1) *The customer groups are correlated*: Interestingly, the demographic group `female-under-25` has almost identical preferences as that of the group `female-over-35`. Perhaps surprisingly, `female-26-35` is "closer" to `male-26-35` than to other female groups. Even more surprisingly, `all-female` is almost orthogonal to `all-male`.
2) *Yet, each customer group has its own bias*: From the first quadrant of Fig. 9, the `female-under-25` group favors a `light` design more than any other criteria. By contrast, in the third quadrant of the same figure, the `male-over-35` group prefers a design that appears to be `robust` and `traditional`.

Intransitivity of preferences can now be explained. The demographic group X = `male-under-25` prefers the design **C**1-4 over the design **A**4-2. (Geometrically, the axis for male-under-25, in the fourth quadrant of Fig. 9, receives a higher value in the projection from the point **C**1-4 than does from **A**4–2). It also turns out that the same X = `male-under-25` identifies more with a design being `innovative` and `feminine` (perhaps surprisingly) than does the group Y = `male-26-35` which identifies more with designs that are perceived as `petite`. Finally, that the group Z = `male-over-35` is almost orthogonal to Y = `male-26-35` makes transitivity of preferences, from X, to Y, to Z, almost impossible – even without the deformation arising from arbitrary projections.

ACKNOWLEDGMENT

The construction of the partitions for the 48 cellular telephone designs and the survey involved over a dozen graduate students at the National Taiwan University of Science and Technology, whose talents and dedication have made the work possible.

REFERENCES

[1] (1999) Compaq, Houston, TX. [Online]. Available: http://www.compaq.com
[2] M. P. DoCarmo, *Differential Geometry of Curves and Surfaces*. Englewood Cliffs, NJ: Prentice-Hall, 1976.
[3] A. Gray, *Modern Differential Geometry of Curves and Surfaces*. Boca Raton, FL: CRC Press, 1993.
[4] J. J. Koenderink, *Solid Shape*. Cambridge, MA: MIT Press, 1990.
[5] C. Lanczos, *The Variational Principle of Mechanics*, 4th ed. New York: Dover, 1970.
[6] (1999) Levi's Online Store: Men's Jackets. Levi Strauss & Co.. [Online]. Available: http://store.us.levi.com
[7] (1999) Mattel's Online Privacy Policy, Legal Terms/Conditions. Mattell, Inc, East Aurora, NY. [Online]. Available: http://www.barbie.com
[8] J. L. Synge and A. Schild, *Tensor Calculus*. New York: Dover, 1949.

Single-Channel Speech Enhancement in Variable Noise-Level Environment

Chin-Teng Lin

***Abstract*—This paper discusses the problem of single-channel speech enhancement in variable noise-level environment. Commonly used, single-channel subtractive-type speech enhancement algorithms always assume that the background noise level is fixed or slowly varying. In fact, the background noise level may vary quickly. This condition usually results in wrong speech/noise detection and wrong speech enhancement process. In order to solve this problem, we propose a new subtractive-type speech enhancement scheme in this paper. This new enhancement scheme uses the RTF (refined time-frequency parameter)-based RSONFIN (recurrent self-organizing neural fuzzy inference network) algorithm we developed previously to detect the word boundaries in the condition of variable background noise level. In addition, a new parameter (MiFre) is proposed to estimate the varying background noise level. Based on this parameter, the noise level information used for subtractive-type speech enhancement can be estimated not only during speech pauses, but also during speech segments. This new subtractive-type enhancement scheme has been tested and found to perform well, not only in variable background noise level condition, but also in fixed background noise level condition.**

***Index Terms*—Filter bank, noise estimation, recurrent network, time-frequency analysis, word boundary detection.**

I. INTRODUCTION

Background noise acoustically added to speech can decrease the performance of digital signal processing used for applications such as speech compression and recognition. The main objective of speech enhancement is to reduce the influence of noise [1]. Adaptive noise cancellation (ANC) [2]–[4] uses a secondary input to measure the noise source such that the estimated noise can then be subtracted from the primary channel resulting in the desired signal. A spectral subtraction method [5] which does not need a second microphone can also reduce the influence of noise. In this method, noise magnitude spectrum is estimated during speech pauses, and it is subtracted from the noisy speech magnitude spectrum in order to estimate the clean speech. A method based on nonlinear spectral subtraction is presented [6], [7], and it needs to estimate the signal-to-noise ratio (SNR). Gurgen and Chen [8] performed speech enhancement based on Fourier–Bessel coefficients of speech and noise signals. Jensen and Hansen [9] proposed a sinusoidal model based algorithm for enhancement of speech degraded by additive broad-band noise.

An important problem in subtractive-type speech enhancement is to detect the presence of speech in noisy environment. The above single-channel speech enhancement algorithms always require that the background noise level is fixed or slowly varying in order to correctly detect the presence of speech in noisy environment, but the background noise level may vary quickly in real world. The speech enhancement method proposed by Sameti *et al.* [10] contains a noise adaptation algorithm which can cope with noise level variation as well as different noise types. The method proposed in [11] updates the noise estimation during speech pauses in order to calculate the masking threshold

Manuscript received September 23, 2000. This paper was recommended by Associate Editor M. S. Obaidat.

The author is 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).

Digital Object Identifier 10.1109/TSMCA.2003.811115

correctly. Logan and Robinson [12] modeled the speech and noise statistics using autoregressive hidden Markov models for speech enhancement. Rezayee and Gazor [13] proposed an adaptive tracking algorithm for enhancement of speech degraded by colored additive interference. However, the detection algorithm used in these schemes is not reliable in a nonstationary noise environment. In many applications, the environment is further complicated by nonstationary backgrounds, where there may exist concurrent noises, due to movements of desks, door slams, etc. This condition usually results in incorrect speech/noise detection of speech signal, and then results in wrong speech enhancement process.

The problem of detecting the presence of speech in noisy environment was also attacked in robust word boundary detection algorithms [14]–[17]. 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 environment, even if more complex decision strategies are used [18]. Up to date, several other parameters were proposed such as, linear prediction coefficient (LPC), linear prediction error energy [19], [20], pitch information [21], and time-frequency (TF) parameter [18]. However, these parameters still cannot be adapted to variable-level background noise well.

In this paper, we focus on the problem of single-channel subtractive-type speech enhancement in the variable-level noise condition. To avoid the previous problems, we use the RTF-based RSONFIN algorithm developed by us [22]. Since the RTF parameter can extract useful frequency energy and the RSONFIN [23], [24] can process the temporal relations, this RTF-based RSONFIN algorithm can detect the word boundaries well in the condition of variable background noise level. This new algorithm has been tested and found to perform well not only in variable background noise level condition, but also in fixed background noise level condition. Another problem is to estimate the noise information in the speech segments. Commonly used, single-channel subtractive-type speech enhancement algorithms estimate the noise magnitude spectrum during speech pauses. Since the noise magnitude spectrum may vary in the speech segments, we should also estimate it in the speech segments. We propose a *minimum-frequency-energy* (MiFre) parameter which can estimate the varying background noise level by adaptively choosing the proper bands from the mel-scale frequency bank. Based on this parameter, the background noise information used for subtractive-type speech enhancement can be estimated, not only during speech pauses, but also during speech segments.

This paper is organized as follows. The new MiFre parameter is derived in Section II. In Section III, we introduce the RTF-based RSONFIN algorithm, and propose a new subtractive-type speech enhancement scheme. This enhancement scheme uses the new MiFre parameter and the RTF-based RSONFIN word boundary detection algorithm. In addition, some experiments are done in this section. Finally, the conclusions of our work are summarized in Section IV.

II. Minimum Frequency Energy

In this section, we propose a *minimum-frequency-energy* (MiFre) parameter which can estimate the varying background noise level by adaptively choosing the proper bands from the mel-scale frequency bank. Based on this parameter, the background noise level can be estimated, not only during speech pauses, but also during speech segments.

A. Auditory-Based Mel-Scale Filter Bank

There is an evidence from auditory psychophysics that the human ear perceives speech along a nonlinear scale in the frequency domain [25]. 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 normal frequency (Hz) is described by the following equation:

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

where mel is the mel-frequency scale and f is in Hz. The filter bank is then designed according to the mel scale, where the filters of 20 bands are approximated by simulating 20 triangular band-pass 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, $f(i,\ k)$, also represents the weighting factor of the frequency energy at the kth point of the ith band.

With this mel-scale frequency bank given in Fig. 1(a), 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 n–th point of the m–th frame. We first find the spectrum, $x_{\text{freq}}(m,\ k)$, of this signal by Discrete Fourier Transform (128-point DFT).

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

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

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 m–th frame.

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

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. We further smooth it by using a three-point median filter to get $\hat{x}(m,\ i)$.

$$\begin{aligned}\hat{x}(m,\ i) &= \text{SMOOTHING}(X(m,\ i)) \\ &= \frac{x(m-1,i) + x(m,i) + x(m+1,i)}{3}.\end{aligned} \tag{5}$$

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:

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

B. Background Noise Level Estimation

To estimate the background noise level, we need a parameter to stand for the amount of word signal information of each band. Before we propose a way to estimate the background noise level, we first

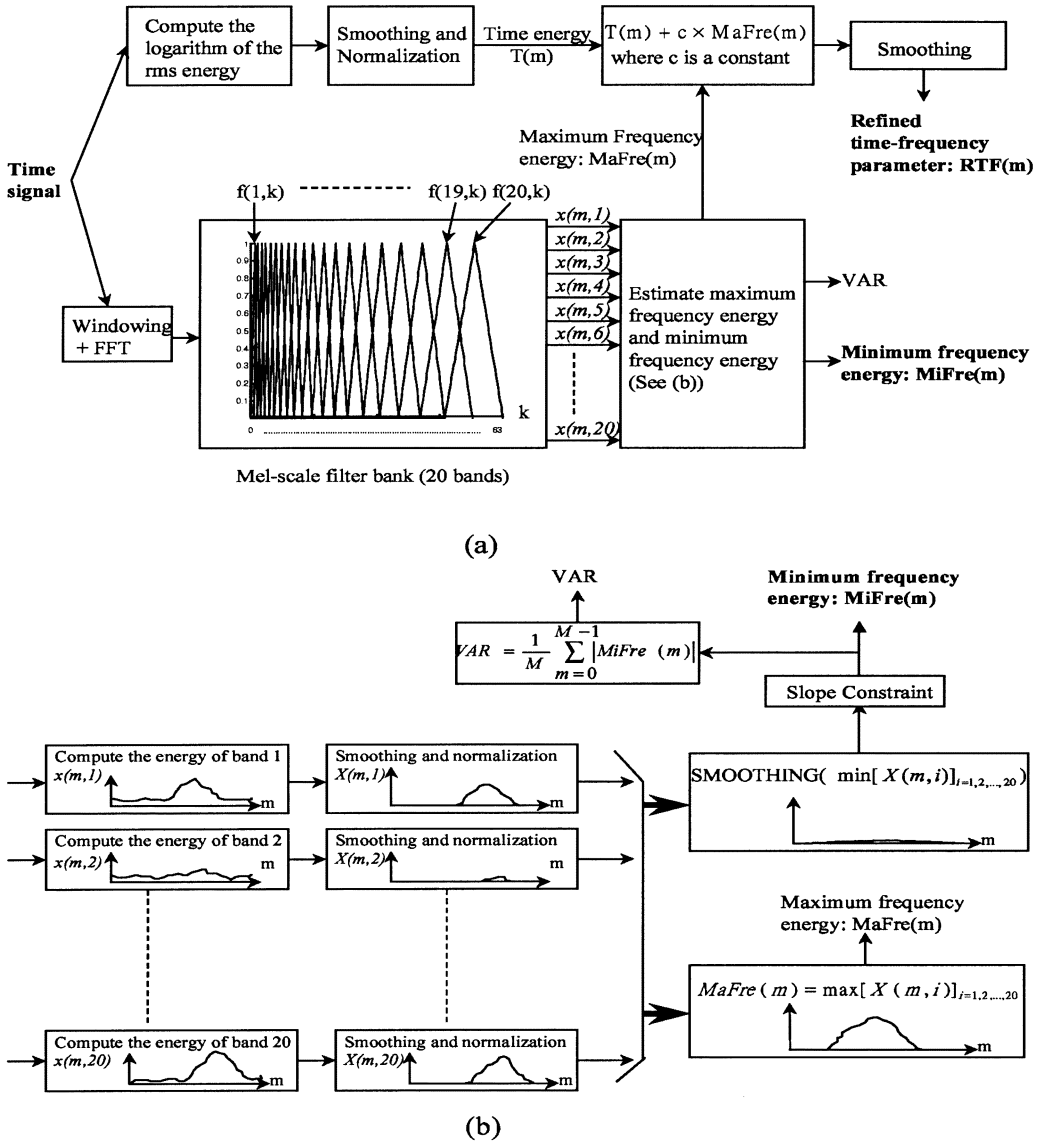

Fig. 1. (a) Flowchart for computing the RTF(m) and MiFre(m) parameters. (b) Procedure for estimating the maximum frequency energy and minimum frequency energy in (a).

make some observations on the effect of additive noise on each frequency band. In Fig. 2(a), we try to add white noise (0 dB) to the clean speech signal to see the effects of adding white noise on each band. For illustration, the smoothed and normalized frequency energies of a speech signal, $X(m, i)$ in (6), for 20 bands ($i = 1, 2, \ldots, 20$) and 166 frames ($m = 0, 1, \ldots, 165$) are shown in Fig. 2(b) and (c). We find that the energy of the first word signal ($m = 30, 41, \ldots, 50$) mainly focuses on the 5th band. Since the 8th$\sim$20th bands are seriously corrupted by the additive white noise, these bands have little information of word signal. In order to estimate the background noise of the first word signal segment correctly, we shall adopt the bands between band indexes 8 and 20 to estimate the white noise level. In addition, the energy of the second word signal ($m = 70, 71, \ldots, 90$) mainly focuses on the 7th band, and the energy of the third word signal ($m = 120, 121, \ldots, 140$) mainly focuses on the 9th band. Hence, we cannot adopt the 7th and 9th bands in estimating the noise levels in the second and third word signal segments. Obviously, some bands have small frequency energy $X(m, i)$ and should be adopted to estimate the background noise level. However, these small-energy bands may change under different word signals and noise conditions. This is because different word signals and noise focus their frequency energy on different bands; some focus on low frequency bands, and others on high frequency bands.

Based on the above discussion and illustrations, we propose a new parameter, MiFre, to estimate the variation of background noise level and reduce the effect of word signal. We adopt the minimum $X(m, i)$ and smooth it by a three-point median filter to be $\hat{X}(m)$:

$$\hat{X}(m) = \text{SMOOTHING}(\min[X(m, i)]_{i=1, 2, \ldots, 20}). \tag{7}$$

Finally, we put the slope constraint on $\hat{X}(m)$ to get the $\text{MiFre}(m)$ parameter to stand for the background noise level:

$$\text{MiFre}(m) = \text{Slope-Constraint}(\hat{X}(m)), \tag{8}$$

$$= \begin{cases} \dfrac{m}{30} + 5, & \text{if } \hat{X}(m) > \dfrac{m}{30} + 5 \\ \hat{X}(m), & \text{if } \dfrac{m}{30} + 5 \geq \hat{X}(m) \geq -\dfrac{m}{30} - 5 \\ -\dfrac{m}{30} - 5, & \text{if } -\dfrac{m}{30} - 5 > \hat{X}(m). \end{cases} \tag{9}$$

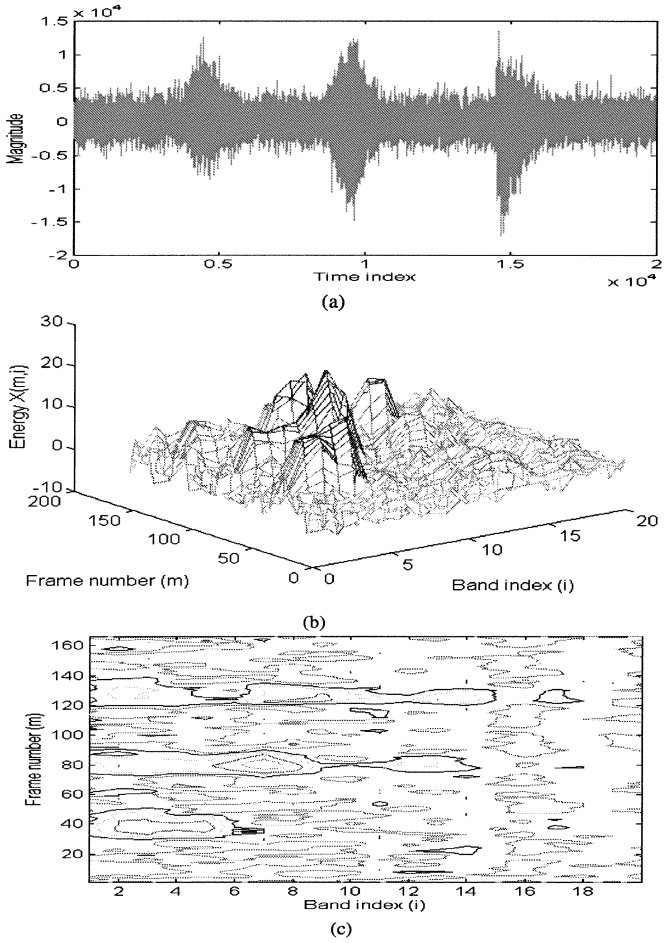

Fig. 2. (a) Speech waveform recorded in additive white noise of 0 dB. (b) Smoothed and normalized frequency energies, $\boldsymbol{X(m,\ i)}$, on 20 frequency bands. (c) The contour of (b).

If the values of $\hat{X}(m)$ increase or decrease largely, the slope constraint will reduce the variations of $\hat{X}(m)$.

The detailed procedure to calculate the MiFre parameter is illustrated in Fig. 1, and the RTF parameter in this figure is used for the RTF-based RSONFIN algorithm we developed [22] as described in the next section. In addition, the procedure for estimating the maximum frequency energy and minimum frequency energy in Fig. 1(a) is shown in Fig. 1(b). In order to see the effect of MiFre parameter, we make a test as follows. The speech signal with additive increasing-level white noise ($\mathrm{SNR} = 10$ dB) is shown in Fig. 3(a), and 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). According to (9), the values of MiFre parameter can be obtained and shown in Fig. 3(c). The root-mean-square energy of the background noise is shown in Fig. 3(d). We can easily find that the values of MiFre parameter in Fig. 3(c) are increasing and do reflect the variations of background noise in Fig. 3(d).

III. New Speech Enhancement Algorithm

In this section, we propose a new speech enhancement scheme in variable background noise-level environment. This enhancement scheme uses MiFre parameter to estimate the varying background noise level, and uses the RTF-based RSONFIN algorithm to detect the word boundaries in the condition of variable background noise level.

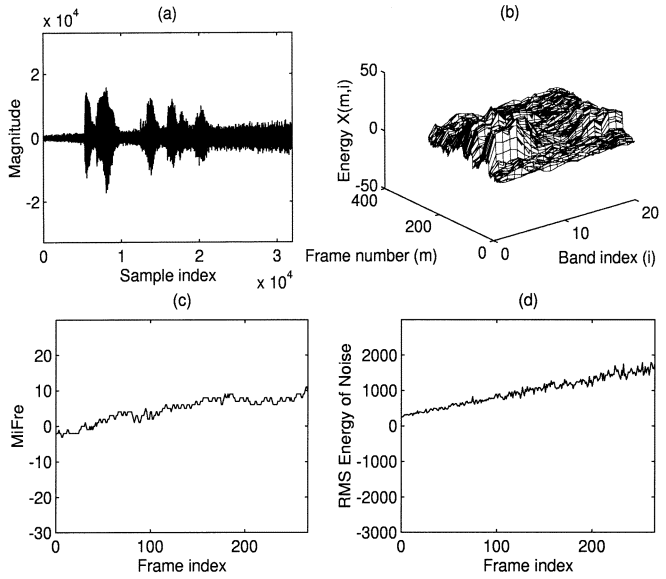

Fig. 3. (a) Speech signal with additive increasing-level white noise ($\mathbf{SNR} = \mathbf{10}$ dB). (b) Smoothed and normalized frequency energy, $\boldsymbol{X(m,\ i)}$, on 20 frequency bands. (c) Values of MiFre parameter. (d) Root-mean-square energy of the background noise.

A. RTF-Based RSONFIN Algorithm for Word Boundary Detection

The structure of the RSONFIN is shown in Fig. 4(a). With the learning ability of temporal relations, a procedure of using the RSONFIN for word boundary detection in variable background noise level condition is illustrated in Fig. 4(b). The input feature vector of the RSONFIN consists of the average of the logarithmic root-mean-square (rms) energy on the first five frames of recording interval (Noise_time), RTF parameter, and zero-crossing rate (ZCR). These three parameters in an input feature vector are obtained by analyzing a frame of a speech signal. Hence there are three (input) nodes in layer 1 of RSONFIN. Before entering the RSONFIN, the three input parameters are normalized to be in [0, 1]. For each input vector (corresponding to a frame), the output of RSONFIN indicates whether the corresponding frame is a word signal or noise. For this purpose, we used two (output) nodes in layer 5 of RSONFIN, where the output vector of (1, 0) standing for word signal, and (0, 1) for noise.

In the training process, the noisy speech waveform is sampled, and each frame is transformed into the desired input feature vector of RSONFIN (Noise_time, RTF parameter, and zero-crossing rate). These training vectors are classified as word signal or noise by using waveform, spectrum displays and audio output. Among these training vectors, some are from word sound category with the desired RSONFIN output vector being (1, 0), and the others are from noise category with the desired RSONFIN output vector being (0, 1).

The RSONFIN after training is ready for word boundary detection. As shown in Fig. 4(b), the outputs of RSONFIN are processed by a decoder. The decoder decodes the RSONFIN's output vector (1, 0) as value 100 standing for word signal, and (0, 1) as value 0 standing for noise. We observed that the decoding waveform (i.e., the outputs of the decoder) contained impulse noise sometimes. Hence, we let the output waveform of the decoder pass through a three-point median filter to eliminate the isolated "impulse" noise. Finally, we recognize the word-signal island as the part of the filtered waveform whose magnitude is greater than 30, and duration is long enough (by setting a threshold value). We then regard the parts of the original signal corresponding to the allocated word-signal island as the word signal, and the other ones as the background noise. The details of the RTF-based RSONFIN algorithm for word boundary detection can be found in [22].

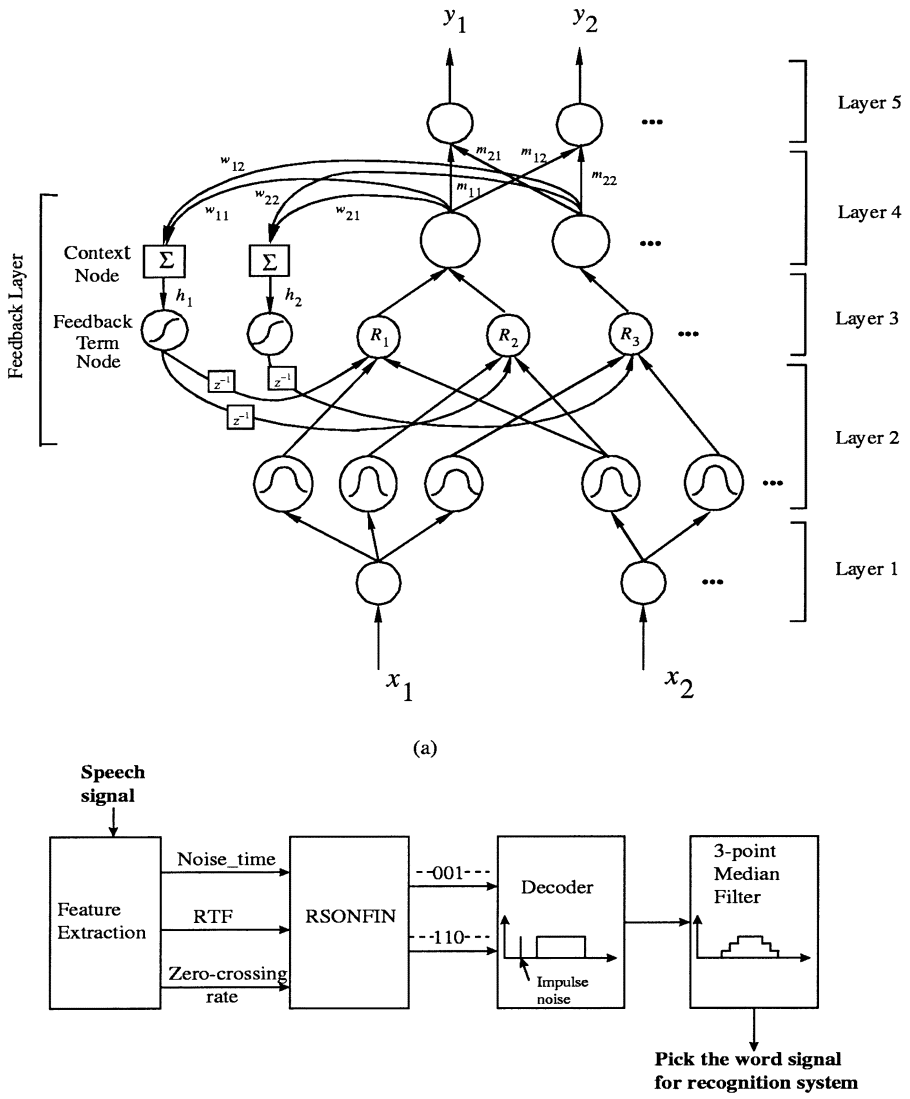

Fig. 4. (a) Structure of the Recurrent Self-Organizing Neural Fuzzy Inference Network (RSONFIN). (b) RTF-based RSONFIN algorithm for automatic word boundary detection.

B. New Speech Enhancement Scheme

The flowchart of the proposed speech enhancement scheme in variable background noise-level environment is shown in Fig. 5. Consider a speech signal $s(n)$ corrupted by additive noise $d(n)$.

$$y(n) = s(n) + d(n) \tag{10}$$

where the speech and noise signals are assumed to be uncorrelated. Taking the Fourier Transform of (10) gives

$$Y(e^{jw}) = S(e^{jw}) + D(e^{jw}). \tag{11}$$

We further smooth the magnitudes of $Y(e^{jw})$ by using a three-point median filter to get $|\overline{Y}(e^{jw})|$

$$|\overline{Y}_i(e^{jw})| = \frac{|Y_{i-1}(e^{jw})| + |Y_i(e^{jw})| + |Y_{i+1}(e^{jw})|}{3}, \tag{12}$$

where i means the i–th time window. The spectral magnitude $|\hat{Y}(e^{jw})|$ is obtained by subtracting the noise spectral magnitude estimate $|\hat{D}(e^{jw})|$ from the smoothed noisy speech spectral magnitude $|\overline{Y}(e^{jw})|$.

$$|\hat{Y}(e^{jw})| = |\overline{Y}(e^{jw})| - |\hat{D}(e^{jw})|. \tag{13}$$

Based on the RTF-based RSONFIN algorithm described in the last subsection, the noise spectral magnitude estimate $|\hat{D}(e^{jw})|$ can be updated reliably during speech pauses. Commonly used single-channel subtractive-type speech enhancement algorithms estimate the noise magnitude spectrum during speech pauses. However, since the noise magnitude spectrum may vary in the speech segments, we use the MiFre parameter to estimate it not only during speech pauses but also during speech segments as described in Section II. In addition, we define a parameter, VAR, to represent the sum of the MiFre values over all frames (see Fig. 1).

$$\mathrm{VAR} = \frac{\sum_{m=0}^{M-1} |\mathrm{MiFre}(m)|}{M}. \tag{14}$$

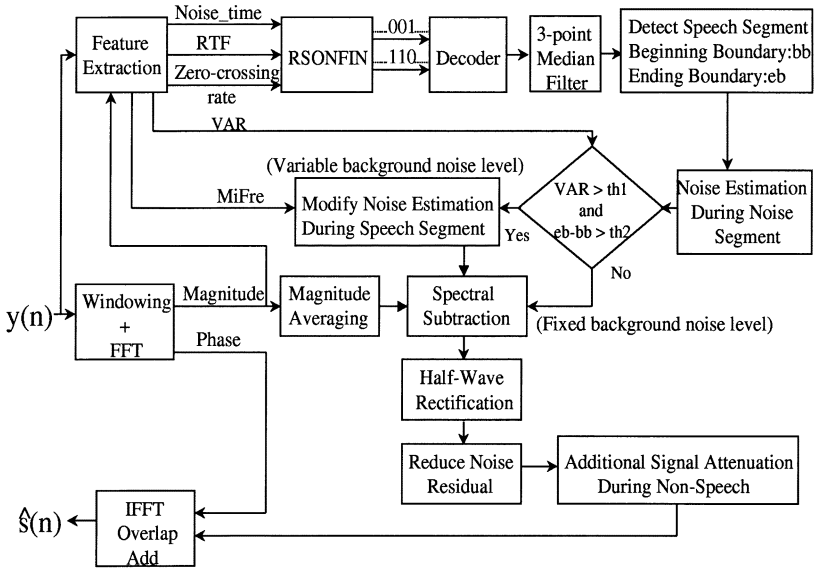

Fig. 5. Proposed speech enhancement scheme in variable noise-level environment.

This VAR parameter can indicate the average variation of background noise level.

Threshold $\mathrm{th1}$ in Fig. 5 is used to check whether the background noise level is fixed or variable. We set the beginning boundary of the speech segment to be "bb" and the ending boundary to be "eb." Threshold $\mathrm{th2}$ in Fig. 5 is used to check whether the speech segment is long enough. If $\mathrm{VAR} \leq \mathrm{th1}$, the variation of background noise level in the recording interval is small. If $(\mathrm{ee} - \mathrm{bb}) \leq \mathrm{th2}$, the MiFre values are not sufficient to stand for the variation of the background noise level in speech segment. In these two cases, the noise spectral magnitude estimate $|\hat{D}(e^{jw})|$ obtained during speech pauses will not be modified in the speech segment. However, if $\mathrm{VAR} > \mathrm{th1}$ and $(\mathrm{eb} - \mathrm{bb}) > \mathrm{th2}$, the variation of background noise level in the corresponding speech segment is large, and the MiFre values can stand for the variation of the background noise level in speech segment. In this case, the noise spectral magnitude estimate $|\hat{D}(e^{jw})|$ obtained during speech pauses should be modified in the speech segment as follows:

$$|\hat{D}_{\mathrm{modified}}(e^{jw})| = |\hat{D}(e^{jw})| \times \mathrm{weight} \tag{15}$$

$$\mathrm{weight} = 1 + \frac{\mathrm{MiFre}(m) - \mathrm{MiFre}(bb) - \mathrm{coef1}}{\mathrm{coef2}} \tag{16}$$

where, by trial and error, we choose $\mathrm{th1} = 5, \mathrm{th2} = 5400, \mathrm{coef1} = 2$ and $\mathrm{coef2} = 1500$ in our speech enhancement scheme. In this case, (13) should be modified accordingly as follows:

$$|\hat{Y}(e^{jw})| = |\overline{Y}(e^{jw})| - |\hat{D}_{\mathrm{modified}}(e^{jw})|. \tag{17}$$

To reduce the effect of noise, we apply half-wave rectification to $|\hat{Y}(e^{jw})|$; for each frequency w, where $|\hat{Y}(e^{jw})|$ obtained by (17) is less than zero, the output is set to zero.

$$|\hat{Y}_{half}(e^{jw})| = \begin{cases} |\hat{Y}(e^{jw})|, & \text{if } |\hat{Y}(e^{jw})| > 0 \\ 0, & \text{if } |\hat{Y}(e^{jw})| \leq 0. \end{cases} \tag{18}$$

In the next step, the methods of "reducing noise residual" and "additional signal attenuation" used by Boll [5] during nonspeech segments are implemented to get the final enhanced spectral magnitude $|\hat{S}(e^{jw})|$. In the process of reducing noise residual, the noise residual is suppressed by replacing its current value with its minimum value chosen from the adjacent analysis frames, and in the process of additional signal attenuation, the noise is attenuated by a fixed factor. Finally, we take the inverse fourier transform to get the enhanced speech signal in time domain.

C. Experiments

This section tests the performance of the proposed speech enhancement scheme. The sampling rate is 8 kHz, and the frame size is 240 samples (30 ms) with 50% overlap. Each speech signal covered by additive noise is a Mandarin speech sentence with length of 4 s, and there are totally 100 noisy sentences for testing. The added noise signals are from the noise database provided by the NATO Research Study Group on Speech Processing (RSG.10) NOISE-ROM-0 [26]. 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.

We first see the performance of the proposed scheme on a speech signal with additive increasing-level white noise ($\mathrm{SNR} = 10$ dB) in Fig. 6. Obviously, the noise in the rear part of recording interval is larger than the noise in the front part of recording interval. The noise makes the distinction between speech and background noise ambiguous. In Fig. 6(b), two speech segments are found, and the word boundaries detected by the RTF-based RSONFIN algorithm are shown by solid lines. Since the RTF parameter can extract useful frequency energy and the RSONFIN [23] can process the temporal relations, the RTF-based RSONFIN algorithm can find the variation of background noise level and detect correct speech segments in the increasing background noise level condition. For contrast, the enhanced speech signal produced by the new speech enhancement scheme without noise estimation during speech segments is shown in Fig. 6(c). Since the noise estimation is done only during speech pauses, the effect of additive increasing-level white noise is obvious in the rear part of the second speech segment. The enhanced speech signal produced by the new speech enhancement scheme with noise estimation during speech segments is shown in Fig. 6(d). Since the noise estimation is done not only during speech pauses but also during speech segments, the increasing-level white noise can be removed reasonably. The rear part of the second speech segment has no obvious noise component. This

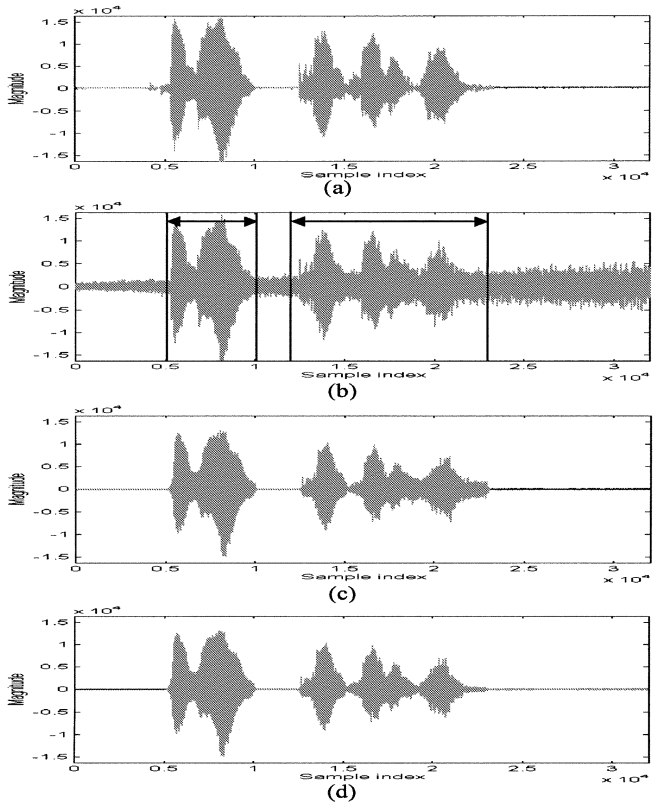

Fig. 6. (a) Original clean speech signal. (b) Speech signal with additive increasing-level white noise ($\mathbf{SNR} = \mathbf{10}$ dB). The word boundaries detected by the RTF-based RSONFIN algorithm are shown by solid lines. (c) Enhanced speech signal without noise estimation during speech segments. (d) Enhanced speech signal with noise estimation during speech segments.

observation demonstrates the efficiency of the proposed new speech enhancement scheme in variable background noise-level condition.

The amount of noise reduction in variable background noise level condition is measured by the objective evaluation:

$$\text{Input SNR} = 10\log \frac{\sum_{n=1}^{K} s^2(n)}{\sum_{n=1}^{K} d^2(n)} \tag{19}$$

$$\text{Output SNR} = 10\log \frac{\sum_{n=1}^{K} s^2(n)}{\sum_{n=1}^{K} [s(n) - \hat{s}(n)]^2} \tag{20}$$

where the "input SNR" is the SNR value of the input noisy speech signal standing for the amount of the additive noise, the "output SNR" is the SNR value of the output enhancement speech signal standing for the efficiency of the speech enhancement scheme, K is the frame-length, $s(n)$ is the clean speech signal, $d(n)$ is the additive noise, and $\hat{s}(n)$ is the enhanced speech signal. In our test, the input SNR values are from 0 to 15 dB, and the output SNR values calculated by (20) are shown in Fig. 7. This figure shows that the proposed scheme with noise estimation during speech segments produces the enhanced speech signals with higher SNR values at various input SNR values than that without noise estimation during speech segments.

IV. Conclusions

Two major characteristics of the new speech enhancement scheme proposed in this paper can be observed. 1) Since the RTF parameter can extract useful frequency information and the RSONFIN can recognize the temporal relations automatically and implicitly, the RTF-based RSONFIN algorithm can find the variation of background noise level and detect correct speech/noise segments in variable noise-level environment. The recurrent property of the RSONFIN makes it more suitable for dealing with temporal problems. 2) Since the MiFre parameter can estimate the varying background noise level, the background noise information required in our subtractive-type speech enhancement scheme can be estimated not only during speech pauses but also during speech segments. This new subtractive-type speech enhancement scheme has been tested and found to perform well not only in variable background noise level condition but also in fixed background noise level condition.

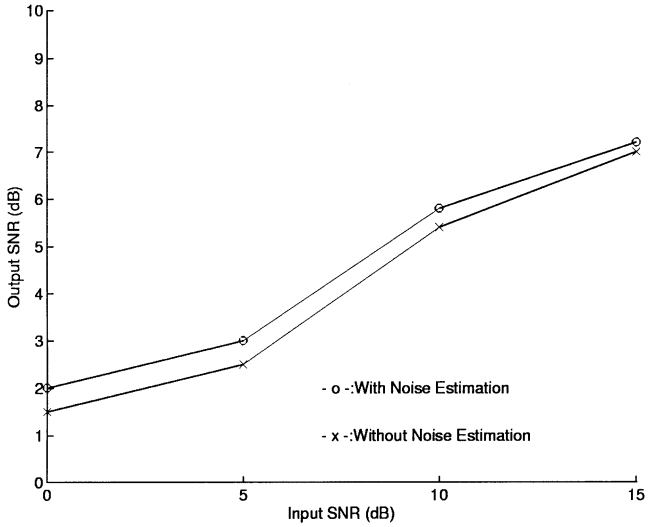

Fig. 7. Comparison of speech enhancement algorithm with noise estimation and that without noise estimation during speech segments in variable background noise level condition.

References

[1] J. S. Lim, *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] B. Widrow and S. D. Strarns, "Adaptive inference cancellation," in *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[3] W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 433–438, Nov. 1995.

[4] C. T. Lin and C. F. Juang, "An adaptive neural fuzzy filter and its applications," *IEEE Trans. Syst., Man, Cybern. B*, vol. 27, pp. 635–656, Aug. 1997.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–27, pp. 113–120, Feb. 1979.

[6] P. Lockwood and J. Boundy, "Experiments with a nonlinear spectral subtraction (NSS), hidden Markov models and the projection for robust speech recognition in cars," *Speech Commun.*, vol. 11, pp. 215–228, 1992.

[7] M. Lorber and R. Hoeldrich, "A combined approach for broadband noise reduction," in *IEEE ASSP Workshop*, 1997, pp. 1–4.

[8] F. Gurgen and C. S. Chen, "Speech enhancement by Fourier-Bessel coefficients of speech and noise," *Institute Elect. Eng. Proc. Comm., Speech Vision*, pt. 1, vol. 137, no. 5, pp. 290–294, Oct. 1990.

[9] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 731–740, Oct. 2001.

[10] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 445–455, Sept. 1998.

[11] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.
[12] B. Logan and T. Robinson, "Adaptive model-based speech enhancement," *Speech Commun.*, vol. 34, no. 4, pp. 351–368, July 2001.
[13] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.
[14] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297–315, 1975.
[15] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, pp. 777–785, June 1981.
[16] M. H. Savoji, "A robust algorithm for accurate endpoint of speech," *Speech Commun.*, vol. 8, pp. 45–60, 1989.
[17] B. Reaves, "Comments on an improved endpoint detector for isolated word recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 526–527, Mar. 1991.
[18] 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.
[19] 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.
[20] S. J. Kia and G. G. Coghill, "A mapping neural network and its application to voiced-unvoiced-silence classification," in *Proc. First New Zealand International Two-Stream Conference Artificial Neural Networks Expert Systems*, 1993, pp. 104–108.
[21] M. Hamada, Y. Takizawa, and T. Norimatsu, "A noise robust speech recognition," in *Int. Conf. Spoken Language Processing*, 1990, pp. 893–896.
[22] G. D. Wu and C. T. Lin, "A recurrent neural fuzzy network for word boundary detection in variable noise-level environments," *IEEE Trans. Syst., Man, Cybern. B*, vol. 31, pp. 84–97, Feb. 2000.
[23] C. F. Juang and C. T. Lin, "A recurrent self-organizing neural fuzzy inference network," *IEEE Trans. Neural Networks*, vol. 10, no. 4, pp. 828–845, July 1999.
[24] C. T. Lin and C. S. G. Lee, *Neural Fuzzy Systems: A Neural-Fuzzy Synergism to Intelligent Systems*. Englewood Cliffs, NJ: Prentice-Hall, May 1996.
[25] J. B. Allen, "Cochlear modeling," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 2, pp. 3–29, 1985.
[26] 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.