A Robust Word Boundary Detection Algorithm for Variable Noise-Level Environment in Cars

Chin-Teng Lin, *Senior Member, IEEE*, Jiann-Yow Lin, and Gin-Der Wu

***Abstract*—This paper discusses the problem of automatic word boundary detection in the presence of variable-level background noise in cars. Commonly used robust word boundary detection algorithms always assume that the background noise level is fixed and sets fixed thresholds to find the boundary of word signal. In fact, the background noise level in cars varies in the procedure of recording due to speed change and moving environment, and some thresholds should be tuned according to the variation of background noise level. This is the major reason that most robust word boundary detection algorithms cannot work well in the condition of variable background noise level. To solve this problem, we propose a *minimum mel-scale frequency band* (MiMSB) parameter which can estimate the varying background noise level in cars by adaptively choosing one band with minimum energy from the mel-scale frequency bank. With the MiMSB parameter, some preset thresholds used to find the boundary of word signal are no longer fixed in all the recording intervals. These thresholds will be tuned according to the MiMSB parameter. We also propose an *enhanced time–frequency* (ETF) parameter by extending the time–frequency (TF) parameter proposed by Junqua *et al.* from single band to multiband spectrum analysis, where the frequency bands help to make the distinction between speech signal and noise. The ETF parameter can extract useful frequency information by choosing some bands of the mel-scale frequency bank. Based on the MiMSB and ETF parameters, we finally propose a new robust algorithm for word boundary detection in variable noise-level environment. The new algorithm has been tested over a variety of noise conditions in cars and has been found to perform well not only under variable background noise level condition, but also under fixed background noise level condition. The new robust algorithm using the MiMSB and ETF parameters achieved higher recognition rate than the TF-based robust algorithm, which has been shown to outperform several commonly used algorithms, by about 5% in variable background noise level condition. It also reduced the recognition error rate due to endpoint detection to 25%, compared to an average of 34% obtained with the TF-based robust algorithm.**

***Index Terms*—Mel-scale frequency, multiband, spectrum analysis, time-frequency, word boundary detection.**

I. Introduction

THE WIDESPREAD use of mobile telephones has motivated the development of robust speech recognition systems in cars [1]. A major source of errors in automatic speech recognition systems is the inaccurate detection of the beginning and ending boundaries. In cars, the problem is further complicated by nonstationary backgrounds where there may exist concurrent noises due to movements, engine running, speed change, braking, slams, etc. These background noises can be broadly classified into three classes: impulse noise, fixed-level noise, and variable-level noise. Decreasing the distance between the mouth and microphone is one way of minimizing the effects of such transient background noise. However, this method is not user-friendly. In order to solve this problem, many researchers proposed robust word boundary detection algorithms in the presence of noise. However, they focused only on the impulse noise and fixed-level background noise. The main aim of this paper is to develop a new robust word boundary detection algorithm to attack the problem of variable-level background noise in cars.

Among the three classes of background noises, the impulse noise can be solved by the parameter of time duration. The problem of fixed-level background noise was first attacked by commonly used robust word boundary detection algorithms [1]–[5]. These algorithms usually use energy (in time domain), zero crossing rate, and time duration to find the boundary between the word signal and background noise. However, it has been found that the energy and zero-crossing rate are not sufficient to get reliable word boundaries in noisy environments, even if more complex decision strategies are used [6]. Currently, several other parameters were proposed such as linear prediction coefficient (LPC), linear prediction error energy [7], [8] and pitch information [9]. Although the LPCs are quite successful in modeling vowels [10], they are not particularly suitable for nasal sounds, fricatives, etc. The reliability of the LPC parameter depends on the noise environment. The pitch information can help to detect the word boundary, but it is not easy to extract the pitch period correctly in a noisy environment.

Four-endpoint detection algorithms were compared in [6]: an energy-based algorithm with automatic threshold adjustment [4], [5], use of pitch information [9], a noise adaptive algorithm, and a voiced activation algorithm. These four algorithms are strongly dependent on the noise condition. The reliability of the parameters used by the four algorithms also depends on the noise condition. In the connection, Junqua *et al.* [6] proposed the time–frequency (TF) parameter. They used the frequency energy in the fixed frequency band 250–3500 Hz to enhance the time energy information. The TF parameter is the result obtained after smoothing the sum of the time energy and frequency energy. The frequency energy helps us to make the distinction between speech and noise. Based on the TF parameter, a robust algorithm was proposed in [6] to get more precise word boundary in noisy environment. This robust algorithm includes

Manuscript received February 29, 2000; revised August 31, 2001. This work was supported in part by the Lee and MTI Center for Networking Research and in part by the National Research Council, R.O.C., under Grant NSC90-2213-E-009-096. The Associate Editor for this paper was R. Kohno.

The authors are with the Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: ctlin@fnn.cn.nctu.edu.tw).

Publisher Item Identifier S 1524-9050(01)10749-0.

noise classification, a refinement procedure, and some preset thresholds. Although this algorithm outperforms several commonly used algorithms for word boundary detection in the presence of noise, it could work well only for the impulse noise and fixed-level background noise. For the condition of variable-level background noise in cars, this algorithm usually results in inaccurate detection of the beginning or ending boundaries in the recording interval. There was little research about specific algorithm for processing the variable-level background noise in cars. The existing robust algorithms usually set thresholds from the first few frames of the recording interval. Then the algorithms used these preset thresholds to determine the word boundary of speech signal. These thresholds are fixed in all the recording interval.

In cars, the background noise level varies in the recording interval due to the dynamically moving environment. It is not reasonable to make all preset thresholds fixed in all the recording intervals. If the variation of background noise level is large, these fixed preset thresholds will result in incorrect location of word boundaries. In order to avoid this problem, we need to have a parameter which can efficiently reflect the variation of background noise level. Then we can use this parameter to tune the preset thresholds. Based on this concept, this paper first proposes a minimum mel-scale frequency band (MiMSB) parameter. The MiMSB parameter comes from the mel-scale frequency bank (20 bands). The 20 frequency bands are spaced on a nonlinear frequency scale (mel scale). The MiMSB parameter corresponds to the band with the lowest frequency energy, and can efficiently extract the information of background noise level. With the MiMSB parameter, some preset thresholds used to find the boundary of word signal are no longer fixed in all the recording intervals. They are tuned from time to time according to the MiMSB parameter.

In addition to being tuned for variable noise level, the thresholds in the word boundary detection algorithm are also expected to be tuned reliably according to variable types of background noises. In the TF parameter proposed by Junqua *et al.* [6], the frequency information is extracted on a single frequency band (250–3500 Hz). Since the frequency energy (i.e., magnitudes of the spectrum) of different types of noises focus on different frequency bands, more accurate frequency information can be obtained by considering multiband analysis of noisy speech signals. With this motivation, we propose a new robust parameter, called *enhanced time–frequency* (ETF) parameter, for word boundary detection in noisy environment. Like the TF parameter, the ETF parameter represents both the time and frequency features of noisy speech signals. The ETF parameter is to extend the TF parameter from single-band to multiband spectrum analysis, so it inherits the ability of the TF parameter for detecting the impulse noise. Besides, the undesired impulse noise can be further smoothed by the three-point median filter used in our algorithm. A procedure is proposed such that the ETF parameter can extract more informative frequency energy than the single-band approach to compensate the time-energy information by adaptively choosing some frequency bands. The ETF parameter is the result obtained after smoothing the sum of the time energy and frequency energy. It makes the word signal more obvious than the TF parameter that uses single frequency band.

Based on the MiMSB and ETF parameters, we propose a robust word boundary detection algorithm for variable background noise level. If the background noise level changes gradually in the recording interval, the proposed robust algorithm will automatically tune its thresholds to find the word boundary. The new proposed algorithm has been tested over a variety of noise conditions in cars and has been found to perform well not only in variable background noise level environment but also in fixed background noise level environment. To simulate the varying noise-level conditions, a normal way is to use the continuous increasing/decreasing changing noises, which cover the whole spectrum of varying noise levels under consideration. The increasing/decreasing changing noises can also mimic the accelerating/decelerating behaviors of cars in the real environment. In our experiments, we take four typical types of noise for speech contamination. They are vehicle noise, cockpit noise, multitalker babble noise, and white noise. These noisy signals are added to the recorded speech signals with different signal-to-noise-ratios (SNRs) including 5 dB, 10 dB, 15 dB, 20 dB, and ∞ dB. The experimental results show that the new robust algorithm with the MiMSB and ETF parameters achieved higher recognition rate than the TF-based robust algorithm in [6], which has been shown to outperform several commonly used algorithms, by about 5% in variable background noise level condition. It also reduced the recognition error rate due to endpoint detection to 25%, compared to an average of 34% obtained with the TF-based robust algorithm.

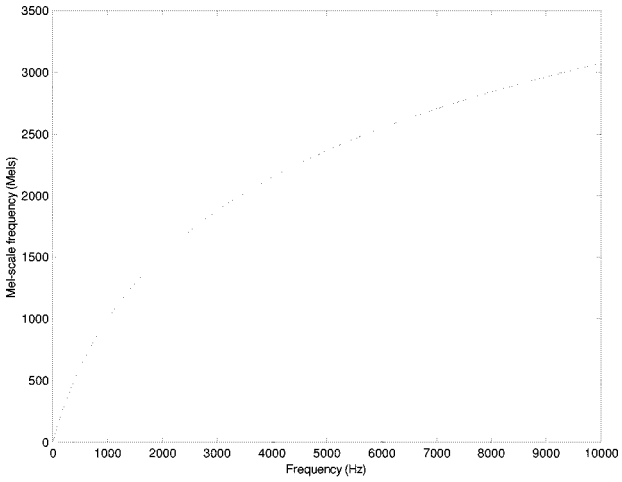
This paper is organized as follows. The minimum mel-scale frequency band (MiMSB) parameter which can estimate the variation of background noise level is derived in Section II. In Section III, we derive the ETF parameter which helps us to make the distinction between speech signal and noise. Based on the MiMSB and ETF parameters, a new robust word boundary detection algorithm for variable background noise level is proposed in Section IV. The performance evaluation and comparisons of the proposed robust algorithm are performed extensively also in Section IV. Finally, the conclusions of our work are summarized in Section V.

II. MiMSB Parameter

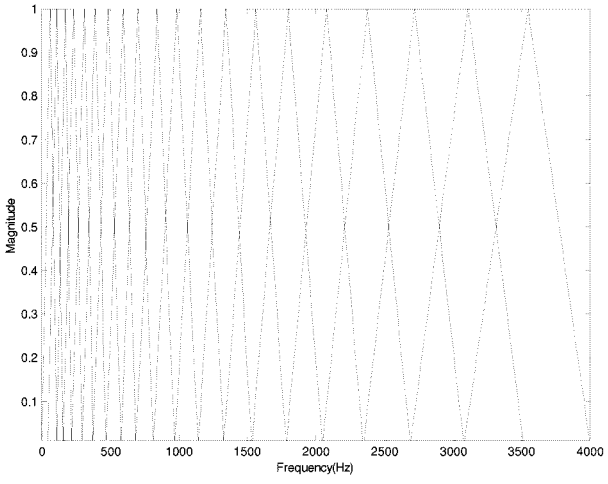
This section derives a parameter which can estimate the variation of background noise level reliably. When the background noise adds to the speech signal, we cannot clearly get the background noise level in the presence of word signal. In this section, we propose the MiMSB parameter to estimate the background noise level in the segment of word signal. The MiMSB parameter is obtained by adaptively choosing one band with minimum frequency energy from the mel-scale frequency bank. A procedure to calculate the MiMSB parameter is proposed as follows.

A. Auditory-Based Mel-Scale Filter Bank

Loosely speaking, it has been found that the perception of a particular frequency f by the auditory system is influenced by the energy in a critical band of frequencies around f [11]. Hence, an auditory-based spectrum obtained by summing



(a)



(b)

Fig. 1. (a) The relation between mel-scale frequency (Mels) and normal frequency (Hz). (b) A mel-scale filter-bank in which each filter has a triangular bandpass frequency response with bandwidth and spacing determined by a constant mel-frequency interval.

the energies in each critical band is a perceptually relevant characterization. It is also known that critical band filtering of the speech spectrum using parallel bandpass filters functionally represents an aspect of auditory processing. There is an evidence from auditory psychophysics that the human ear perceives speech along a nonlinear scale in the frequency domain. One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on a nonlinear, warped frequency scale, such as the mel scale. The relation between mel-scale frequency and frequency (hertz) is shown in Fig. 1(a), and described by the following equation [12]:

$$\text{mel} = 2595 \log(1 + f/700) \tag{1}$$

where mel is the mel-frequency scale and f is in hertz. The filter bank is then designed according to the mel scale as shown in Fig. 1(b), where the filters of 20 bands are approximated by simulating 20 triangular bandpass filters, $f(i, k)$ $(1 \le i \le 20$, $0 \le k \le 63)$, over a frequency range of 0–4000 Hz. Hence, each filter band has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval by (1). The value of the triangular function, $f(i, k)$, in the figure also represents the weighting factor of the frequency energy at the kth point of the ith band.

With the mel-scale frequency bank given in Fig. 1(b), we can now calculate the energy of each frequency band for each time frame of a speech signal. Consider a given time-domain noisy speech signal, $x_{\text{time}}(m, n)$, representing the magnitude of the nth point of the mth frame. We first find the spectrum, $x_{\text{freq}}(m, k)$, of this signal by discrete Fourier transform (128-point DFT)

$$x_{\text{freq}}(m, k) = \sum_{n=0}^{N-1} x_{\text{time}}(m, n) W_N^{kn}, \quad 0 \le k \le N-1;\ 0 \le m \le M-1 \tag{2}$$

$$W_N = \exp(-j2\pi/N) \tag{3}$$

where $x_{\text{freq}}(m, k)$ is the magnitude of the kth point of the spectrum of the mth frame, N is 128 in our system, and M is the number of frames of the speech signal for analysis. We then multiply the spectrum $x_{\text{freq}}(m, k)$ by the weighting factors $f(i, k)$ on the mel-scale frequency bank and sum the products for all k to get the energy $x(m, i)$ of each frequency band i of the mth frame

$$x(m, i) = \sum_{k=0}^{N-1} |x_{\text{freq}}(m, k)| f(i, k), \quad 0 \le m \le M-1;\ 1 \le i \le 20 \tag{4}$$

where i is the filter band index, k is the spectrum index, m is the frame number, and M is the number of frames for analysis.

We found in our experiments that the energy $x(m, i)$ obtained in (4) usually had some undesired impulse noise and was covered by the energy of background noise. Hence, we further smooth it by using a three-point median filter to get $\hat{x}(m, i)$

$$\hat{x}(m, i) = \frac{x(m-1, i) + x(m, i) + x(m+1, i)}{3}. \tag{5}$$

Finally, the smoothed energy, $\hat{x}(m, i)$, is normalized by removing the frequency energy of the beginning interval, Noise_freq, to get $X(m, i)$, where the energy of the beginning interval is estimated by averaging the frequency energy of the first five frames of the recording

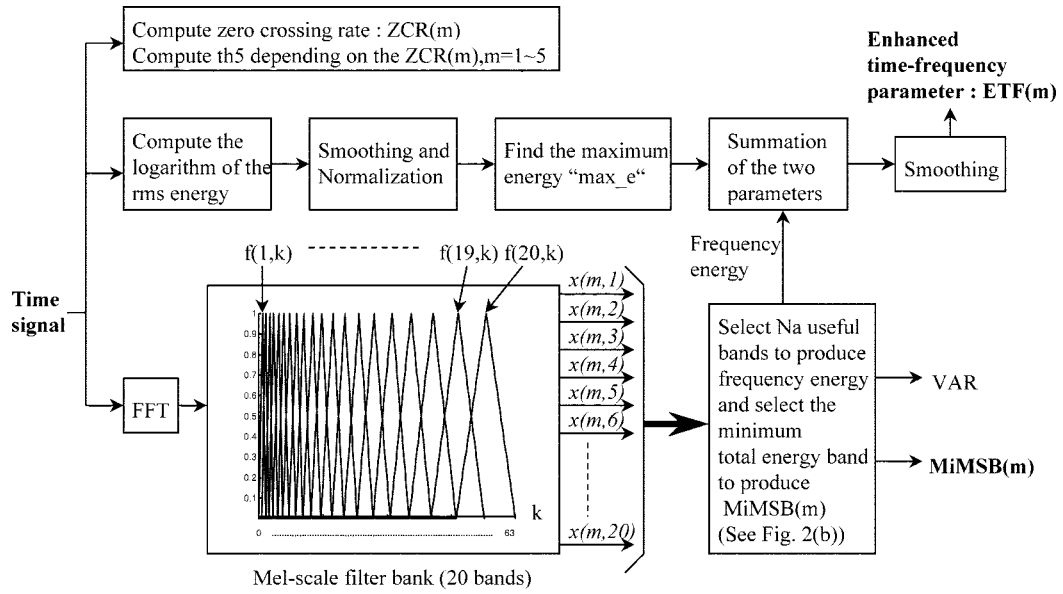
$$X(m, i) = \hat{x}(m, i) - \text{Noise\_freq} = \hat{x}(m, i) - \frac{\sum_{j=0}^{4} \hat{x}(j, i)}{5}. \tag{6}$$

With the smoothed and normalized energy of the ith band of the mth frame, $X(m, i)$, we can calculate the total energy of the nearly pure speech signal at the ith band as $E(i)$

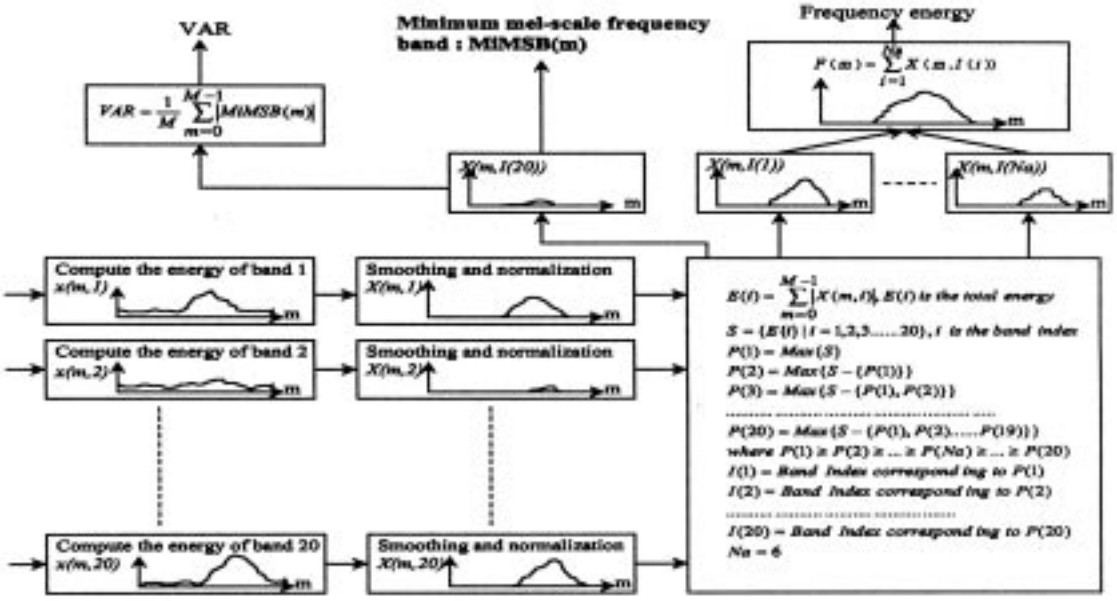
$$E(i) = \sum_{m=0}^{M-1} X(m, i). \tag{7}$$

B. Minimum-Energy Band Selection

Since our goal is to extract the information of variation of background noise level, we need a parameter to stand for the



(a)



(b)

Fig. 2. (a) Flowchart for computing the parameters and thresholds in the proposed robust word boundary detection algorithm. (b) Minimum band selection procedure [in (a)] for computing the MiMSB and VAR parameters, and adaptive band selection procedure [in (a)] for computing the frequency energy.

amount of the background noise. It is understood that $E(i)$ in (7) cannot represent the total (frequency) energy of the exactly pure speech signal, since the part of the word signal covered by background noise is also removed in the normalization procedure. However, $E(i)$ is still a good indicator for the amount of speech information, since the more the word signal information is covered by the noise, the smaller the $E(i)$ is. In other words, the larger the $E(i)$ is, the more word signal information the ith band has. Hence, we use the total energy, $E(i)$, to stand for the amount of the word signal information in band i. In order to extract the information of background noise and reduce the effect of word signal, we choose the band having the smallest $E(i)$ to stand for the background noise.

Since the band with smaller $E(i)$ contains less pure speech information, we shall sort the 20 mel-scale frequency bands according to their $E(i)$ values. This is also a preparatory task for the adaptively band-chosen method developed in the following section. Let S be the set of all $E(i)$

$$S = \{E(i) | i = 1, 2, 3, \ldots, 20\}. \tag{8}$$

The sorting is performed as follows:

$$\begin{aligned} P(1) &= \max\{S\} \\ P(2) &= \max\{S - \{P(1)\}\} \\ P(3) &= \max\{S - \{P(1), P(2)\}\} \\ &\vdots \\ P(20) &= \max\{S - \{P(1), P(2), \ldots, P(19)\}\} \end{aligned} \tag{9}$$

where $P(1)$ is the maximum total energy, and $P(20)$ is the minimum total energy. Let the band index corresponding to $P(i)$ be

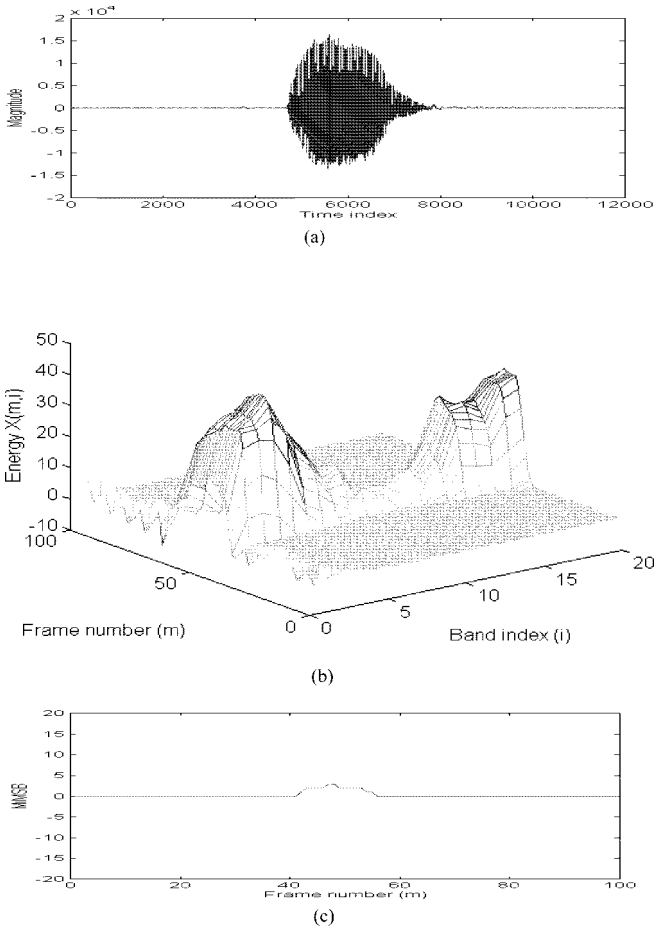


Fig. 3. (a) Speech waveform recorded in silent environment (no additive noise). (b) Smoothed and normalized frequency energy, $X(m, i)$, on 20 frequency bands, where $X(m, 13)$ has the minimum total energy. (c) The values of MiMSB parameter obtained by $X(m, 13)$.

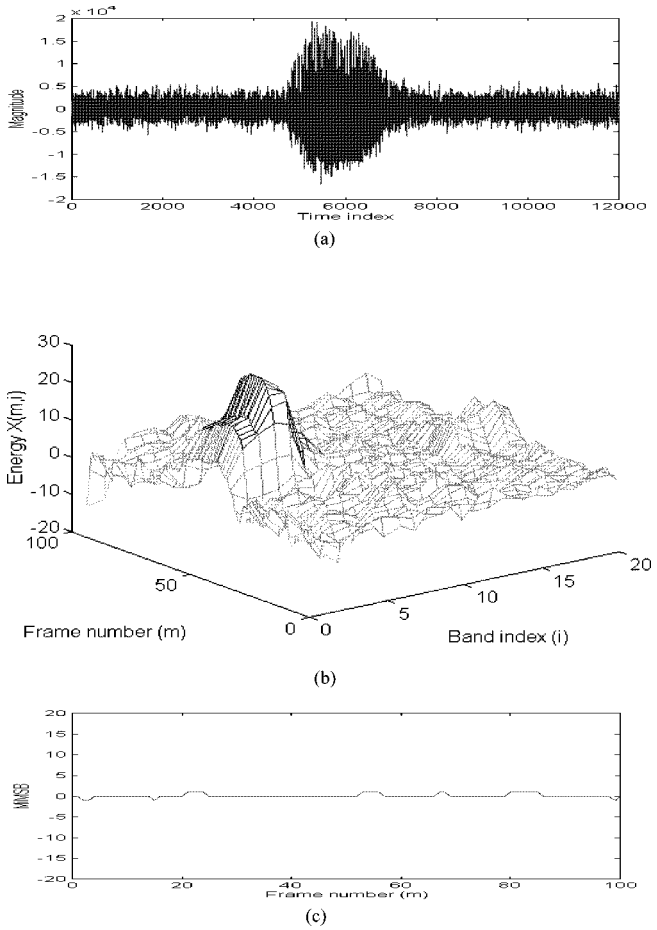


Fig. 4. (a) Speech waveform recorded in fixed noise-level environment with SNR being 5 dB. (b) Smoothed and normalized frequency energy, $X(m, i)$, on 20 frequency bands, where $X(m, 13)$ has the minimum total energy. (c) The values of MiMSB parameter obtained by $X(m, 13)$.

represented by $I(i)$, for $i = 1, 2, \ldots, 20$. That is, $I(1)$ is the index of band having the maximum total energy $P(1)$, and $I(20)$ is that having the minimum total energy $P(20)$.

From the above analysis, the output of the band $X(m, I(20))$ is a good indicator for the variation of background noise level. We name it the minimum mel-scale frequency band parameter of the mth frame $MiMSB(m)$. The procedure to get the value of MiMSB parameter is illustrated in Fig. 2(a). The details of the block with label "Select the minimum total energy band" of this figure is shown in Fig. 2(b). Finally, we define a parameter, VAR, to be the sum of the MiMSB values over all frames

$$\text{VAR} = \frac{\sum_{m=0}^{M-1} |\text{MiMSB}(m)|}{M} \tag{10}$$

where M is the number of frames of the speech signal for analysis. The VAR parameter can tell us the average variation of background noise level.

To demonstrate the efficiency of MiMSB parameter, Figs. 3–7 show the experimental results in white noise background with different noise levels. We first see its performance in silent environment. Fig. 3(a) shows a clean speech signal. The corresponding smoothed and normalized frequency energies, $X(m, i)$ [see (6)], on 20 mel-scale frequency bands and 100 frames are shown in Fig. 3(b), which indicates that $X(m, 13)$ has the minimum total energy. The values of MiMSB parameter can be obtained by $X(m, 13)$ as shown in Fig. 3(c). It appears that the MiMSB parameter is almost constant (zero) and does reflect the level of background noise. In Fig. 4(a), the speech is recorded in the condition of fixed background noise level with SNR being 5 dB. The corresponding frequency energies $X(m, i)$ and values of MiMSB parameter are shown in Fig. 4(b) and (c), respectively. Again, the MiMSB parameter is nearly constant in the recording interval. It matches the situation of fixed background noise level. In Figs. 5 and 6, the speech signals are corrupted by background noise with increasing and decreasing levels, respectively. Accordingly, the corresponding values of MiMSB parameter form an increasing curve and a decreasing curve, respectively, in Figs. 5(c) and 6(c). From these observations, we see that the MiMSB parameter can efficiently reflect the background noise level, either in fixed noise-level background (including silent environment) or in variable noise-level background.

In the above, we focused on the band with the minimum total energy. In fact, the bands which have larger total energy are also useful. These bands can help us to make the distinction between speech signal and noise in noisy environment. We shall introduce this concept in the next section.

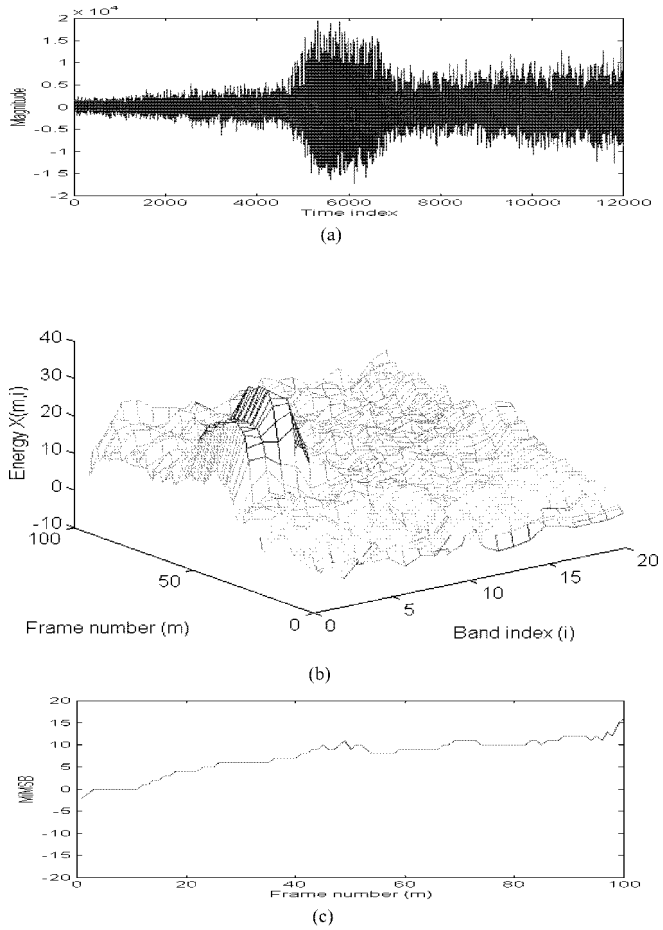


Fig. 5. (a) Speech waveform recorded in increasing noise-level environment with SNR being 5 dB. (b) Smoothed and normalized frequency energy, $X(m, i)$, on 20 frequency bands, where $X(m, 12)$ has the minimum total energy. (c) The values of MiMSB parameter obtained by $X(m, 12)$.

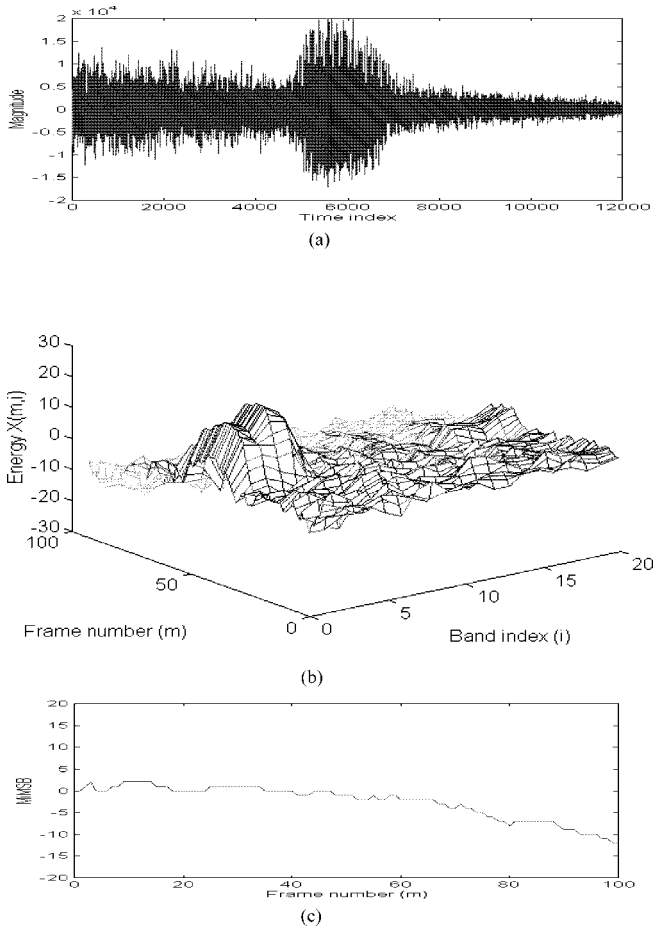


Fig. 6. (a) Speech waveform recorded in decreasing noise-level environment with SNR being 5 dB. (b) Smoothed and normalized frequency energy, $X(m, i)$, on 20 frequency bands, where $X(m, 11)$ has the minimum total energy. (c) The values of MiMSB parameter obtained by $X(m, 11)$.

III. ETF Parameter

In general, the word boundary is susceptible to noise corruption because the additive noise obscures the distinction between the word signal and noise. The general solution is to compensate the strength of word signal in noisy environment. It has been found that the information of frequency energy of a noisy speech signal can enhance the normally used time energy to make the distinction between word signal and background noise more obvious. In [6], Junqua *et al.* extracted the frequency energy of the signal on a single frequency band (250–3500 Hz) to form the TF parameter. In this section, we generalize the single-band analysis of the TF parameter to multiband analysis based on mel-scale frequency bank and propose a new ETF parameter. The ETF parameter is obtained by smoothing the sum of the time energy and frequency energy, where the frequency energy is contributed by six adaptively chosen frequency bands. Based on our experiments, the ETF parameter improves the word boundary detection accuracy not only in noisy environment, but also in silent background.

A. Effect of Additive Noise

Since our goal is to select some bands having the maximum word signal information, we need a parameter to stand for the amount of word signal information in each band. Based on the analysis in the previous section, we know that $E(i)$ in (7) is a good indicator for the amount of speech information. In other words, the larger the $E(i)$ is, the more word signal information the ith band has.

Before we consider the adaptive choices of suitable bands for extracting useful frequency information of word signal, we first make some observations on the effect of additive noise on each frequency band. Obviously, larger background noise will add more noise component into each band, and thus reduce each $E(i)$. Especially at low SNR, we found the total energy $E(i)$ of each band i become small. However, some bands are corrupted more seriously than the others. These seriously obscured bands have little word signal information left, and are not useful, if not harmful, for word boundary detection. We denote the number of bands useful for producing reliable frequency energy as N_a. We also observed that even at the same noise energy level (SNR), the useful bands were different under different noise conditions. This is because different noise sources focus their energy on different frequency bands; some focus on low frequency bands, and others on high frequency bands. The effect can be detected by the total frequency energy $E(i)$ in (7) .

We try to add white noise (10 dB) to the clean speech signal to see the effects of adding noise on each band. For illustration, the

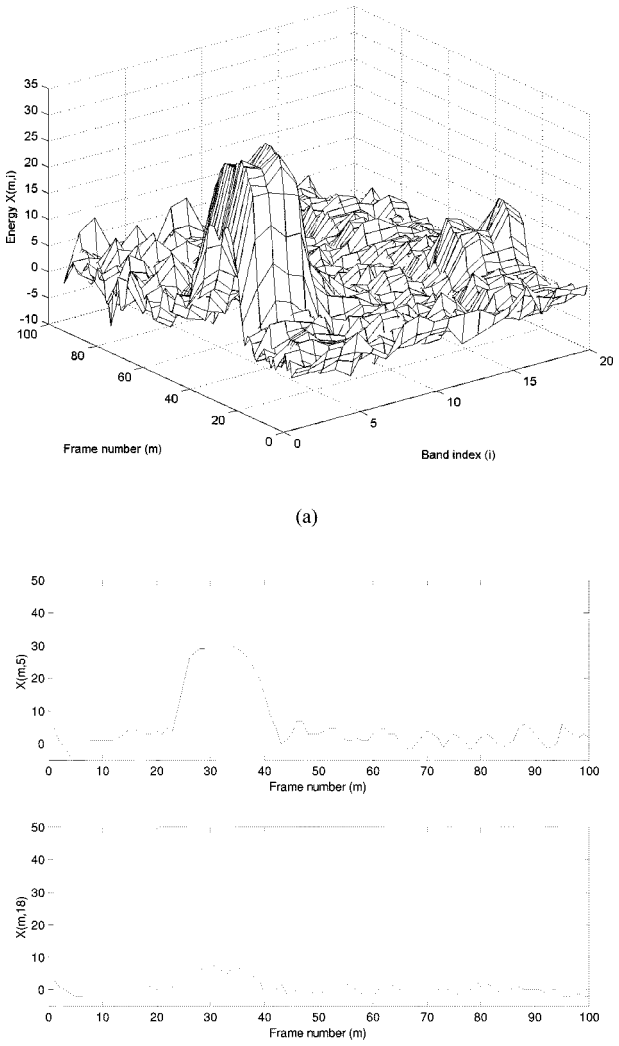


Fig. 7. Multiband spectrum analysis of the speech signal with additive white noise of 10 dB. (a) Smoothed and normalized frequency energies, $X(m, i)$, on 20 frequency bands. (b) Smoothed and normalized frequency energies, $X(m, 5)$ and $X(m, 18)$, on the 5th and 8th frequency bands.

smoothed and normalized frequency energies of a speech signal, $X(m, i)$ in (6), for 20 bands ($i = 1, 2, \ldots, 20$) and 100 frames ($m = 0, 1, \ldots, 99$) are shown in Fig. 7(a). Specifically, the energies of the 5th and 18th bands, $X(m, 5)$ and $X(m, 18)$, are shown in Fig. 7(b). From the figure, we observe that the additive noise reduces $X(m, 5)$ and $X(m, 18)$, and thus reduces $E(5)$ and $E(18)$, but we still have $E(5) \geq E(18)$. Hence, both the bands are corrupted by the additive noise. However, Fig. 7(b) shows that the 18th band is corrupted by the added noise more seriously than the 5th band. The word signal is still clear in the 5th band whose maximum $X(m, 5)$ value is about 30, but the word signal is ambiguous in the 18th band whose maximum $X(m, 18)$ value falls below 10. As a result, we cannot extract helpful word signal information from the 18th band, and we shall not treat this band as a useful frequency band. On the other hand, the 5th band is still a useful frequency band in the added white-noise environment.

B. Robust Parameter in Noisy Environment

Based on the above discussion and illustrations, we now propose a way to adaptively extract helpful frequency information of word signal. More precisely, after ordering the band indexes according to their total frequency energy ($E(i)$) as in (9), we want to decide the number N_a such that the first N_a bands ($I(1), I(2), \ldots, I(N_a)$) can produce helpful frequency energy, ($P(1) = E(I(1)), P(2) = E(I(2)), \ldots, P(N_a) = E(I(N_a))$).

By trial and error, we observed that the first 6 bands (after ordering) could provide the maximum improvement for word boundary detection in noisy environment. With $N_a = 6$, we then sum the total energies of the first N_a bands (after ordering) in (9) to get the final frequency energy, $F(m)$, of frame m:

$$F(m) = \sum_{i=1}^{Na} X(m, I(i)). \tag{11}$$

The proposed ETF parameter of the mth frame is the result obtained after smoothing the sum of the frequency energy $F(m)$ in (11) and time energy $T(m)$:

$$\text{ETF}(m) = \text{SMOOTHING}(T(m) + cF(m)) \tag{12}$$

where SMOOTHING is performed by a three-point median filter as in (5), and constant c is a proper weighting factor to adjust the scale of the ETF parameter. Different c values around 1 affect the smoothing process slightly. The typical c value that we used in the smoothing process is 1.1. The time energy $T(m)$ in (12) is given by smoothing and normalizing the logarithm of the root-mean-square (rms) energy of the time-domain speech signal:

$$x_{\text{rms}}(m) = \log \sqrt{\frac{\sum_{n=0}^{L-1} x_{\text{time}}^2(m, n)}{L}} \tag{13}$$

$$\hat{x}_{\text{rms}}(m) = \frac{x_{\text{rms}}(m-1) + x_{\text{rms}}(m) + x_{\text{rms}}(m+1)}{3} \tag{14}$$

$$\begin{aligned} T(m) &= \hat{x}_{\text{rms}}(m) - \text{Noise\_time} \\ &= \hat{x}_{\text{rms}}(m) - \frac{\sum_{j=0}^{4} \hat{x}_{\text{rms}}(j)}{5} \end{aligned} \tag{15}$$

where L is the length of the frame, which is 120 (15 ms) in our system. The procedure to calculate the ETF parameter is illustrated in Fig. 2(a). The details of the block with label "Select N_a useful bands to produce frequency energy" of this figure is shown in Fig. 2(b).

Up to now, we have proposed the MiMSB and ETF parameters to indicate the variable background noise level and the amount of word signal information, respectively. We shall next propose a new robust word boundary detection algorithm using these two parameters for variable background noise level in the next section.

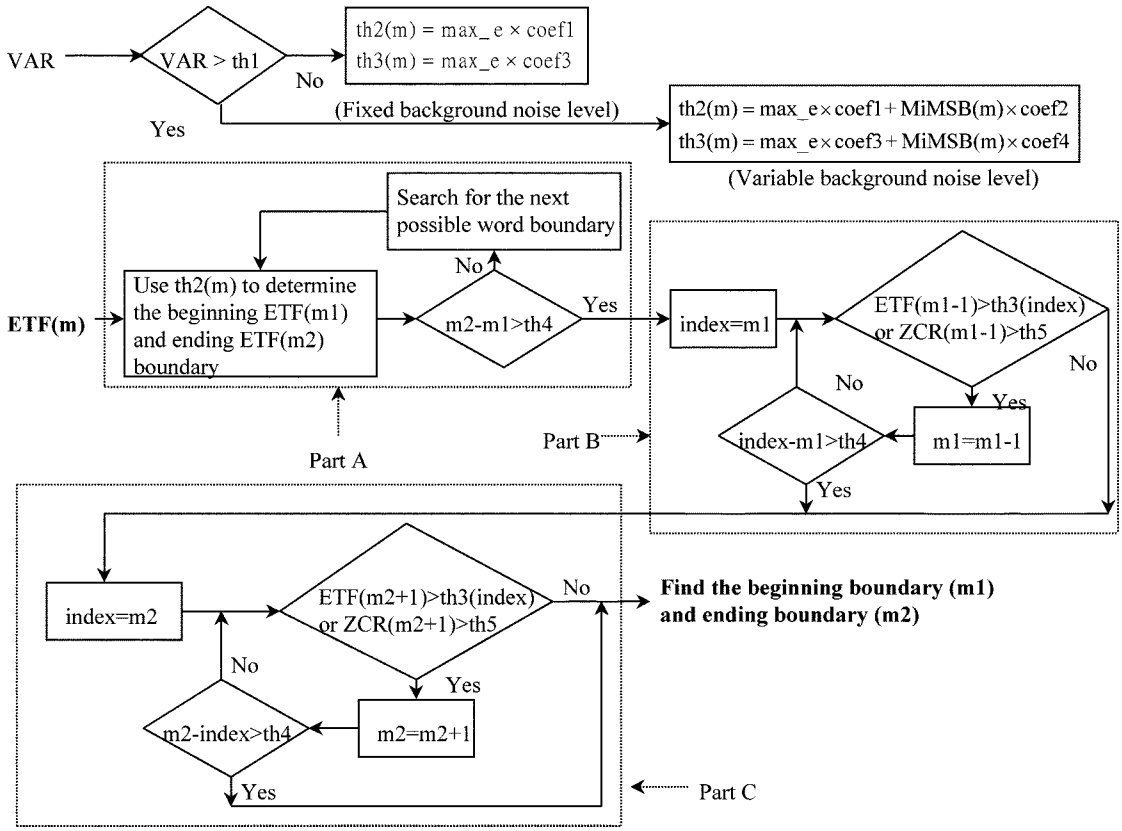


Fig. 8. Flowchart of the proposed robust algorithm for word boundary detection. Part A is to find the rough reliability boundary, Part B is to tune the rough beginning boundary, and Part C is to tune the rough ending boundary.

IV. Robust Algorithm for Variable Noise-Level Environment

In this section, we propose a new robust algorithm using the MiMSB and ETF parameters for word boundary detection. If the background noise level changes gradually in the recording interval, the proposed robust algorithm will automatically tune its thresholds to find the word boundary. The new algorithm works well in fixed noise-level environment as well as in variable noise-level environment.

A. New Robust Word Boundary Detection Algorithm

Most algorithms for word boundary detection cannot find proper boundary of the word signal in variable background noise level condition, since they cannot get the correct information of the background noise level and use it to tune some preset thresholds in all the recording intervals. Improper thresholds will result in incorrect location of the boundaries. In previous sections, we have proposed the MiMSB parameter to estimate the background noise level and the ETF parameter to make the distinction between speech signal and background noise clear. The next problem is how to use these parameters in variable noise-level background. We shall deal with this problem by proposing a new word boundary detection algorithm.

The new robust algorithm of using MiMSB and ETF parameters for word boundary detection is outlined in Fig. 8. The VAR parameter in (10) is used to stand for the average variation of background noise level, and threshold $th1$ is used to judge whether the background noise level is fixed or variable. If $\mathrm{VAR} \leq th1$, the average variation of background noise level in all the recording intervals is small. In this case, the preset thresholds $th2(m)$ and $th3(m)$ are not tuned and kept constant in all the recording intervals

$$\begin{aligned} th2(m) &= \max\_e \times coef1 \\ th3(m) &= \max\_e \times coef3 \end{aligned} \tag{16}$$

where $\max\_e$ is the maximum time energy, and $coef1$ and $coef3$ are the weighting factors for determining the thresholds $th2(m)$ and $th3(m)$ to find the word boundary in the condition of fixed background noise level. If $\mathrm{VAR} > th1$, the average variation of background noise level in the recording interval is large. In this case, thresholds $th2(m)$ and $th3(m)$ are tuned properly in the recording interval

$$\begin{aligned} th2(m) &= \max\_e \times coef1 + \mathrm{MiMSB}(m) \times coef2 \\ th3(m) &= \max\_e \times coef3 + \mathrm{MiMSB}(m) \times coef4 \end{aligned} \tag{17}$$

where the MiMSB parameter is used to estimate the background noise level, and $coef2$ and $coef4$ are used to determine the change amount of $th2$ and $th3$ due to the variable background noise level. In other words, $coef2$ and $coef4$ are used to tune the thresholds $th2(m)$ and $th3(m)$ to find the word boundary in the condition of variable background noise level. Since the ETF parameter can extract useful frequency information, it is used to find the word boundary in the noisy environment. In Part A of Fig. 8, thresholds $th2(m)$ and $th4$ are used to find the rough reliability boundary. In Part B of Fig. 8, thresholds $th3(m)$ and $th5$ are used to tune the rough beginning reliability boundary. In Part C of Fig. 8, thresholds $th3(m)$ and $th5$ are used to tune

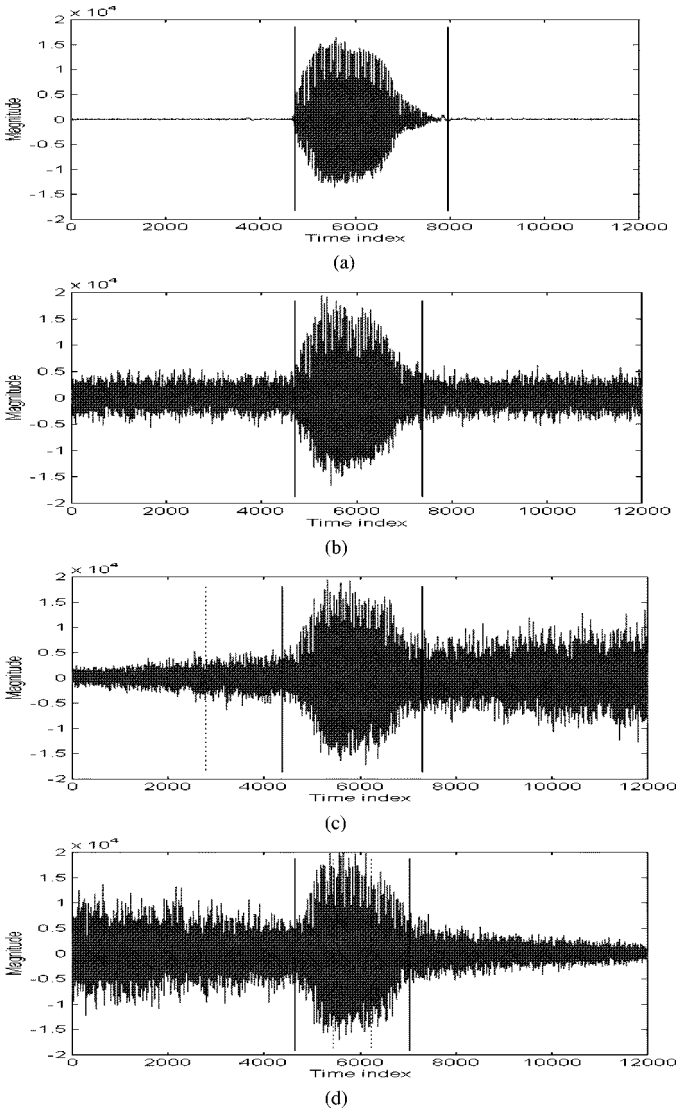


Fig. 9. Performance illustration of word boundary detection algorithms under different background noise conditions, where the word boundaries detected by the proposed MiMSB–ETF-based algorithm are shown by solid lines, and those by the TF-based algorithm are shown by dotted lines. (a) The condition of silent background. (b) The condition of fixed background noise level. (c) The condition of increasing background noise level. (d) The condition of decreasing background noise level. Noted that the solid lines and dotted lines coincide in Figs. (a) and (b), and the right-hand-side dotted line in Fig. (c) is missing. This means that the word ending boundary was not found by the TF-based algorithm.

the rough ending reliability boundary. Finally, we can obtain the beginning boundary $m1$ and ending boundary $m2$. By trial and error, we choose $coef1 = 0.7$, $coef2 = 0.8$, $coef3 = 0.25$, $coef4 = 1$, $th1 = 5$, and $th4 = 6$ in the proposed new robust algorithm.

We call the word boundary detection algorithm in [6] as the TF-based algorithm. Compared to the TF-based algorithm, the proposed MiMSB–ETF-based robust algorithm gives more accurate word beginning and ending boundaries in the condition of variable background noise level. In Fig. 9, we add white noise with different noise levels to demonstrate the efficiency of the proposed robust word boundary detection algorithm. The word boundaries determined by the TF-based algorithm are shown as dotted lines and those by the MiMSB–ETF-based algorithm are shown as solid lines. The background noise level is fixed in Fig. 9(a) and (b). It is observed that the two algorithms find nearly the same word boundaries in the fixed background noise level condition, where dotted lines and solid lines coincide. Under the condition of variable background noise level in Fig. 9(c) and (d), the TF-based algorithm fails to find the correct word boundary because its preset thresholds cannot be tuned properly according to the variation of background noise level. In Fig. 9(c), the TF-based algorithm finds the wrong beginning boundary and cannot find the ending boundary due to the increasing noise level. Also, in Fig. 9(d), the TF-based algorithm finds the wrong location of the boundaries because the decreasing background noise level. In the proposed MiMSB–ETF-based robust algorithm, the preset thresholds $th2(m)$ and $th3(m)$ are tuned by the MiMSB parameter according to the variation of background noise level. The MiMSB parameter makes these thresholds proper from time to time to find the correct location of word boundaries as shown in Fig. 9(c) and (d).

B. Experimental Evaluation

There are two possible ways to evaluate the correctness of a word boundary detection algorithm; one is to compare the detected results to hand labeled ones, and the other is to pass the detected words into a speech recognizer to see the recognition rate. The latter approach is the most common one due to its subjective nature. In this section, we shall test the performance of the proposed MiMSB–ETF-based algorithm and compare it to the TF-based robust algorithm in [6]. In order to observe the effects of the proposed MiMSB and ETF parameters, respectively, we use the TF parameter instead of the ETF parameter in the MiMSB–ETF-based algorithm to form another word boundary detection algorithm, called MiMSB–TF-based algorithm for performance comparison. In addition, we used the ETF parameter instead of the TF parameter in the TF-based algorithm to form the ETF-based algorithm. Recognition rates of these four word boundary detection algorithms (MiMSB–ETF-based algorithm, MiMSB–TF-based algorithm, ETF-based algorithm, and TF-based algorithm) will be obtained in the following tests. The tests are performed in the variable background noise conditions in cars. Since inaccurate detection of word boundary is harmful to recognition, the performance of the word boundary detection process is examined by the recognition rate of speech recognizer. In the following, we shall introduce the used speech recognizer, test database, and the evaluation results.

Speech Recognition System: The speech recognition system used in this paper for evaluating the performance of word boundary detection algorithms is a robust isolated word recognition system consisting of two parts, feature extractor and classifier. In the feature extractor, the modified two-dimensional cepstrum (Modified TDC—MTDC) [13]–[16] is used as the speech feature. The MTDC can simultaneously represent several types of information contained in the speech waveform: static and dynamic features, as well as global and fine frequency structures. To represent an utterance, only some MTDC coefficients need to be selected to form a feature vector instead of the sequence of feature vectors. The MTDC has

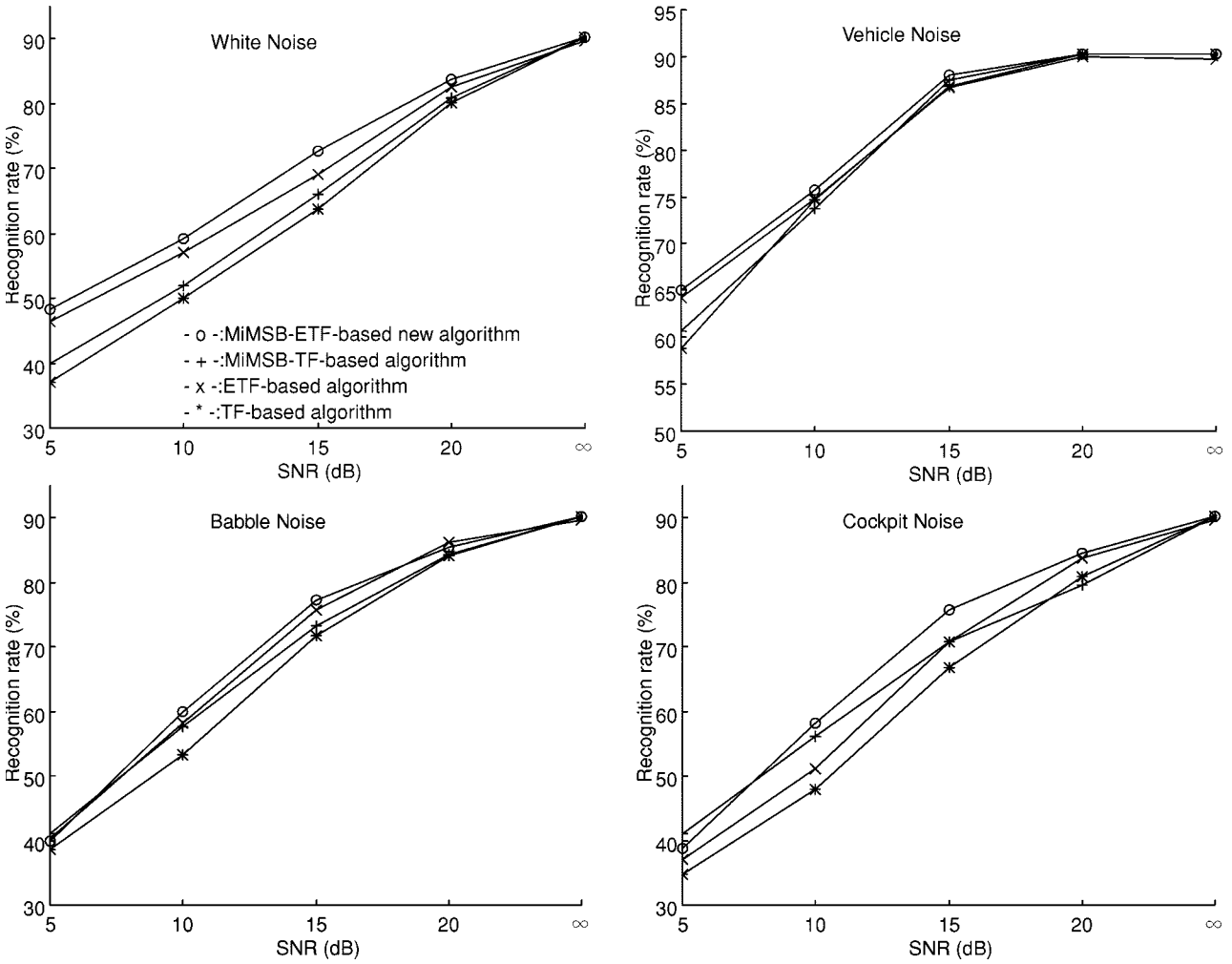


Fig. 10. Recognition rates of four word boundary detection algorithms (MiMSB–ETF-based new algorithm, MiMSB–TF-based algorithm, ETF-based algorithm, and TF-based algorithm) in the condition of variable background noise level.

the advantage of simple computation and is suitable for noisy speech recognition due to its choices of robust coefficients. In the classifier, a Gaussian clustering algorithm is used. The training was done on clean speech pronounced in a clean environment (without background noise). In the training phase, each model is trained by a mixture of four Gaussian distribution density functions. We use a total of 1000 utterances for training. The details of the above isolated word recognition system can be found in [16].

Test Environment and Noise Speech Database: In the recognition procedure, the frame window used for obtaining the MTDC features is 30 ms in length, with 15-ms overlap between two frames. In the word boundary detection procedure, the frame length is set to be 15 ms in order to get more accurate endpoint location. The sampling rate of our system is 8 kHz. The noise signals are taken from the noise database provided by the NATO Research Study Group on Speech Processing (RSG.10) NOISE-ROM-0 [17]. The database consists of 24 noise sources in order to offer as wide as possible variations in characteristics. Among these noise sources, we take four typical types of noise for speech contamination in our experiments. They are vehicle noise, cockpit noise, multitalker babble noise, and white noise. The original NOISE-ROM-0 data were sampled at 19.98 kHz and stored as 16-bit integers. In our experiments, they are prepared for use by downsampling to 8 kHz and applying attenuation on them. The attenuation was applied to enable the addition of noise without causing an overflow of the 16-bit integer range. The speech data used for our experiments are the set of isolated Mandarin digits. They are ten digits spoken by ten speakers and each speaker pronounced the ten digits 20 times. The recording sampling rate is 8 kHz and stored as 16-bit integer. To set up the noisy speech database for testing, we added the prepared noisy signals to the recorded speech signals with different SNRs including 5, 10, 15, 20, and ∞ dB. To test the proposed robust algorithm in the variable background noise condition, we change the amplitude of a given noise signal between 0.4 and 2.5 times of its nominal energy value linearly under a desired SNR level. In other words, we change the power level of the noise signal between 0.16 and 6.25 times of its nominal power value linearly. For example, if the desired SNR is 10 dB, then we change the noise level such that the SNRs vary from 1.6 to 6.25 dB. The noise level changing could be in increasing, decreasing, increasing–decreasing, or decreasing–increasing order. The duration of each utterance used for testing the performance of the word boundary detection algorithm is about

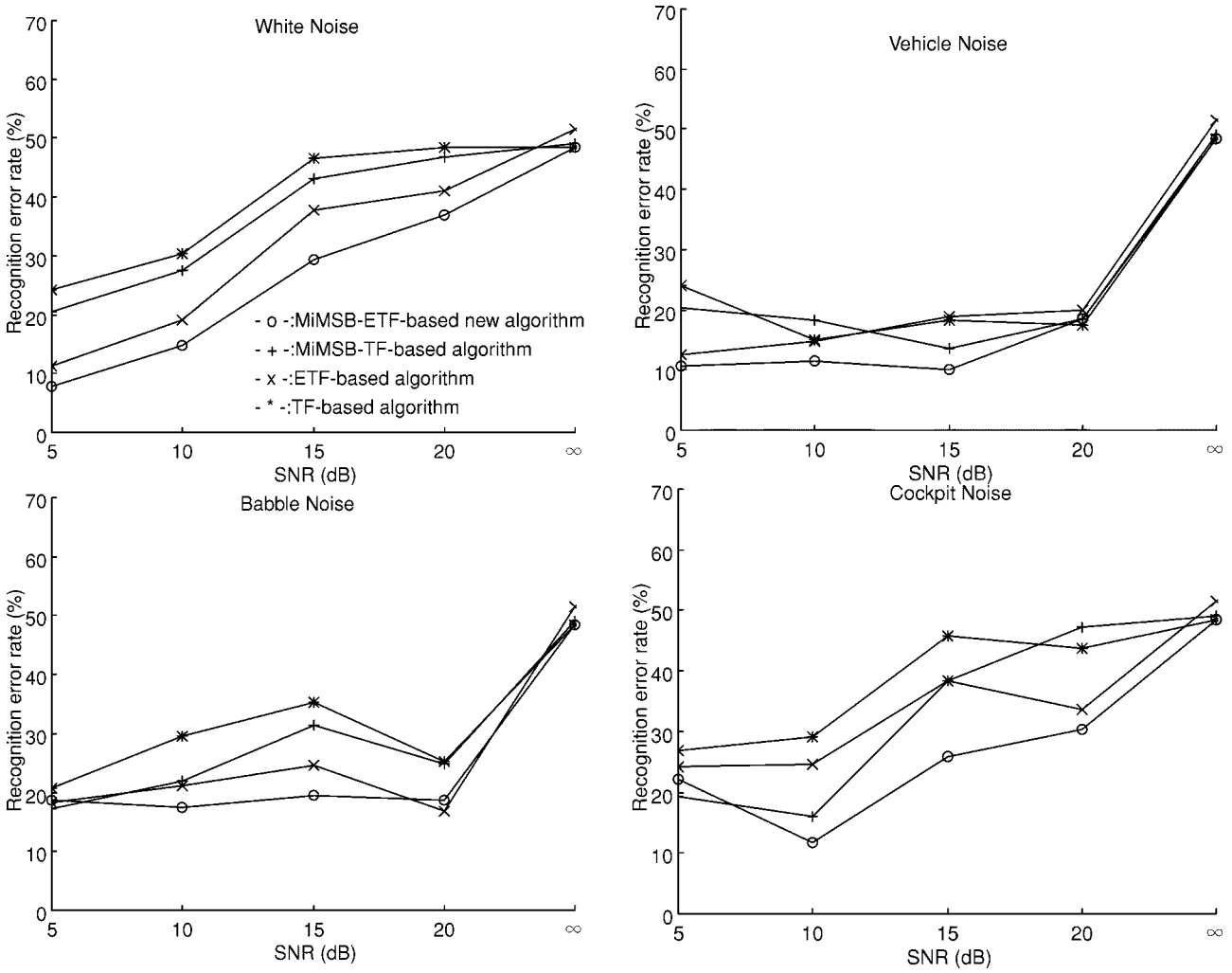


Fig. 11. Recognition error rates of four word boundary detection algorithms (MiMSB–ETF-based new algorithm, MiMSB–TF-based algorithm, ETF-based algorithm, and TF-based algorithm) in the condition of variable background noise level.

1 s (including silence). A total of 600 utterances were used in our experiments.

Experimental Results: Four word boundary detection algorithms (MiMSB–ETF-based algorithm, MiMSB–TF-based algorithm, ETF-based algorithm, and TF-based algorithm) are tested in the variable background noise condition, and the results are shown in Fig. 10. There are totally 600 utterances used in this test to simulate the variable background noise level conditions in cars; 300 utterances are recorded in the increasing background noise level condition and 300 utterances are recorded in the decreasing background noise level condition. We first make some observations on the effect of the MiMSB parameter. Since the MiMSB parameter tunes some preset thresholds according to the variation of background noise level, the MiMSB–TF-based algorithm outperforms the TF-based algorithm. Since the ETF parameter can extract more useful frequency information of word signal than the TF parameter, the ETF-based algorithm also outperforms the TF-based algorithm. By using both the MiMSB and ETF parameters, the proposed MiMSB–ETF-based algorithm outperforms the other three algorithms.

Considering another performance index, we examine the recognition error rates averaged across the four noise conditions due to incorrect word boundary detection as a function of SNRs. The results are shown in Fig. 11. Here, the recognition error rate is the ratio of the recognition errors due to incorrect word boundary detection (taking the recognition scores obtained by hand labeling as a reference) to the total number of recognition errors of the detection algorithm [6]. More precisely, let the recognition errors obtained by hand labeling be E_{hl}, and the recognition errors obtained by using the automatic word boundary detection algorithm be E_{al}. Then the recognition error rate is given by $(E_{\mathrm{al}}-E_{\mathrm{hl}})/E_{\mathrm{al}}$. This index represents the percentage of recognition errors attributable to word boundary detection errors relative to the total number of errors, where the recognition rate with hand-labeled boundaries is used as a reference. By averaging the experimental results obtained in Fig. 11 with both different background noise types (vehicle noise, cockpit noise, multitalker babble noise, and white noise) and different SNR decibels, (5, 10, 15, 20, and ∞ dB) we get the averaged recognition error rates of these four algorithms which are 25%, 31%, 30%, and 34%, respectively. In other

words, these rates are obtained by averaging the point values on the curves corresponding to each algorithm in Fig. 11. It shows that the proposed MiMSB–ETF-based algorithm reduced the recognition error rate due to endpoint detection to 25%, compared to an average of 31%, 30%, and 34% obtained by the MiMSB–TF-based algorithm, ETF-based algorithm, and TF-based algorithm, respectively. The MiMSB–ETF-based algorithm still outperforms the other three algorithms.

As a summary, since the MiMSB parameter tunes some preset thresholds in all the recording intervals and the ETF parameter extracts more useful frequency information of word signal than the TF parameter, the proposed MiMSB–ETF-based algorithm achieves higher recognition rate than the TF-based algorithm by about 5% in the variable background noise level condition. It also reduces the recognition error rate due to endpoint detection to 25%, compared to an average of 34% obtained with the TF-based robust algorithm.

In the proposed robust word boundary detection algorithm, we use some segmental parameters such as $E(i)$, VAR, and MiMSB to help the detection of word boundaries. Since most of these parameters are independent of those used in the speech recognizer, they need extra computation for each frame. Although this two-phase process (i.e., word detection and word recognition) is a normal mechanism existing in a speech recognition system, our word detection scheme is more time consuming than the normal approaches. However, this is the reasonable expense paid for a robust word boundary detection scheme suitable, especially, for varying background noise like the in-car environment. In all the segmental parameters of our algorithm, only the VAR parameter cannot be obtained on-line; it needs to be calculated after a set of speech frames has arrived. The VAR is to detect the average variation of background noise level and to determine if the tuning of the two thresholds, $th2$ and $th3$ in (17), are necessary. Hence, the VAR parameter need not be calculated at any time; once it detects the varying noise environment, we can switch on the tuning phase of thresholds and keep this phase running for a period of time without recalculating VAR. So, the proposed algorithm can *on-line* detect word boundaries practically in adverse environments. This makes it have potential for real-time operation, depending the computation power of the used hardware platform.

V. Conclusions

In this paper, we first proposed a MiMSB parameter which can efficiently estimate the variation of background noise level in cars. This parameter adaptively chooses one band with minimum frequency energy from the mel-scale frequency bank. We also proposed a reliable parameter, ETF, that possesses both the time and frequency features for word boundary detection in noisy environment. This parameter adaptively adopts six useful bands from 20 mel-scale frequency bands for producing useful frequency features to enhance time features of word signal in noisy environment. Based on the MiMSB and ETF parameters, we proposed a new robust word boundary detection algorithm. In contrast to the commonly used robust word boundary detection algorithms which always fix all preset thresholds in the recording interval, the proposed MiMSB–ETF-based algorithm does not use fixed preset thresholds; they are tuned adaptively according to the MiMSB parameter. This makes the algorithm more reliable in the noisy environment with variable noise level. The MiMSB–ETF-based algorithm has been tested over a variety of noise conditions in cars and has been found to perform well in both fixed and variable noise-level environments. Also, the results are compared to those of other word boundary detection schemes under the same *well-behaved* speech recognizer. In our experimental evaluation, the MiMSB–ETF-based algorithm achieved higher recognition rate than the TF-based algorithm by about 5% in the variable background noise level condition. It also reduced the recognition error rate due to endpoint detection to 25%, compared to an average of 34% obtained with the TF-based robust algorithm. In our future work, we will perform some advanced experiments of the proposed robust algorithm in a real car on site.

References

[1] C. E. Mokbel and G. F. A. Chollet, "Automatic word recognition in cars," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 346–356, Sept. 1995.

[2] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297–315, Feb. 1975.

[3] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, vol. 8, pp. 45–60, 1989.

[4] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "An improved endpoint detector for isolated word recognition," *IEEE ASSP Mag.*, vol. 29, pp. 777–785, Aug. 1981.

[5] B. Reaves, "Comments on an improved endpoint detector for isolated word recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 526–527, Feb. 1991.

[6] J. C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 406–412, July 1994.

[7] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 250–255, Apr. 1993.

[8] S. J. Kia and G. G. Coghill, "A mapping neural network and its application to voiced-unvoiced-silence classification," in *Proc. 1st New Zealand Int. Two-Stream Conf. Artificial Neural Networks Expert Systems*, 1993, pp. 104–108.

[9] M. Hamada, Y. Takizawa, and T. Norimatsu, "A noise robust speech recognition," in *Proc. ICSLP*, 1990, pp. 893–896.

[10] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[11] J. B. Allen, "Cochlear modeling," *IEEE Acoust., Speech, Signal Processing*, vol. 2, pp. 3–29, 1985.

[12] D. O'Shaughnessy, *Speech Communication*. Reading, MA: Addison-Wesley, 1987, p. 150.

[13] Y. Ariki, S. Mizuta, and T. Sakai, "Spoken-word recognition using dynamic features analyzed by two-dimensional cepstrum," *Proc. Inst. Elec. Eng.*, pt. I, vol. 136, no. 2, Apr. 1989.

[14] H. F. Pai and H. C. Wang, "A study on two-dimensional cepstrum approach for speech recognition," *Comput. Speech Lang.*, vol. 6, pp. 361–375, 1992.

[15] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[16] C. T. Lin, H. W. Nein, and J. Y. Hwu, "GA-based noisy speech recognition using two-dimensional cepstrum," *IEEE Trans. Speech Audio Processing*, to be published.

[17] A. 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, pp. 247–251, 1993.

Chin-Teng Lin (S'88–M'91–SM'99) received the B.S. degree in control engineering from the National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., in 1986 and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1989 and 1992, respectively.

Since August 1992, he has been a Professor and Chairman of the Electrical and Control Engineering Department at the College of Electrical Engineering and Computer Science, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C. From 1998 to 2000, he served as the Deputy Dean of the Research and Development Office of the National Chiao-Tung University, Hsinchu, Taiwan, R.O.C. His current research interests include fuzzy systems, neural networks, intelligent control, human–machine interface, image processing, pattern recognition, video and audio (speech) processing, and intelligent transportation systems (ITS). He is the coauthor of *Neural Fuzzy Systems—A Neuro-Fuzzy Synergism to Intelligent Systems* (Prentice Hall), and the author of *Neural Fuzzy Control Systems with Structure and Parameter Learning* (World Scientific). He has published over 60 journal papers in the areas of soft computing, neural networks, and fuzzy systems, including about 40 IEEE TRANSACTIONS papers.

Dr. Lin is a member of Tau Beta Pi and Eta Kappa Nu. He is also a member of the IEEE Computer Society, the IEEE Robotics and Automation Society, and the IEEE Systems, Man, Cybernetics Society. He has been the Executive Council Member of the Chinese Fuzzy System Association (CFSA) since 1995, and the Supervisor of the Chinese Automation Association since 1998. He is the Chairman of IEEE Robotics and Automation Society, Taipei Chapter, since 2000, and the Associate Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, CYBERNETICS since 2001. Dr. Lin won the Outstanding Research Award granted by National Science Council (NSC), Taiwan, from 1997 to 2001; the Outstanding Electrical Engineering Professor Award granted by the Chinese Institute of Electrical Engineering (CIEE) in 1997; and the Outstanding Engineering Professor Award granted by the Chinese Institute of Engineering (CIE,) in 2000. He was also elected to be one of the 38th Ten Outstanding Young Persons in Taiwan, R.O.C., (2000).

Jiann-Yow Linn received the B.S. degree in control engineering from the National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., in 1997. He is currently working toward the Ph.D. degree in the Department of Electrical and Control Engineering at the same university.

His current research interests are digital signal processing, neural networks, fuzzy control, and learning systems.

Gin-Der Wu received the B.S. degree in engineering science from the National Cheng-Kung University, Taiwan, R.O.C., in 1996 and the Ph.D. degree in electrical and control engineering from the National Chiao-Tung University, Taiwan, R.O.C., in 2000.

His research interests include speech recognition and enhancement in noisy environments, adaptive signal processing, neural networks, and fuzzy control.