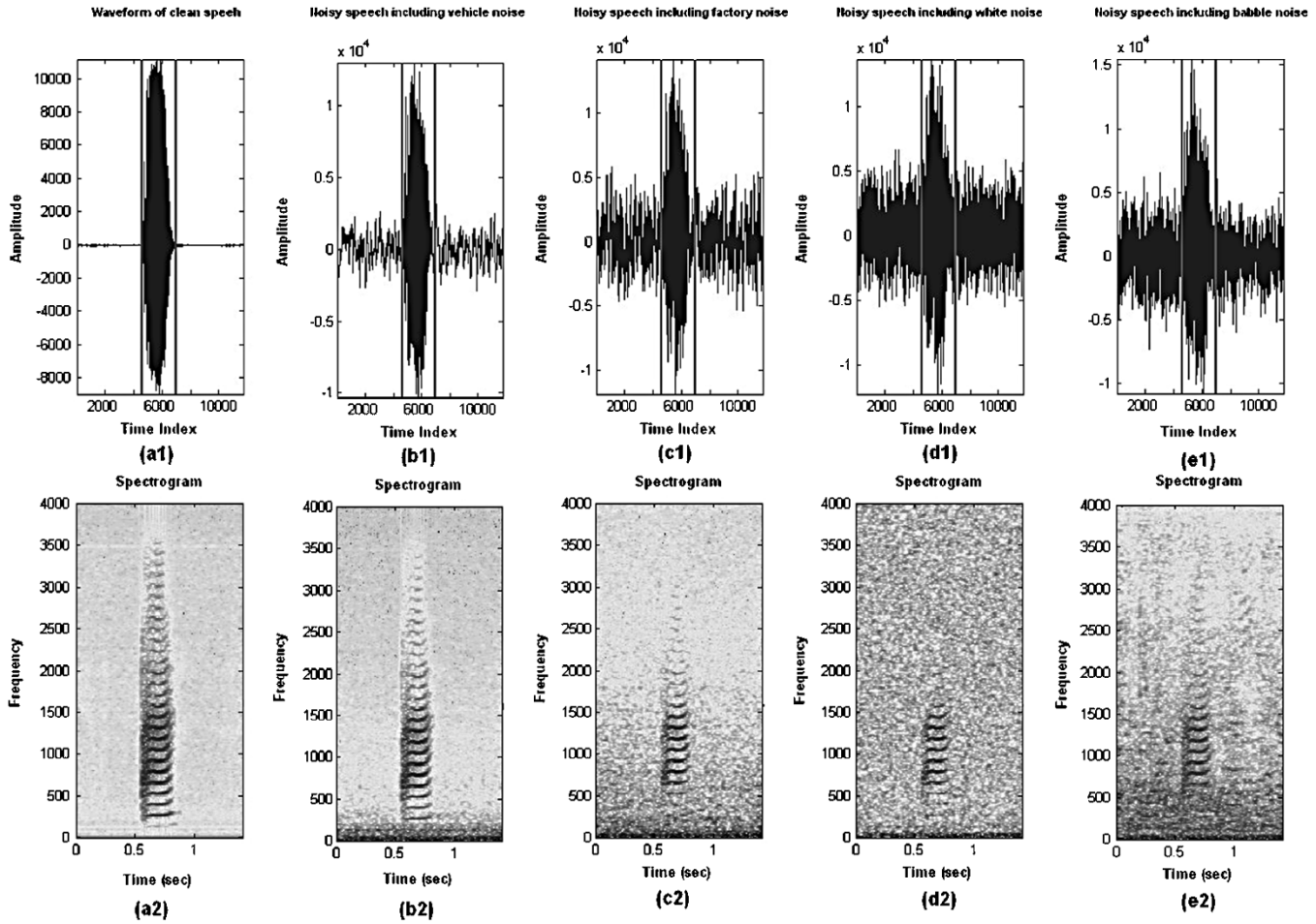

Fig. 3. Illustration of the banded nature against various types of noise.

noise, white noise, and multitalker babble) at 0 dB. It is also observed that the banded nature on speech spectrogram is robust against various types of additive noise.

B. Band-Partitioning Spectral Entropy

Shen *et al.* [13] first used an entropy-based parameter for endpoint detection under adverse conditions. Their experimental results reveal that the spectral entropy of a speech signal differs from that of a nonspeech signal. The procedure for calculating a spectral entropy parameter is described as follows. The STFT of a given time frame $S(n,l)$ is given by

$$X(k,l)=\sum_{n=1}^{N} H(n)\cdot S(n,l)\cdot \exp(-j2\pi kn/N) \qquad 1\le k\le N \quad (2)$$

where $X(k,l)$ represents the spectral magnitude of the kth frequency bin of the lth frame. N is the total number of frequency bins in STFT for each frequency frame ($N=256$ in the proposed system). $H(n)$ is a Hamming window and overlapping size is 128. The spectral energy of each frame, $X_{\text{energy}}(k,l)$, is described as follows:

$$X_{\text{energy}}(k,l)=|X(k,l)|^2, \quad 1\le k\le N/2 \qquad (4)$$

Then, the probability associated with each spectral energy component $P(i,l)$ can be estimated by normalizing

$$P(i,l)=\frac{X_{\text{energy}}(i,l)}{\sum_{k=1}^{N/2} X_{\text{energy}}(k,l)}, \quad 1\le i\le N/2. \qquad (5)$$

Following normalization, the corresponding spectral entropy for a given frame is defined as follows:

$$H(l)=\sum_{i=1}^{N/2} P(i,l)\cdot \log[1/P(i,l)] \qquad (6)$$

where $H(l)$ is the spectral entropy of the lth frame. 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. However, the magnitude associated with each point in the spectrum is easily contaminated by noise and then the performance of endpoint detection would be degraded at seriously low SNRs. This study addresses the multiband analysis of noisy

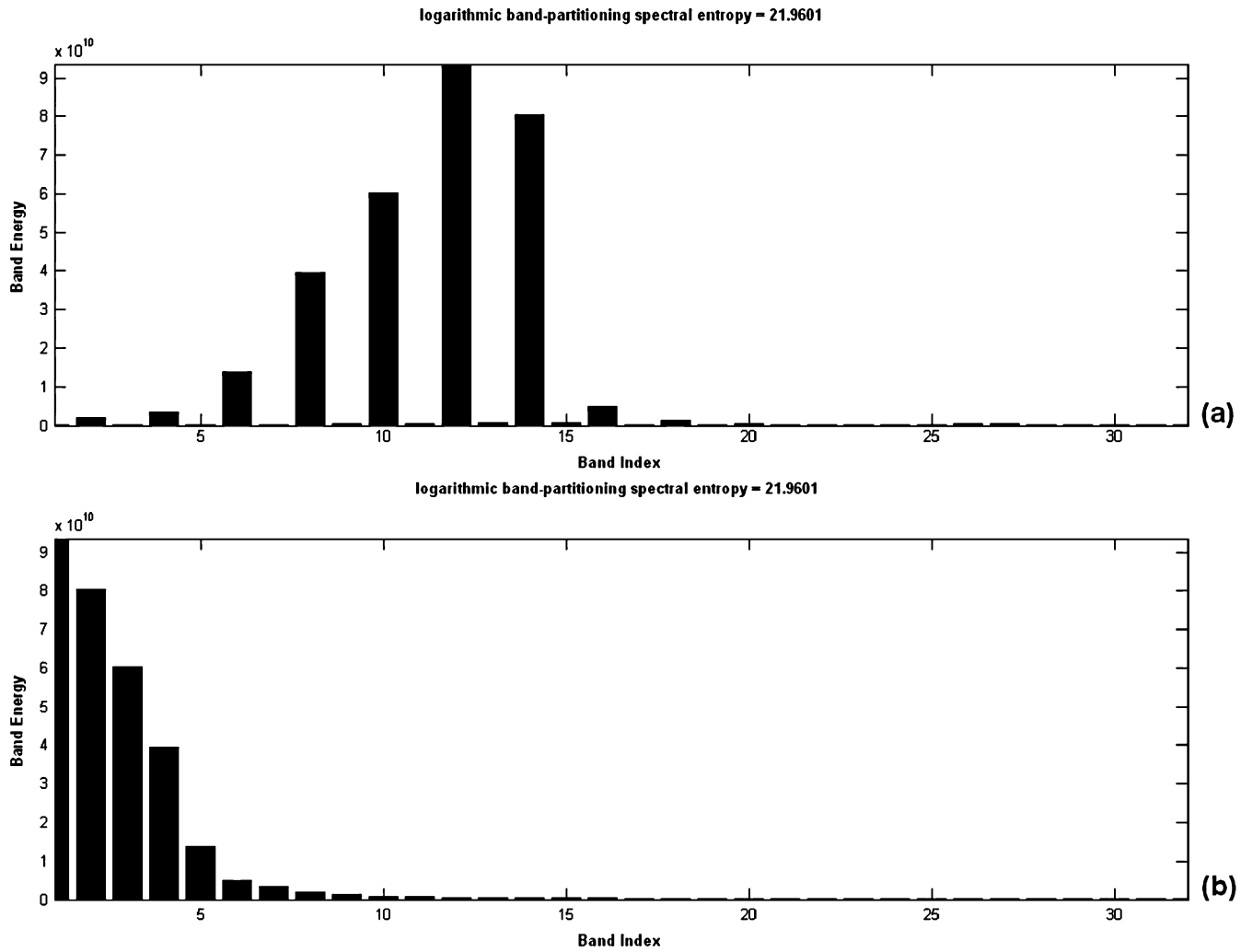

Fig. 4. Different distribution of band energy with the same entropy value (BSE $= 21.9601$). (a) Distribution of band energy during a speech segment. (b) Distribution of band energy during a nonspeech segment.

speech signals to overcome the sensitivity of the spectral magnitude in noisy environments. The band energy of each band for a given frame is described as follows:

$$E_b(m,l) = \sum_{k=1+(m-1)*4}^{1+(m-1)*4+3} X_{\text{energy}}(k,l), \quad 1 \le m \le N_b \quad (7)$$

where N_b is the total band size of each frame ($N_b = 32$), and $E_b(m,l)$ represents the band energy of the mth band. Consequently, the probability associated with band energy modified from (5) is described as follows:

$$P_b(m,l) = \frac{E_b(m,l)}{\sum_{k=1}^{N_b} E_b(k,l)}, \quad 1 \le m \le N_b. \quad (8)$$

The BSE parameter $H_b(l)$ is represented as follows:

$$H_b(l) = \sum_{m=1}^{N_b} P_b(m,l) \cdot \log[1/P_b(m,l)]. \quad (9)$$

In fact, the well-known entropy cannot indicate a distribution (spatial information) of the data sequence. Fig. 4 plots different distributions of band energy with the same entropy. The distribution of band energy during a speech segment is shown in Fig. 4(a), which is similar to Fig. 1(d). The distribution of band energy during a nonspeech segment is presented in Fig. 4(b). The observation is clearly shown that the original entropy does not sufficiently specify the organization of the banded nature on the speech spectrogram. A set of weighting factors $W(m,l)$ can be further employed to compensate for the above drawback and are defined by (10) and (11), shown at the bottom of the page, where $W(m,l)$ indicates the weight of the mth

$$W(m,l) = \text{var}[P_{b\_\text{offset}}(m-1,l), P_{b\_\text{offset}}(m,l), P_{b\_\text{offset}}(m+1,l)] \quad (10)$$

$$\begin{cases} P_{b\_\text{offset}}(m-1,l) = \frac{\min\{P_b(l)\}}{P_b(m-1,l)}, & \text{for all bands of the } l\text{th frame} \\ P_{b\_\text{offset}}(m,l) = \frac{\min\{P_b(l)\}}{P_b(m,l)}, & \text{for all bands of the } l\text{th frame} \\ P_{b\_\text{offset}}(m+1,l) = \frac{\min\{P_b(l)\}}{P_b(m+1,l)}, & \text{for all bands of the } l\text{th frame} \end{cases} \quad (11)$$

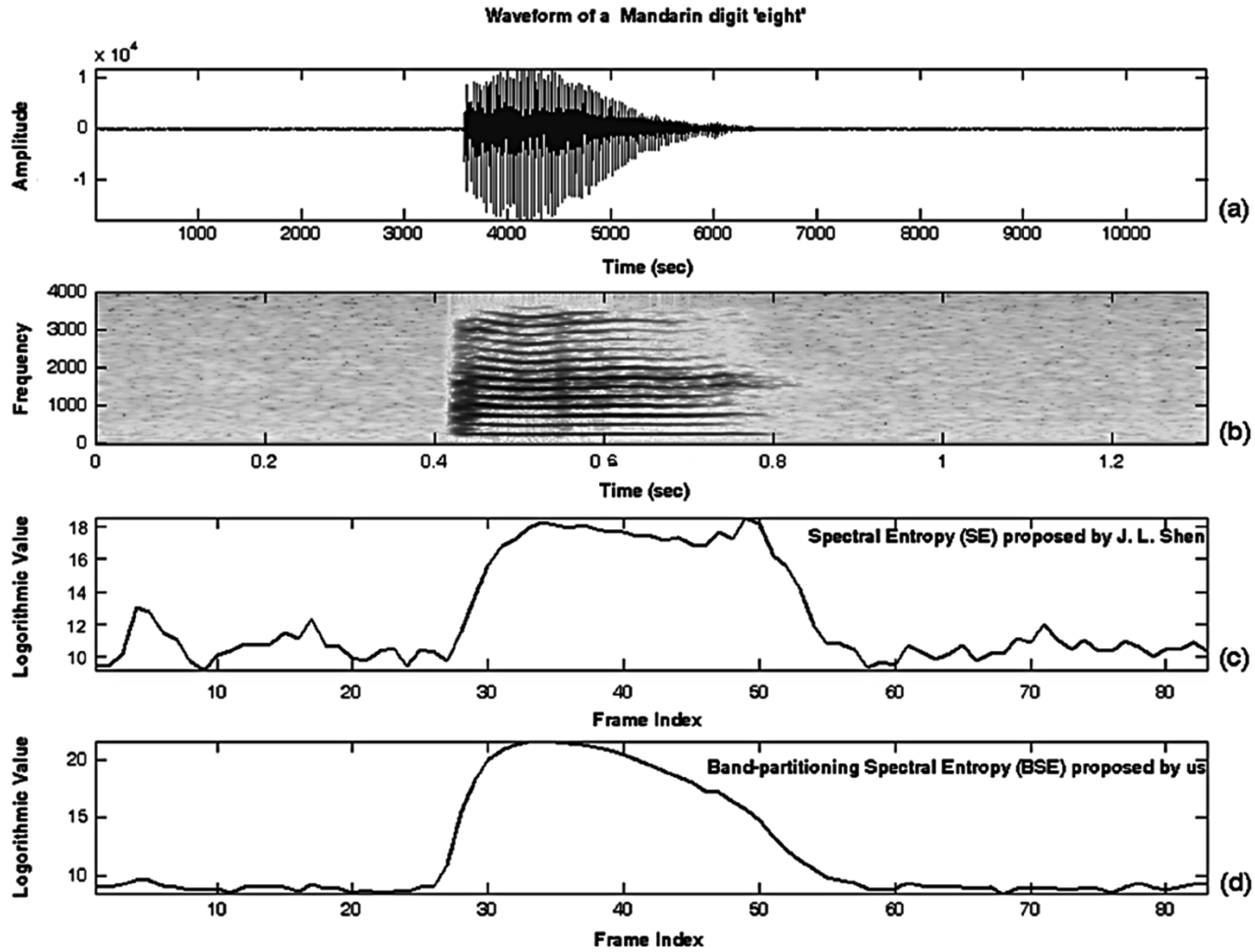

Fig. 5. Illustration of characterizing speech signals by using entropy-based feature parameter. (a) Waveform of a Mandarin digit "eight." (b) The corresponding spectrogram. (c) Contour of SE proposed by J. L. Shen *et al.* [13]. (d) Contour of the proposed BSE.

band, and $P_{b\_\text{offset}}(\cdot)$ means the normalization of band energy of each band. If the distribution of the band energies from the $(m-1)$th band to the $(m+1)$th band is concave or protruding, then a banded structure appears in these bands for the given frame. Similarly, if the distribution of the band energies from the $(m-1)$th band to the $(m+1)$th band is fairly flat, then this states that these bands do not include any banded structure. By using the set of weighting factors $W(m,l)$, the proposed BSE parameter is finally represented as follows:

$$H_b(l) = \sum_{m=1}^{N_b} W(m,l) \cdot P_b(m,l) \cdot \log[1/P_b(m,l)]. \quad (12)$$

Fig. 5 clearly indicates that the first proposed BSE parameter, using the method of band decomposition along with a set of weighting factors $W(m,l)$, more sufficiently characterizes the speech signals than other entropy-based parameter such that SE parameter proposed Shen *et al.* [13].

C. Adaptive Band-Partitioning Spectral Entropy

In fact, the frequency energies of difference types of noise are concentrated on different frequency bands [12], as shown in Fig. 6. This observation demonstrates that the bands with larger noisy energy more contaminate the useful frequency information than do the other bands. The bands with larger noisy energy (called the harmful bands) must be discarded accurately to yield more accurate frequency information. Although the BSE remains a good feature parameter, the detection sometimes fails at seriously low SNRs, especially when relatively harmful bands are involved. How to discard the harmful bands or preserve the useful bands becomes a serious task. Wu *et al.* [12] showed that the number of harmful bands (or useful bands) is related to the background noise level. A MiMSB parameter proposed in [15] was used to estimate the varying noise level, which by adaptively choosing one band with minimum energy. In this paper, we must make the number of useful bands closely correlate with a MiMSB parameter. Regardless of changing level of noise, a normalized minimum band energy (NMinBE) parameter is proposed to precisely decide the number of useful bands. An NMinBE parameter is determined as follows:

$$\text{NMinBE}(l)^{=} - \log\left[\min\{E_b(\text{m},l)\} \Big/ \sum_{m=1}^{N_b} E_b(\text{m},l)\right] \quad (13)$$

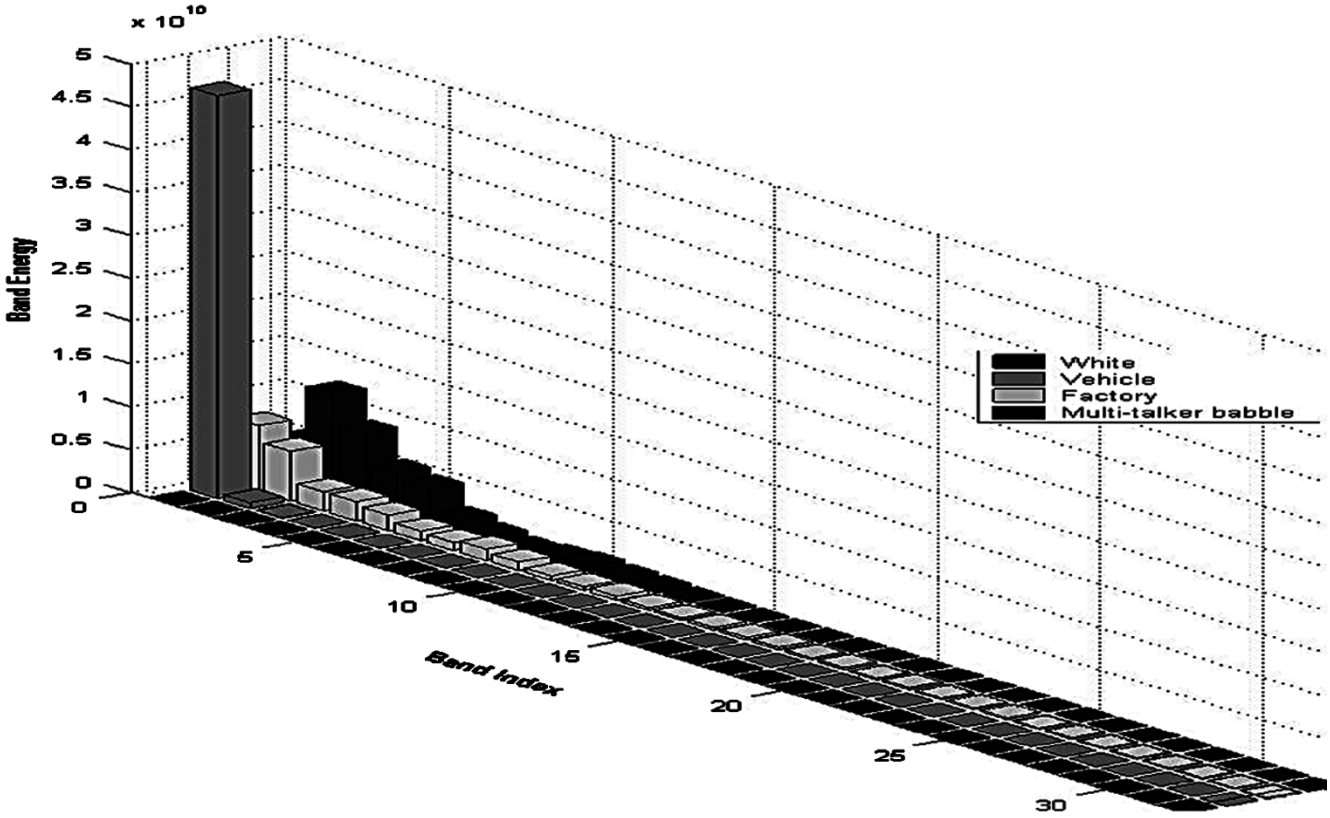

Fig. 6. Different types of noises focusing on different frequency bands.

where the $\min\{\cdot\}$ operator selects the minimum band energy among all 32 band energies for a given frame, and $\log[\cdot]$ is the logarithmic operation. The number of useful bands, $N_{\rm ub}(l)$, required to yield reliable information. Fig. 7 displays the relation between $N_{\rm ub}(l)$ and $\mathrm{NMinBE}(l)$ [see (14), shown at the bottom of the bottom of the page].

It is observed that a large $N_{\rm ub}(l)$ should be used at a low noise level (corresponding to a high SNR), and a small $N_{\rm ub}(l)$ should be used at a high noise level (corresponding to a low SNR). According to (14), for the lth frame, the first $(32 - N_{\rm ub}(l))$ frequency bands with larger energies are adaptively selected to remove noise component. The BSE parameter with an adaptive band selection method, called the adaptive band-partitioning spectral entropy (ABSE) parameter, is defined as follows:

$$H_b(l) = \sum_{m=1}^{N_{\rm ub}(l)} W(m,l) \cdot P_b(m,l) \cdot \log[1/P_b(m,l)]. \quad (15)$$

Generally, variation in the background noise level causes $N_{\rm ub}(l)$ to vary with time. This results in a varying $N_{\rm ub}(l)$ over an entire signal. To evaluate the efficiency of adaptive band selection, factory noise was added to a recorded speech signal, and then the results of ABSE-based method are compared with that of the BSE-based one. The ABSE parameter can be approximately evaluated by manually selecting useful bands according to the relationship between $N_{\rm ub}(l)$ and noise level, as described in (14). Fig. 8(a) and (b) plots the waveform of someone's saying the Mandarin digit "eight" with increasing level of factory noise and the corresponding spectrogram,

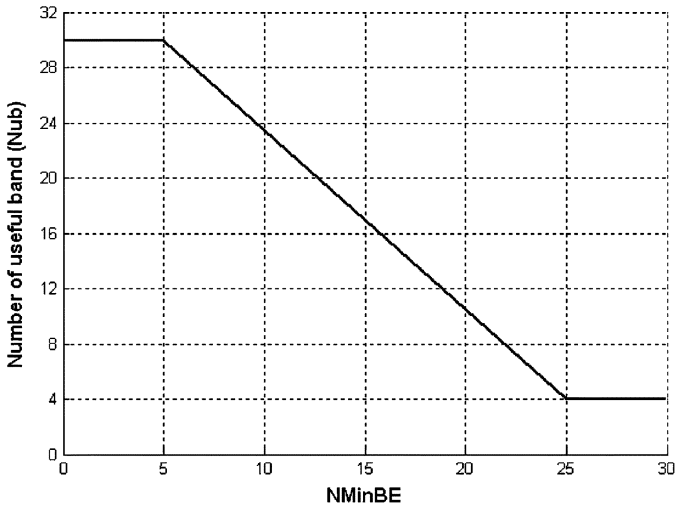

Fig. 7. Relation between $N_{\rm ub}$ and NminBE parameter.

$$N_{\rm ub}(l) = \begin{cases} 30, & \mathrm{NMinBE}(l) < 5 \\ \left[36.5 + \frac{\mathrm{NMinBE}(l)}{(25-5)} \times (4-30)\right], & 5 < \mathrm{NMinBE}(l) < 25 \\ 4, & \text{otherwise.} \end{cases} \quad (14)$$

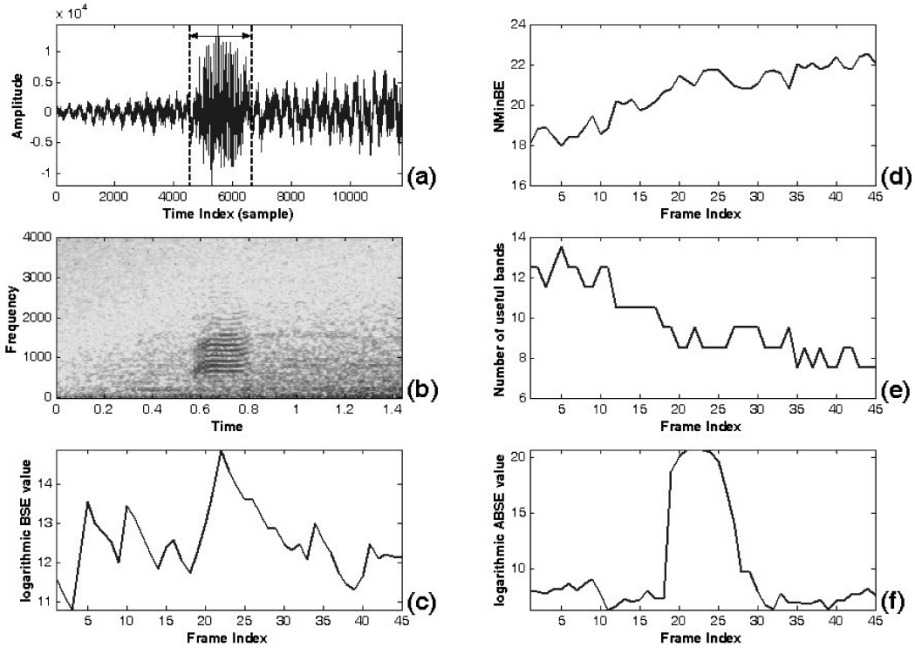

Fig. 8. Illustration of the efficiency of NMinBE parameter for applying in BSE parameter. (a) Waveform of the Mandarin digit "eight" at SNR −5 dB with increasing level of factory noise. (b) The corresponding spectrogram. (c) The contour of logarithmic BSE value under changing level of noise. (d) The NMinBE parameter proposed in this study. (e) The number of useful bands obtained from the relation between $N_{\rm ub}$ and NMinBE, as shown in Fig. 7. (f) The contour of logarithmic ABSE value obtained by manual selecting useful bands.

respectively. The endpoint is not easily detected in an adverse environment due to that the BSE parameter considers some harmful bands, as shown in Fig. 8(c). Observing Fig. 8(d), it is shown that the proposed NMinBE parameter can represent accurately the variation of noise level and provide the number of useful bands of corresponding frame, shown in Fig. 8(e). Fig. 8(f) states that the ABSE with selection of useful bands can greatly improve the performance of endpoint detection, especially at low SNRs. In order to make the ABSE-based be reliable in practical applications, the next section will develops a procedure for selecting useful bands adaptively for on-line implementation.

III. Endpoint Detection Algorithm

Generally, endpoint detection is the crucial part in speech recognition systems. It is required to enable the systems to operate smoothly in a practical test. Although existing endpoint detection algorithms are extremely accurate, they all depend on complicated computation and are not reliable in real applications. For example, Wang *et al.* [10] proposed a robust algorithm, which is an off-line method. Nemer *et al.* [17] used higher-order-statistics (HOS) parameter to detect speech, but the calculation of this parameter required too much computing time. Wu *et al.* [12] suggested an ABS method as a noise cancellation to perform ATF-based endpoint detection; however, their ABS depends on all information of the entire recorded signals. Although those algorithms are inappropriate for practical implementation, some ideas related to those algorithms are adopted herein. The ABS method proposed by Wu *et al.* [12] is strong with respect to noise cancellation. The ABS was used to preserve the useful bands (or discard the harmful bands) for each frame, but the band selection depends on an entire recorded signal. The drawbacks of ABS are, thus, as follows.

— First, the decision of band selection is not immediately determined. Since the method is an off-line strategy, its decision must be determined by analyzing an entire recorded signal.
— Second, for practical purposes, the indexes associated with the harmful bands vary with time for entire recorded signals; however, Wu *et al.* did not address this issue.

The ABS is modified herein, to overcome these two drawbacks, in the development of RABS. Previous work has shown that accurately selecting useful bands greatly improves the performance of endpoint detection in noisy environments. Wu *et al.* [12] assumed that the indexes of harmful bands were fixed; however, this assumption does not hold. In fact, the indexes of harmful bands vary with time. How can we detect that when index of harmful bands vary with time? From the Fig. 8(f), we observe that the selected bands are not contaminated by noise; the entropy value in nonspeech segments is small and smoothly and slightly varies with time. Similarly, reviewing Fig. 8(c), the selected bands are contaminated by background factory noise. However, the entropy value in nonspeech segments is large and its variation is also violent. A comparison of foregoing two observations reveals that the determined entropy value is quite large and violently varying whenever the considered bands include harmful bands. In contrast, the determined entropy value is small and its variation is very smooth if the considered bands do not include harmful bands. This finding provides a hint about how to detect whenever the indexes of harmful bands vary with time.

A. On-Line Detection

In order to detect the indexes of harmful bands immediately, an algorithm which can be performed in on-line is essential. An on-line speech detection method which may be worked in real time with minimal processing delay was implemented in [16]. Owing

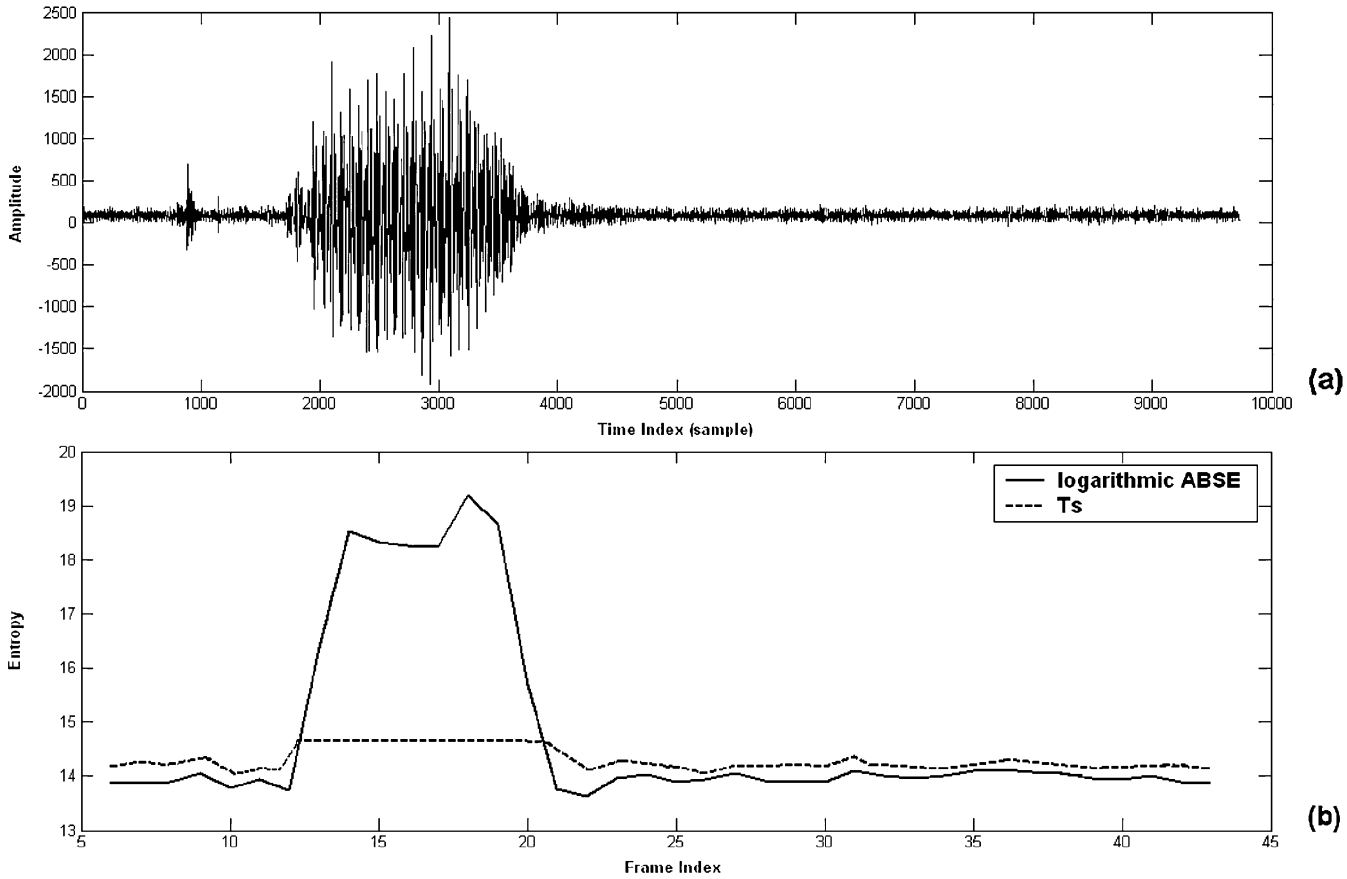

Fig. 9. An on-line detection. (a) Waveform of an utterance of the digit "one." (b) Detection of speech segments together with the logarithmic ABSE value and speech thresholds T_s.

to time-varying noise environments, the detection method sets an adaptive speech threshold to classify speech or nonspeech. The adaptive decision can be described as follows. During a short initialization period, the mean and variance of the logarithmic ABSE value is estimated over the nonspeech segments. The initial speech threshold is computed from the local noise statistics. An adaptive speech threshold T_s is determined by

$$T_s = \mu + \alpha \cdot \sigma \tag{16}$$

where μ and σ are the mean and variance of the logarithmic ABSE value during the noise-only frame, respectively, and α is an adjustment coefficient by experiment. Then, the threshold T_s is compared to the value of the most recent frame. Whenever the difference surpasses a specified threshold, speech is detected. If a given frame is detected to fall in a nonspeech period, the speech threshold T_s is updated. During nonspeech period, the relative mean and variance of the logarithmic ABSE value are updated as follows:

$$\mu_{\text{new}} = \beta \cdot \mu + (1-\beta) \cdot H_b(l) \tag{17}$$

$$\sigma_{\text{new}} = \sqrt{|H^2_{b,\text{mean}}(l) - \mu^2_{\text{new}}|} \tag{18}$$

$$H^2_{b,\text{mean}}(l) = \beta \cdot H^2_{b,\text{mean}}(l-1) + (1-\beta) \cdot H^2_b(l) \tag{19}$$

$$H^2_{b,\text{mean}}(l-1) = \frac{1}{\text{init\_period}} \sum_{k=1}^{k=\text{init\_period}} H^2_{b,\text{mean}}(l-k) \tag{20}$$

where β is also experimentally determined.

On the contrary, the speech threshold cannot be updated during speech period. Fig. 9 depicts an on-line algorithm for speech detection in an utterance of the Mandarin word for "one," together with the logarithmic ABSE value and speech threshold T_s. The results indicate clearly that the speech threshold T_s is updated during a nonspeech segment and maintained during a speech segment.

B. Refined Adaptive Band Selection

Observing from Fig. 9, we can find that the detection is robust against a tolerable variation of ABSE. If the contour of ABSE appears an abrupt peak which is caused by considering the harmful bands, the on-line detection will fails. According to the above statement, the occurrence of abrupt ABSE value indicates the possibility of variation of harmful bands. To stand for possibility of performing band selection on each frame, we propose a decision of band selection (DBS) parameter. The DBS is defined as follows:

$$\text{DBS} = \begin{cases} 1, & H_b(l) > T_s \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

The DBS parameter can be used to deal with an abrupt change in contour of ABSE to avoid wrong decision. If DBS is low for a given frame, it implies that the considered bands do not include any harmful band and the indexes of frequency bands with noise power do not vary violently with time. The indexes can be maintained from the previous frame to the current frame until DBS

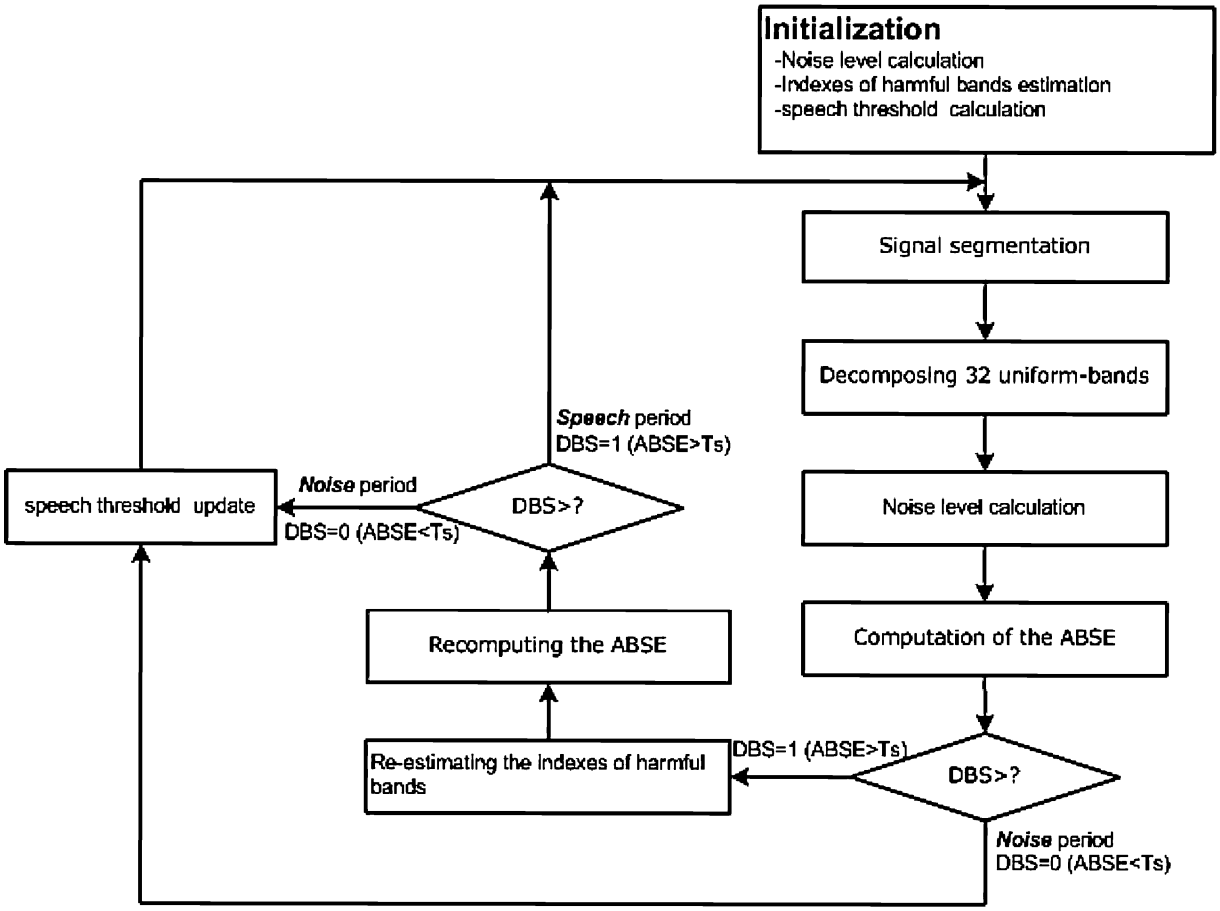

Fig. 10. Flowchart of the proposed ABSE-based endpoint detection algorithm.

TABLE I
CPU TIME OF RUNNING VARIOUS TASKS IN MATLAB PLATFORM

	Band selection method		Endpoint detection algorithm	
	RABS	ABS	ABSE-based	ATF-based [12]
CPU Time	1.046 ms	73.25 ms	65.23 ms	126.4 ms

is high; then, the task of harmful bands detection during lowness of DBS period is cancelled to reduce required computing power. Similarly, if DBS is high for a given frame, then the considered bands include some harmful bands, which result in an abrupt change in contour of ABSE. We refer that the indexes of harmful bands vary violently with time. Therefore, the task of harmful bands detection is performed in this frame. Table I lists the CPU time of various tasks performed on a Pentium-4 2.6 G, for 100 frames of a noisy speech. The following observation shows that the CPU time of RABS is much less than that of ABS. Although the complexity of computing ABSE parameter that is related the information of deciding band selection is more complicated than that of computing ATF parameter, the function of the proposed DBS parameter can make the waste of computing power reduce greatly. The decision of band selection in RABS method, which is an on-line algorithm and may be implemented in real time, is only depended on the current frame, whereas the ABS method is an off-line strategy and its decision is made from all frames (100 frames). The CPU time of the ATF-based endpoint detection algorithm is about two times of that of the ABSE-based one and it increases with the total frame size. However, the CPU time of the ABSE-based algorithm is less dependent on the frame size.

C. ABSE-Based Endpoint Detection

Flowchart diagram of the proposed ABSE-based endpoint detection algorithm is presented in Fig. 10. The proposed algorithm using RABS is described as follows.

1) Initialization: After an initial period of noise only, noise level, indexes of harmful bands and initial speech threshold are calculated in turn.
   a) Assuming that the previous five frames contain only noise, the bands with larger frequency energy are the harmful bands. The NMinBE parameter determines the number of useful bands and the indexes of these harmful bands are discarded. The ABSE parameters for previous five frames are, thus, obtained by (15).
   b) The initial speech threshold T_s is determined by (16), and μ_n and σ_n are computed by (22) and (23)

$$\mu_n = \frac{1}{5}\sum_{l=1}^{5} H_b(l) \tag{22}$$

$$\sigma_n = \frac{1}{5-1}\sum_{l=1}^{5} (H_b(l) - \mu_n)^2. \tag{23}$$

TABLE II
PERFORMANCE BETWEEN THE PROPOSED ABSE-BASED ALGORITHM AND ATF-BASED ONE IN [12] FOR VARIOUS NOISE CONDITIONS

Noise Conditions		Proposed ABSE-based algorithm		ATF-based algorithm proposed by Wu *et al.* [12]	
Type	**SNR(dB)**	**Probability of correct detection, P_c (%)**	**Probability of false detection, P_f (%)**	**Probability of correct detection, P_c (%)**	**Probability of false detection, P_f (%)**
Vehicle Noise	***40***	98.4	1.2	96.5	1.8
	20	97.2	2.3	92.3	5.7
	10	94.2	3.9	86.3	9.8
	0	91.2	6.8	74.4	14.6
Babble Noise	***40***	96.1	2.6	94.8	3.8
	20	93.7	3.4	89.2	5.9
	10	89.2	5.8	80.4	10.8
	0	81.8	9.2	65.8	16.8
Factory Noise	***40***	97.9	2.1	95.5	3.2
	20	94.6	2.8	90.6	5.7
	10	91.7	4.6	83.4	10.5
	0	82.1	8.9	70.4	15.3
White Noise	***40***	99.8	0.9	95.3	3.4
	20	98.7	1.2	89.4	6.8
	10	96.6	1.9	82.6	9.5
	0	93.3	2.7	72.2	14.9
Average		93.5	3.8	84.9	8.7

2) *Updating of speech threshold*: The speech threshold is updated for each frame in the following manner.
   - When the ABSE value determined for a given frame exceeds T_s, one of two possibilities obtains: the first is that the segment contain speech, and the other is that ABSE in error, as occurs when harmful bands are considered. Moreover, harmful bands must again be detected. If the determined ABSE remains greater than T_s, then the speech is detected and T_s is not updated.
   - Similarly, if the determined ABSE is less than T_s last, then the noise characteristics vary too violently to assume that the frequency bands in which noise is concentrated are the same as in the previous frame; then, the mean and variance of the logarithmic ABSE value are updated by (17)–(20) to form a new speech threshold T_s.
3) The starting point of speech is detected if the ABSE value continues to exceed T_s.
4) Finally, the end point of the speech is detected when ABSE is less than T_s.

IV. PERFORMANCE ANALYSIS

In this section, the performance of the ABSE-based endpoint detection algorithm is evaluated. The probabilities of correct and false detection of four kinds of noise (vehicle, multitalker babble, factory, and white noises) are calculated and compared with those associated with the ATF-based algorithm at various SNRs. The speech database used in the experiments includes a set of isolated utterances of the ten digits in Mandarin, spoken by 100 speakers. The sampling rate was 8 KHz and the speech was stored as 16-bit integers. The following metrics are defined to evaluate the performance of the proposed algorithm.

— Probability of correctly detecting speech frames, P_c: Computed as the ratio of correct speech detection to the total number of hand-labeled speech frames.

— Probability of falsely detecting speech frames, P_f: Computed as the ratio of incorrectly classified speech frames or noise frames to the total number of frames.

The practical implementation of the ABSE-based endpoint detection algorithm is also compare to that of others in a real car environment with musical background noise.

A. Artificially Added Noise

Four noise signals with various SNR levels were used in this experiment. The four noise signals-vehicle noise, multitalker babble, factory noise and white noise - were taken from the NOISEX-92 database. The noise signals were added to the recorded speech signals with different SNRs including 0, 10, 20, and 40 dB to generate noisy speech signals. Using these various types of noise and different SNRs, the P_c and P_c of proposed algorithm were compared with those of the ATF-based algorithm. Table II compares the performance of ABSE-based and ATF-based endpoint detection algorithms. The ABSE-based algorithm is observably superior to the ATF-based algorithm, especially at low SNRs. At high SNRs, the ATF-based algorithm performs as well as the ABSE-based one; however, at low SNRs the ATF parameter related to pure energy-based feature is no longer effective. Although ABS associated with ATF-based algorithm can extract useful information, the selected bands are not always useful for detecting endpoints. Additionally, in ATF-based endpoint detection, the indexes of the harmful bands are assumed to be fixed with time. This assumption is incorrect. Consequently, the performance of the ATF-based approach is seriously degraded under adverse conditions. In this study, the entropy is used to capture the banded structure on speech spectrogram and further classified speech or noise. For vehicle and white noises, whose frequencies are spread simply over the spectrum, the ABSE-based algorithm is superior to the ATF-based one under this case since their spectrograms do not show obvious a banded structure. The factory noise is as described above. Although the multitalker babbling noise, although the

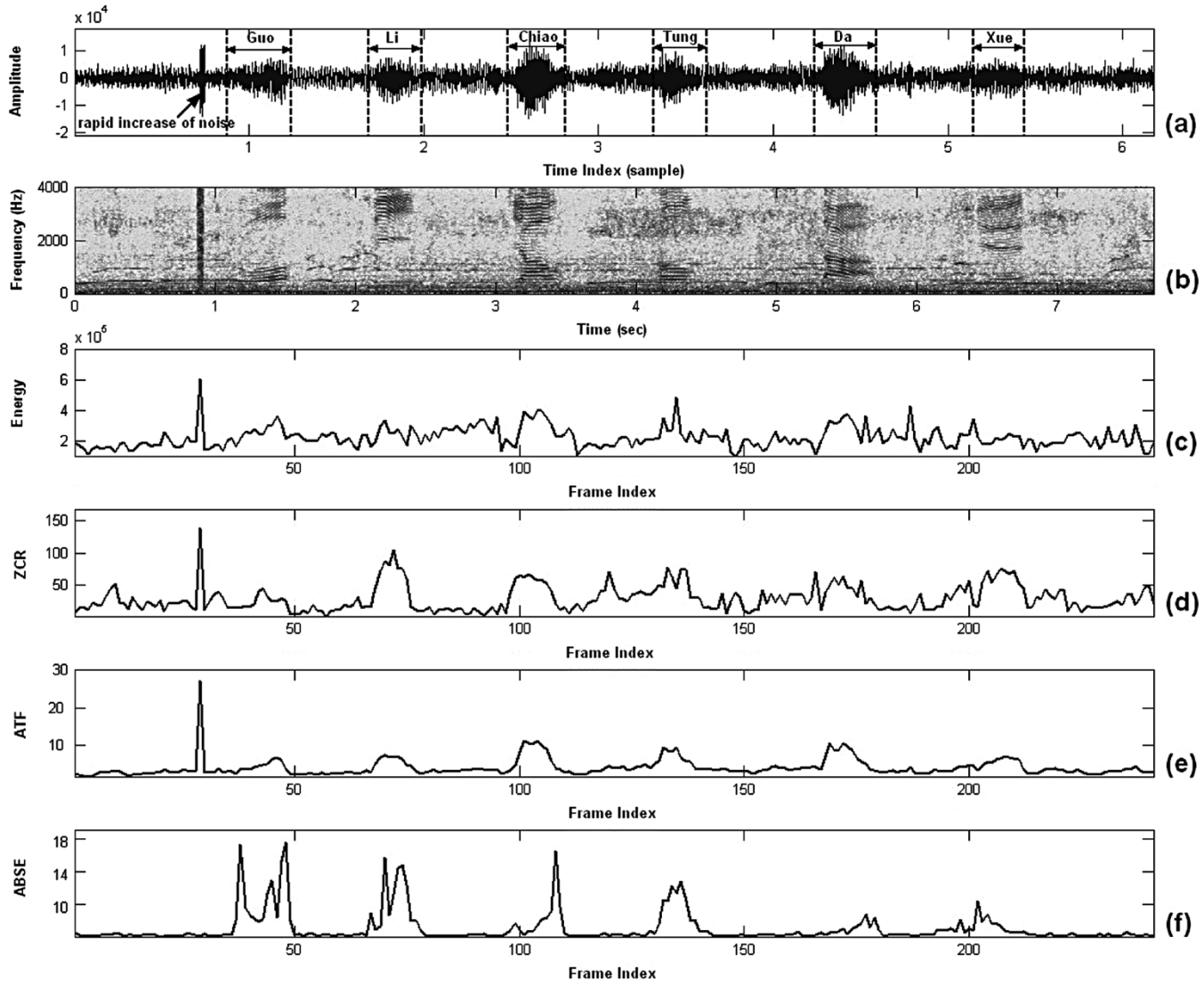

Fig. 11. Compared performance of endpoint detection using different parameters for an utterance with musical background noise inside a car. (a) Waveform of an utterance in Chinese: "Guo Li Chiao Tung Da Xue (National Chiao Tung University)." (b) The corresponding spectrogram (c) Contour of spectral energy. (d) Contour of zero-crossing rate. (e) Contour of ATF. (f) Contour of ABSE.

noise is pronounced by human, the banded nature of speech spectrums is weaker than that in speech signals. Consequently, the proposed ABSE-based algorithm still successfully outperforms the ATF-based algorithm under such conditions. The average probability of correct detection of speech frames using the ABSE-based algorithm exceeds that using the ATF-based algorithm by around 8.6%. Similarly, the average probability of false detection speech frames using the ABSE-based algorithm is less than that using the ATF-based algorithm by around 4.9%, mainly because the proposed ABSE-based parameter is a robust feature against noise, which exploits the inherent characteristic of banded nature on speech spectrogram. Besides, the ABS method associated with ATF-based algorithm has critical weaknesses that have been described above.

B. Recordings in a Car

Endpoint detection was performed using recordings of a real car with musical background noise to evaluate the effectiveness of the proposed ABSE parameter in a real environment and to compare it with other parameters. To show the performance of endpoint detection using different parameters, for an utterance in Chinese, "Guo Li Chiao Tung Da Xue," made with musical background noise in a car in Fig. 11(a). The corresponding spectrogram shows in Fig. 11(b). It is displayed that the banded nature appears only in speech spectrogram not in noise one. Fig. 11(c) and (d) demonstrates that the short-time energy and ZCR both fail in a car environment. Fig. 11(e) shows that the ATF parameter outperforms the other two. The ATF parameter can extract useful frequency information by selecting proper bands; however, it is still a purely energy-based parameter. The ATF parameter fails in a rapid increase of noise. Fig. 11(f) indicates that the ABSE parameter is superior to the other parameters, especially in a rapid increase of noise. Although noise level increases abruptly, the ABSE parameter catches only the banded nature not energy on speech spectrogram. The banded nature has shown that it can excellently specify a speech signal and be robust to

TABLE III
COMPARISON OF THE PROPOSED ABSED-BASED ALGORITHM AND ATF-BASED ONE IN CAR WITH MUSICAL BACKGROUND NOISE

	Total probability of correct detection, P_c (%)	Total probability of false detection, P_f (%)
Proposed ABSE-based algorithm	89.2	3.5
ATF-based algorithm in [12]	78.6	9.4

various types of noise in our experiments. In Table III, the ABSE-based algorithm is compared with the ATF-based one in [12] testing in a car with musical background noise. The speech database used in the experiments contains ten isolated controlled commands in Mandarin Chinese produced by 15 speakers. During the entire process, the car was moving and its radio was on. It is observed that the total probability of correct detection of the proposed ABSE-based algorithm is greater than that of the ATF-based one by about 10.6%, and the total probability of false detection of the proposed ABSE-based algorithm is smaller than that of the ATF-based one by about 5.9%

V. CONCLUSION

The objective of this study is to enhance the banded nature on speech spectrogram to develop a robust endpoint detection algorithm in adverse environments. This study has shown that the frequency energies of various types of noise are concentrated in different frequency bands and the inherent characteristic of banded nature is robust to noise. Based on the above findings, a new feature parameter, BSE, is first proposed in this study. To select useful bands effectively and accurately, a new RABS method, which is modified from ABS, was presented in this study since the indexes and numbers of harmful bands vary with time. The successful RABS method is strongly depended on an on-line detection, which is reliable in practical environment. Finally, the RABS method incorporated the BSE parameter to form a new ABSE-based endpoint detection algorithm that is effective in adverse conditions. Experimental results reveal that the ABSE-based algorithm performs excellently in the presence of four types of noise (vehicle, multitalker babble, factory, and white noise) at various SNRs. It can also be performed successfully in real cars with musical background noise. The entropy-based parameter is related only to the variation of spectral energy but not to the amount of spectral energy, so the ABSE-based algorithm outperforms the energy-based algorithm, especially in changing level of noise. The ABSE-based algorithm achieved a probability of correctly detection that was about 8.6% greater than that of ATF-based one and the probability of false detection that was lower by around 4.9%. Given a rapid increase of noise, the ABSE feature parameter is clearly superior to others, including short-time energy, ZCR, LPCs, and Cepstral features. Our future work will apply the proposed endpoint detection algorithm to a speech recognition system in a real environment, such as in a car, with a view to achieving a high recognition rate. We recommend that the endpoint detection algorithm proposed in this paper be replicated in a voice-controlled environment. The performance of speech recognition is excellent if an endpoint detection algorithm is highly reliable in real, adverse environments. Furthermore, voice-controlled equipment will become popular with consumers. For example, in KTVs, singers will be able to order songs using a microphone, without pushing a button, in an environment with musical background noise and multitalker babble.

REFERENCES

[1] L. Karray, C. Mokbel, and J. Monne, "Solutions for robust speech/non-speech detection in wireless environment," presented at the IVTTA, Sep. 1998.
[2] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[3] 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*, May 1989, pp. 369–372.
[4] L. Lamel, L. Labiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detect for isolated word recognition," *IEEE ASSP Mag.*, vol. 29, no. 4, pp. 777–785, Aug., 1981.
[5] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, vol. 8, pp. 45–60, 1989.
[6] H. Ney, "An optimization algorithm for determining the endpoints of isolated utterances," in *Proc. ICASSP*, 1981, pp. 720–723.
[7] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in *Proc. ICASSP*, May 1977, pp. 323–326.
[8] J. A. Haign and J. S. Mason, "Robust voice activity detection using cepstral features," *Proc. IEEE TEN-CON*, pp. 321–324, 1993.
[9] R. Chengalvarayan, "Robust energy normalization using speech/non-speech discriminator for German connected digit recognition," in *Proc. Eurospeech*, Sep. 1999, pp. 61–64.
[10] J. F. Wang and S. H. Chen, "A voice activity detection algorithm based on perceptual wavelet packet transform and teager energy operator," *Proc. ISCSLP*, pp. 177–180, Aug. 2002.
[11] J. C. Junqua, B. Mak, and B. Revaes, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 406–412, Jul. 1994.
[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, pp. 541–554, May 2000.
[13] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," presented at the ICSLP, 1998.
[14] L. S. Sheng and C. H. Yang, "A novel approach to robust speech endpoint detection in car environments," in *Proc. ICASSP*, 2000, pp. 1751–1754.
[15] C. T. Lin, J. Y. Lin, and G. D. Wu, "A robust word boundary detection algorithm for variable noise-level environment in cars," *IEEE Trans. Intell. Transp. Syst.*, vol. 3, no. 1, pp. 89–101, Mar. 2002.
[16] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Proc. Eurospeech*, vol. 3, 1997, pp. 1095–1098.
[17] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 217–231, Mar. 2001.
[18] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, pp. 1751–1754, 2000.
[19] S. Basu, B. Clarkson, and A. Pentland, "Smart headphones: Enhancing auditory awareness through robust speech detection and source localization," in *Proc. ICASSP*, 2001, pp. 3361–3364.

Bing-Fei Wu (S'89–M'92–SM'02) was born in Taipei, Taiwan, R.O.C., in 1959. He received the B.S. and M.S. degrees in control engineering from National Chiao-Tung University (NCTU), Hsinchu, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1992.

From 1983 to 1984, he was with the Institute of Control Engineering, NCTU, as an Assistant Researcher. From 1985 to 1988, he was with the Department of Communication Engineering, NCTU, as a Lecturer. Since 1992, he has been with the Department of Electrical Engineering and Control Engineering, NCTU, where he is currently a Professor. As an active industry Consultant, he was also involved in the chip design and applications of the flash memory controller and 3C consumer electronics in multimedia. His research interests include vision-based smart car control in ITS, multimedia signal compression, and wavelet analysis and applications.

Prof. Wu is a member of the Chinese Automatic Control Society, the Intelligent Transportation Society of Taiwan, the Chinese Institute of Electrical and Electronic Engineers, and the Chinese Institute of Engineers (CIE). He founded and served as the Chair of the IEEE Systems, Man, and Cybernetics Society, Taipei Chapter, Taiwan, 2003. He was the Director of The Research Group of Control Technology of Consumer Electronics in the Automatic Control Section of National Science Council (NSC), Taiwan, from 1999 to 2000. His research has been honored by the Ministry of Education (MOE) as the Best Industry-Academics Cooperation Research Award. He received the Distinguished Engineering Professor Award from CIE in 2002; the Outstanding Information Technology Elite Award in 2003; the Golden Linux Award in 2004; the Outstanding Research Award in 2004 from NCTU; the Research Awards from NSC in the years of 1992, 1994, and 1996 to 2000; the Golden Acer Dragon Thesis Award sponsored by the Acer Foundation in 1998 and 2003, respectively; the First Prize Award of the We Win (Win by Entrepreneurship and Work with Innovation and Networking) Competition hosted by Industrial Bank of Taiwan in 2003; and the Silver Award of Technology Innovation Competition sponsored by the Advantech Foundation in 2003.

Kun-Ching Wang (S'02) was born in Kaohsiung, Taiwan, R.O.C., in 1976. He received the B.S. degree in electric engineering from Southern Taiwan University of Technology in 1998 and the M.S. degree in electric engineering from Feng Chia University, Taiwan, in 2000. He is currently pursuing the Ph.D. degree in electrical and control engineering at National Chiao Tung University, Hsinchu, Taiwan.

His research interests include adaptive signal processing, speech enhancement, speech recognition, and wavelet analysis and applications.