CROSSCORRELATION COMPENSATED WIENER FILTER FOR SPEECH ENHANCEMENT

Lutfa Akter and Md. Kamrul Hasan

Department of Electrical and Electronic Engineering,
Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh

ABSTRACT

The focus of this paper is on improving the performance of the Wiener filter in speech enhancement by formulating filter gain without neglecting the crosscorrelation between the speech signal and background noise. Comparative results with the conventional Wiener filter and other reported methods confirm the superiority of the proposed method.

1. INTRODUCTION

Speech enhancement is increasingly becoming an important topic of research due to the use of automatic speech processing systems in a variety of real world applications. Speech degraded by background noise adversely affects the performances of speech recognition and coding systems. To enhance the performance of speech processing systems several methods have been reported so far in the literature. The enhancement can be made using the Wiener filtering [1]-[2], spectral subtraction rules [3]-[6], thresholding [7]-[8] and Kalman filtering. Among them the Wiener and spectral subtraction type algorithms are widely used because of their low computational complexity and impressive performance. In general, using the family of spectral *subtraction-type* algorithms the enhanced speech spectrum is obtained by subtracting an average noise spectrum from the noisy speech spectrum or by multiplying the noisy spectrum with a gain function [4]. The phase of the noisy speech is kept unchanged since it is assumed that the phase distortion is not perceived by human ear. The main shortcoming of this method, however, is that it introduces musical noise in the enhanced speech.

It has been observed experimentally that the performance of Wiener filter (WF) is relatively better in terms of objective measures such as overall output SNR and average segmental SNR (AvgSegSNR) than those of the spectral subtraction techniques such as parametric power subtraction (PARA) method [5]. But the Log-Area-Ratio (LAR) measure is poor as compared to PARA method, for example.

This paper introduces a crosscorrelation compensated Wiener filter approach for speech enhancement. In our work, we formulate the Wiener filter without neglecting the crosscorrrelation term between the speech signal and background noise to improve its performance as compared to that of the conventional Wiener filter.

2. CONVENTIONAL WIENER FILTER GAIN

Let the clean speech, noise and the noisy speech signals in the time domain be denoted by $x(t)$, $d(t)$ and $y(t)$, respectively. If it is assumed that the noise is additive, $y(t)$ can be expressed as

$$y(t) = x(t) + d(t). \tag{1}$$

At the n-th frame and k-th frequency bin, the DCT domain representation of (1) is

$$Y_{n,k} = X_{n,k} + D_{n,k} \tag{2}$$

where $X_{n,k}$, $D_{n,k}$ and $Y_{n,k}$ are the clean speech, noise and noisy speech discrete cosine transform (DCT) coefficients, respectively. An estimate of the clean speech component, denoted as $\widehat{X}_{n,k}$, can be obtained as

$$\widehat{X}_{n,k} = W_{n,k} Y_{n,k} \tag{3}$$

where $W_{n,k}$ is the optimal filter gain for modifying the noisy speech component. An expression for optimal $W_{n,k}$ is derived in the minimum mean squared error (MMSE) sense by minimizing the cost function

$$J_w = E\{(\widehat{X}_{n,k} - X_{n,k})^2\}. \tag{4}$$

Using the assumption that $X_{n,k}$ and $D_{n,k}$ are zero-mean and uncorrelated real Gaussian random variables (i.e., $E\{X_{n,k} D_{n,k}\} = 0$), J_w reduces to

$$J_w = (1 - W_{n,k})^2 E\{X_{n,k}^2\} + W_{n,k}^2 E\{D_{n,k}^2\}. \tag{5}$$

Setting $\partial J_w / \partial W_{n,k} = 0$, we obtain

$$W_{n,k} = \frac{\xi_{n,k}}{\xi_{n,k} + 1} \tag{6}$$

where $\xi_{n,k} = E\{X_{n,k}^2\}/E\{D_{n,k}^2\}$. The a priori signal-to-noise ratio (SNR), $\xi_{n,k}$ is calculated according to the decision directed approach reported in [2]. In this paper, we denote $W_{n,k}$ in (6) as the conventional Wiener filter (CWF) gain.

3. IMPROVED WIENER FILTER GAIN

In the derivation of (6) it is assumed that the speech and noise DCT components are uncorrelated, i.e.,

$$E\{X_{n,k} D_{n,k}\} = 0. \tag{7}$$

If $E\{X_{n,k}D_{n,k}\} \neq 0$, the cost function that the Wiener filter minimizes is

$$\begin{aligned}\tilde{J}_w &= (\tilde{W}_{n,k}-1)^2 E\{X_{n,k}^2\} \\ &\quad +2\tilde{W}_{n,k}(\tilde{W}_{n,k}-1)E\{X_{n,k}D_{n,k}\} \\ &\quad +(\tilde{W}_{n,k})^2 E\{D_{n,k}^2\} \end{aligned} \tag{8}$$

where $\tilde{W}_{n,k}$ denotes the crosscorrelation compensated Wiener gain. The proposed Wiener filter will be termed as crosscorrelation compensated Wiener filter (CCWF). Differentiating $\tilde{J}_w$ with respect to $\tilde{W}_{n,k}$ and equating to zero yields

$$\tilde{W}_{n,k} = \frac{\xi_{n,k} + \frac{E\{X_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}}}{\xi_{n,k}+1+2\frac{E\{X_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}}}. \tag{9}$$

A method for estimating $E\{X_{n,k}D_{n,k}\}$ from the noisy speech is required to compute (9) in a practical system. We can write

$$\begin{aligned}E\{Y_{n,k}D_{n,k}\} &= E\{(X_{n,k}+D_{n,k})D_{n,k}\} \\ &= E\{X_{n,k}D_{n,k}\} + E\{D_{n,k}^2\}. \end{aligned} \tag{10}$$

Dividing (10) by $E\{D_{n,k}^2\}$, we obtain

$$\frac{E\{Y_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}} = \frac{E\{X_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}} + 1. \tag{11}$$

Rearranging (11) gives

$$\frac{E\{X_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}} = \frac{E\{Y_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}} - 1 = T_{n,k} - 1 \tag{12}$$

where

$$\begin{aligned}T_{n,k} &= \frac{E\{Y_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}} = \frac{E\{Y_{n,k}(Y_{n,k}-X_{n,k})\}}{E\{D_{n,k}^2\}} \\ &= \frac{E\{Y_{n,k}^2\}}{E\{D_{n,k}^2\}} - \frac{E\{Y_{n,k}X_{n,k}\}}{E\{D_{n,k}^2\}}. \end{aligned} \tag{13}$$

We employ a recursive scheme to estimate $T_{n,k}$:

$$\begin{aligned}\widehat{T}_{n,k} &= \beta_{n,k}\widehat{T}_{n-1,k} + (1-\beta_{n,k})\left[\frac{Y_{n,k}^2}{E\{D_{n,k}^2\}} - \frac{Y_{n,k}X_{n,k}}{E\{D_{n,k}^2\}}\right] \\ &= \beta_{n,k}\widehat{T}_{n-1,k} + (1-\beta_{n,k})\gamma_{n,k}\chi_{n,k} \end{aligned} \tag{14}$$

where $\beta_{n,k}$ $(0 \leq \beta_{n,k} \leq 1)$ is an averaging parameter, *a posteriori* SNR $\gamma_{n,k} = Y_{n,k}^2/E\{D_{n,k}^2\}$ and $\chi_{n,k} = 1 - X_{n,k}/Y_{n,k}$. Substituting (12) into (9), we obtain a crosscorrelation accounted expression for the gain function of the Wiener filter as

$$\tilde{W}_{n,k} = \frac{\xi_{n,k}+T_{n,k}-1}{\xi_{n,k}+1+2T_{n,k}-2} = \frac{\xi_{n,k}-1+T_{n,k}}{\xi_{n,k}-1+2T_{n,k}}. \tag{15}$$

It is evident that now $\tilde{W}_{n,k}$ is not only controlled by $\xi_{n,k}$, as for $W_{n,k}$, but also by $T_{n,k}$. It is also interesting to note that (6) and (15) become identical for $T_{n,k} = 1$.

4. OPTIMAL ESTIMATION OF $\beta_{N,K}$

Here we propose an MMSE estimator for $\beta_{n,k}$ which minimizes the conditional cost function

$$J_\beta = E\left\{(\widehat{T}_{n,k} - T_{n,k})^2 \mid \widehat{T}_{n-1,k}\right\}. \tag{16}$$

The cost function is formulated such that $\widehat{T}_{n,k}$ approximates the true $T_{n,k}$ in the MMSE sense for given $\widehat{T}_{n-1,k}$. Substituting (14) into (16), we obtain

$$\begin{aligned}J_\beta &= E\Big\{\Big(\beta_{n,k}^2\widehat{T}_{n-1,k}^2 + (1-\beta_{n,k})^2\gamma_{n,k}^2\chi_{n,k}^2 \\ &\quad +2\beta_{n,k}(1-\beta_{n,k})\widehat{T}_{n-1,k}\gamma_{n,k}\chi_{n,k} - 2T_{n,k} \\ &\quad \times [\beta_{n,k}\widehat{T}_{n-1,k} + (1-\beta_{n,k})\gamma_{n,k}\chi_{n,k}] + T_{n,k}^2\Big) \\ &\quad \mid \widehat{T}_{n-1,k}\Big\} \\ &= \beta_{n,k}^2\widehat{T}_{n-1,k}^2 + (1-\beta_{n,k})^2 E\left\{\gamma_{n,k}^2\right\}\chi_{n,k}^2 \\ &\quad +2\beta_{n,k}(1-\beta_{n,k})\widehat{T}_{n-1,k}E\left\{\gamma_{n,k}\right\}\chi_{n,k} \\ &\quad -2\beta_{n,k}T_{n,k}\widehat{T}_{n-1,k} \\ &\quad -2(1-\beta_{n,k})T_{n,k}E\left\{\gamma_{n,k}\right\}\chi_{n,k} + T_{n,k}^2. \end{aligned} \tag{17}$$

Using the definition of *a posteriori* SNR given in [4], we can write

$$E\{\gamma_{n,k}^2\} = \frac{E\{Y_{n,k}^4\}}{E\{D_{n,k}^2\}^2} \tag{18}$$

and

$$E\{\gamma_{n,k}\} = \frac{E\{Y_{n,k}^2\}}{E\{D_{n,k}^2\}}. \tag{19}$$

Using (2) , we obtain

$$\begin{aligned}E\{Y_{n,k}^4\} &= E\{X_{n,k}^4\} + E\{D_{n,k}^4\} + 6E\{X_{n,k}^2\}E\{D_{n,k}^2\} \\ &\quad +4E\{X_{n,k}D_{n,k}\}(E\{X_{n,k}^2\} + E\{D_{n,k}^2\}). \end{aligned} \tag{20}$$

The fourth moments of $Y_{n,k}$, $X_{n,k}$ and $D_{n,k}$ are required to be calculated. Since $x(t)$ is a zero-mean real Gaussian random process, the probability density function of $X_{n,k}$ also follows the Gaussian distribution, i.e.,

$$p(X_{n,k}) = \frac{1}{\sqrt{2\pi E\{X_{n,k}^2\}}}\exp\left(-\frac{X_{n,k}^2}{2E\{X_{n,k}^2\}}\right). \tag{21}$$

Then by definition

$$\begin{aligned}E\{X_{n,k}^4\} &= \int_{-\infty}^{\infty} X_{n,k}^4 p(X_{n,k})dX_{n,k} \\ &= C\int_0^{\infty} X_{n,k}^4 \exp\left(-\frac{X_{n,k}^2}{2E\{X_{n,k}^2\}}\right)dX_{n,k} \end{aligned} \tag{22}$$

where C=$2/\sqrt{2\pi E\{X_{n,k}^2\}}$. Using the formula

$$\int_0^{\infty} x^m \exp(-ax^2)dx = \frac{\Gamma(\frac{m+1}{2})}{2a^{\frac{m+1}{2}}},\ a>0,\ m>-1 \tag{23}$$

we obtain

$$E\{X_{n,k}^4\} = 3E\{X_{n,k}^2\}^2. \quad (24)$$

Similarly,

$$E\{D_{n,k}^4\} = 3E\{D_{n,k}^2\}^2. \quad (25)$$

Using (2) and (19), we obtain

$$E\{\gamma_{n,k}\} = \xi_{n,k} + 2\frac{E\{X_{n,k}D_{n,k}\}}{E\{D_{n,k}^2\}} + 1. \quad (26)$$

Combining (12) and (26)

$$E\{\gamma_{n,k}\} = \xi_{n,k} + 2T_{n,k} - 1. \quad (27)$$

Dividing (20) by $E\{D_{n,k}^2\}^2$ and using (12), (24) and (25), we obtain

$$E\{\gamma_{n,k}^2\} = 3(\xi_{n,k}+1)^2 + 4(\xi_{n,k}+1)(T_{n,k}-1). \quad (28)$$

Using (27) and (28), we obtain

$$\begin{aligned} J_\beta &= \beta_{n,k}^2\widehat{T}_{n-1,k}^2 + (1-\beta_{n,k})^2[3(\xi_{n,k}+1)^2 + 4(\xi_{n,k}+1) \\ &\quad \times(T_{n,k}-1)]\chi_{n,k}^2 + 2(1-\beta_{n,k})(\beta_{n,k}\widehat{T}_{n-1,k} - T_{n,k}) \\ &\quad \times(\xi_{n,k}+2T_{n,k}-1)\chi_{n,k} - 2\beta_{n,k}T_{n,k}\widehat{T}_{n-1,k} + T_{n,k}^2. \end{aligned} \quad (29)$$

Now computing $\partial J_\beta/\partial\beta_{n,k}$ and setting it to zero, we obtain

$$\begin{aligned} \beta_{n,k}^{opt} &= \frac{A_{n,k}}{B_{n,k}} \\ \text{where } A_{n,k} &= [3(\xi_{n,k}+1)^2 + 4(\xi_{n,k}+1)(T_{n,k}-1)]\chi_{n,k}^2 \\ &\quad + T_{n,k}\widehat{T}_{n-1,k} - (T_{n,k} + \widehat{T}_{n-1,k}) \\ &\quad \times(\xi_{n,k}+2T_{n,k}-1)\chi_{n,k} \\ B_{n,k} &= \widehat{T}_{n-1,k}^2 + [3(\xi_{n,k}+1)^2 + 4(\xi_{n,k}+1) \\ &\quad \times(T_{n,k}-1)]\chi_{n,k}^2 - 2\widehat{T}_{n-1,k} \\ &\quad \times(\xi_{n,k}+2T_{n,k}-1)\chi_{n,k}. \end{aligned}$$

The parameters $\chi_{n,k} = 1 - X_{n,k}/Y_{n,k}$ and hence $T_{n,k}$ cannot be directly computed as $X_{n,k}$ is unobservable. However, $\chi_{n,k}$ may be replaced by $1 - W_{n,k}$. Whilst $T_{n,k}$ can be replaced by its instantaneous estimate $\bar{T}_{n,k} = \gamma_{n,k}\chi_{n,k}$ following (13).

5. PERFORMANCE EVALUATION AND DISCUSSION

The effectiveness of the proposed CCWF for speech enhancement is evaluated using 5 male and 5 female utterances taken from the TIMIT database. The speech samples were downsampled to 8 KHz. Two different noises, e.g. 'white' and 'babble' were taken from the NOISEX database to corrupt the speech signals. A frame size of 16 ms was used and 256 point DCT was taken on each frame. The overlap-add method with 50% overlap was used for speech decomposition. The expressions of $\beta_{n,k}$ and $T_{n,k}$ used in the simulation are given by

$$\beta_{n,k} = \min\left\{1, \frac{A_{n,k}}{B_{n,k}}\right\} \quad (30)$$

$$\widehat{T}_{n,k} = \max\left\{1.05, \beta_{n,k}\widehat{T}_{n-1,k} + (1-\beta_{n,k})\gamma_{n,k}(1-W_{n,k})\right\}. \quad (31)$$

The 'min' function in (30) is used to constrain estimate of $\beta_{n,k}$ in the range $0 \le \beta_{n,k} \le 1$. The 'max' function in (31) is used to limit estimate of $\widehat{T}_{n,k}$ go below a preassigned value. In order to achieve performance at least equivalent to that of the CWF in case of poor estimation accuracy of $\widehat{T}_{n,k}$, we have chosen the minimum value of $\widehat{T}_{n,k}$ to be slightly greater than 1.

The problem of crosscorrelation between the speech and noise arises only in the case of simultaneous presence of these two quantities in a frame. To implement this idea, we divide the speech frames into signal- and noise-dominant subframes as reported in [8]. The CCWF is used for the signal dominant subframes. In the noise-dominant subframes we ignore crosscorrelation between speech and noise. Therefore, the CWF is optimal for such frames. In these frames we set $\widehat{T}_{n,k} = 1$ so that CCWF performs identically to CWF. To observe the difference in the CCWF and CWF gains, we plot these quantities for an arbitrary signal-dominant frame in Fig. 1. It can be seen that the CCWF gain is larger than the CWF gain implying less attenuation of the noisy speech component in the signal dominant subframes. This in turn means that CCWF is introducing less processing distortion in the enhanced speech.

Comparative performance results of the proposed CCWF with the CWF and PARA are presented in Fig. 2. The results show improvement in AvgSegSNR and overall SNR as compared to those of the CWF and PARA for the proposed scheme for a wide range of input SNRs, e.g. -5 dB to 20 dB. Both the proposed CCWF and PARA show superior performance in terms of the LAR measure than that of the CWF.

Speech enhancement results obtained by different methods in the time and frequency domain are also presented in Figs. 3-7. In this experiment, the female utterance "Don't ask me to carry an oily rag like that" was used. The degraded speech is obtained by adding white noise with it. As expected, the proposed CCWF produces lower residual noise and noticeably less speech distortion in some speech segments.

6. CONCLUSION

This paper has dealt with a single speech enhancement technique in the DCT-domain. The conventional Wiener filter (CWF) has been reformulated without neglecting the crosscorrelation that exists between the speech signal and background noise. Comparative results show the superiority of the proposed crosscorrelation compensated Wiener filter (CCWF) over the CWF in terms of several objective speech quality measures.

7. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Speech Audio Processing*, vol. ASSP-32, pp. 1109-1121, 1984.

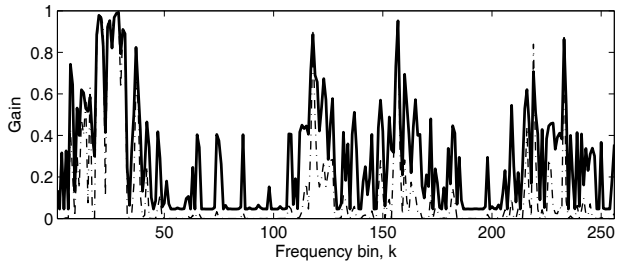

Fig. 1. Plot of instantaneous gain vs. frequency bin for a signal-dominant frame where (−.−) for CWF and (−) for CCWF.

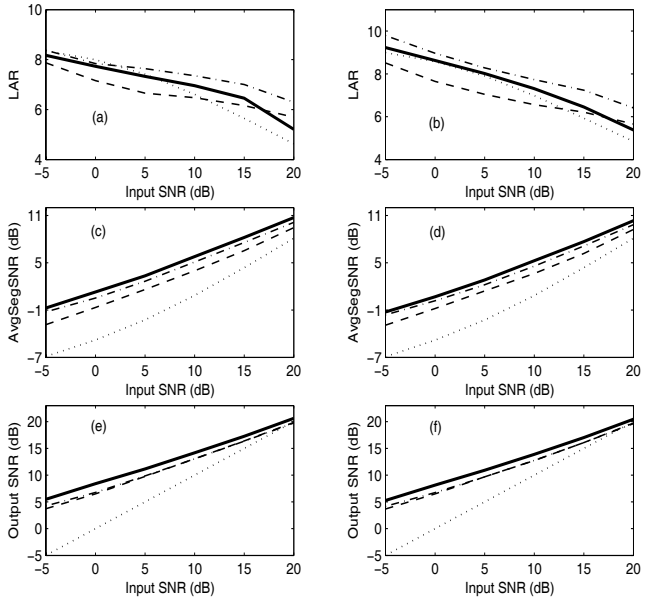

Fig. 2. Average objective quality measures with input SNR where (...) for degraded, (−.−) for CWF, (−−) for PARA and (−) for CCWF ((a), (c), (e): white noise; (b), (d), (f): babble noise).

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113-120, 1979.

[4] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation," *Proc. ICASSP*, pp. 629-632, 1996.

[5] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 328-337, 1998.

[6] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Communication*, vol. 24, pp. 249-257, 1998.

[7] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, Vol. 41, pp. 613-627, May 1995.

[8] S. Salahuddin, S. Z. Al Islam, M. K. Hasan, M. R. Khan, "Soft thresholding for DCT Speech Enhancement," *Electron. lett.*, vol. 38, no.24, pp. 1605-1607, 2002.

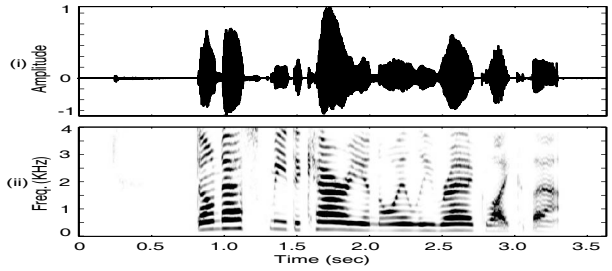

Fig. 3. Clean speech signal: (i) Time-domain; (ii) Spectrogram.

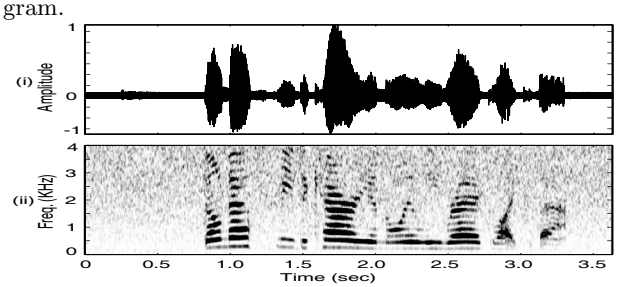

Fig. 4. Degraded speech signal: (i) Time-domain; (ii) Spectrogram.

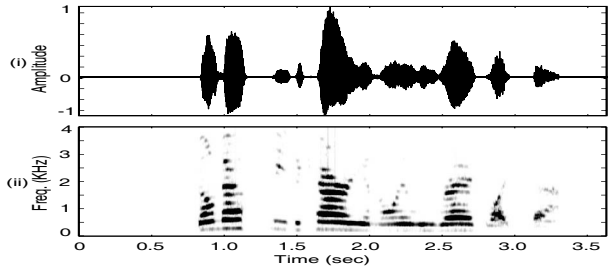

Fig. 5. Enhanced speech signal by the CWF: (i) Time-domain; (ii) Spectrogram.

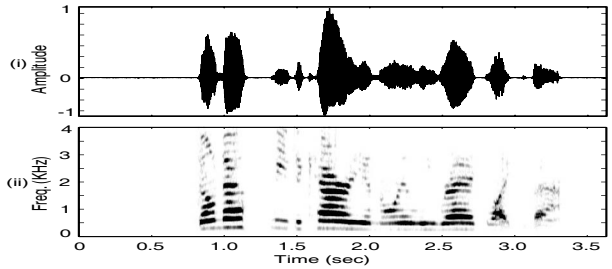

Fig. 6. Enhanced speech signal by the proposed CCWF: (i) Time-domain; (ii) Spectrogram.

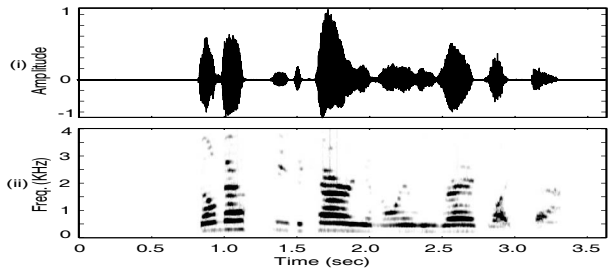

Fig. 7. Enhanced speech signal by PARA: (i) Time-domain; (ii) Spectrogram.