# An Improved Mel-Wiener Filter for Mel-LPC based Speech Recognition

*Md. Babul Islam, Hiroshi Matsumoto, Kazumasa Yamamoto*

Graduate School of Science and Technology, Shinshu University, Japan

{babul, matsu, kyama}@sp.shinshu-u.ac.jp

10.21437/Interspeech.2006-12

## Abstract

We previously proposed a Mel-Wiener filter to enhance Mel-LPC spectra in presence of additive noise. The proposed filter was estimated based on minimization of sum of square error on the linear frequency scale and efficiently implemented in the autocorrelation domain without denoising input speech. In the previously proposed system we segregated speech and noise using an energy based VAD and a very simple flooring technique were used for noise segment. In this present work, we improve the VAD using autoregressive (AR) model of noise and flooring technique as well. In addition, a lag window is applied to the estimated noise autocorrelation function to smooth the fine spectra of high order autocorrelation coefficients. As a result, substantial improvement is obtained over previous result.

**Index Terms**: Noisy speech recognition, Mel-Wiener filter, Mel-LPC analysis, Bilinear transformation, Aurora 2 database

## 1. Introduction

The performance of speech recognition systems has reached to the satisfactory level under controlled and matched training and recognition conditions. However, performance severely degrades when there is a mismatch between training and test conditions, caused for instance by additive noise. So, noise robustness is an important issue for ASR. There are many techniques to enhance the noisy speech signal based on the additive property of noise. The widely used methods to remove additive noise are spectral subtraction with many variants [1],[2] and Wiener filtering [3],[4].

As to front-end of speech recognition system, spectral analysis with auditory-like frequency resolution has been shown to be more effective for speech recognition [5],[6]. In filter-bank based systems, MFCC [5] is widely used. On the other hand, as an LP-based method, we previously proposed a simple and efficient time domain technique to estimate an all-pole model on the mel-frequency scale [7], [8], which is referred to as "Mel-LPC".

Therefore, speech enhancement in auditory-like frequency domain is also advantageous for speech recognition [3]. In the MFCC based system [3], a two-stage mel-warped Wiener filter was proposed, where the mel-warped transfer function was estimated and then converted to a time domain impulse response as Inverse Mel-DCT, which is computationally inefficient. So, this method is not suitable for Mel-LPC based front-end. From the practical viewpoint, it is appropriate to implement the Mel-Wiener filter in the time domain for Mel-LPC based speech analysis because Mel-LPC analysis is a time domain method. So, from this demand previously we proposed Mel-Wiener filter [10] combined with Mel-LPC for noise robust speech recognition. While a conventional Wiener filter might be applied to the frequency warped input signal, we took a novel approach to estimate the Mel-Wiener filter based on the error minimization on the linear frequency scale. The transfer function of the proposed filter is defined by using a first order all-pass filter instead of unit delay. For a given order of filter, the Mel-Wiener filter is expected to have a higher resolution in lower frequency region than that on the linear frequency scale.

In the previously proposed Mel-Wiener filter an energy based VAD was used and a very simple flooring technique was introduced for the noise segment. Though the proposed system outperforms the original two-stage mel-warped Wiener filter [3] for SNR higher than 0 dB, its overall performance was lower than that of ETSI Advanced Front-End (AFE) for Distributed Speech Recognition (DSR) [11]. So, in order to improve the performance of the previously proposed system, we improve VAD and flooring technique. In addition, a lag window is applied to the noise autocorrelation function to smooth the fine spectra of higher order autocorrelation coefficients. Finally, the same blind equalization as in ETSI AFE is applied to the cepstral coefficients to minimize the channel effect. Consequently, remarkable improvement has been achieved and it slightly outperforms the ETSI AFE for DSR [4],[12] as well.

The rest of this paper is organized as follows. In section 2, after overview of the Mel-LPC analysis, formulation of Mel-Wiener filter, estimation of crosscorrelation function between clean and noisy speech with flooring method are presented. The system overview, particularly, VAD, noise estimation, filtering and blind equalization are described in section 3. In section 4, analysis conditions and experimental results are presented. Finally, conclusion is presented in section 5.

## 2. Mel-Wiener Filter

### 2.1. Overview of Mel-LPC Analysis

Our intention is to use Mel-LPC analysis as front-end with Wiener filter. The frequency warped signal $\tilde{x}[n]$ $(n = 0, \ldots, \infty)$ obtained by the bilinear transformation [9] of a finite length windowed signal $x[n]$ $(n = 0, \ldots, N-1)$ is defined by

$$\tilde{X}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{z}^{-n} = X(z) = \sum_{n=0}^{N-1} x[n]z^{-n} \qquad (1)$$

where $\tilde{z}^{-1}$ is the first order all-pass filter,

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}} \qquad (2)$$

Now, the all-pole model on warped frequency scale is defined as

$$\tilde{H}_a(\tilde{z}) = \frac{\tilde{\sigma}_e}{1 + \sum_{k=1}^{p} \tilde{a}_k \tilde{z}^{-k}} \qquad (3)$$

where $\tilde{a}_k$ is the k-th mel-prediction coefficient and $\tilde{\sigma}_e^2$ is the residual energy [7]. To solve for $\tilde{a}_k$ and $\tilde{\sigma}_e$, the generalized autocorre-

lation coefficients of the input signal are required instead of autocorrelation coefficients in the traditional LP analysis [8].

### 2.2. Wiener Filter Formulation on Warped Frequency Scale

First, this section briefly describes the conventional Wiener filtering on a warped frequency domain, which can be implemented by applying the traditional Wiener filter to a frequency warped signal.

Now, we define a Wiener filter $\tilde{H}(\tilde{z})$ on the warped frequency scale as

$$\tilde{H}(\tilde{z}) = \sum_{n=0}^{p-1} \tilde{h}[n]\tilde{z}^{-n} \quad (4)$$

Then, the estimated clean speech $\hat{\tilde{s}}[n]$ is given by

$$\hat{\tilde{s}}[n] = \sum_{k=0}^{p-1} \tilde{h}[k]\tilde{x}[n-k] \quad (5)$$

Now, since the error signal $\tilde{e}[n] = \tilde{s}[n] - \hat{\tilde{s}}[n]$ is an infinite sequence, the sum of the square error is evaluated by

$$\xi\{\tilde{\mathbf{h}}\} = \sum_{n=0}^{\infty} \tilde{e}[n]^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\tilde{S}(e^{j\tilde{\lambda}}) - \tilde{H}(e^{j\tilde{\lambda}})\tilde{X}(e^{j\tilde{\lambda}})\right|^2 d\tilde{\lambda} \quad (6)$$

However, as shown in (1), since the bilinear transformation of a finite sequence results in an infinite sequence, the direct calculation of the autocorrelation coefficients of frequency-warped signal needs to truncate the signal, and thus, is not practical.

### 2.3. Wiener Filter Formulation on Linear Frequency Scale

Now, we define the transfer function of a frequency warped Wiener filter on $z$ domain by

$$\tilde{H}_w(\tilde{z}(z)) = \sum_{n=0}^{p-1} \tilde{h}_w[n]\tilde{z}^{-n} \quad (7)$$

Then, the estimated speech based on filter $\tilde{H}_w(\tilde{z}(z))$ is given by

$$\hat{s}_w[n] = \sum_{k=0}^{p-1} \tilde{h}_w[k]x_k[n] \quad (8)$$

where $x_k[n]$ is the output signal of $k$ cascaded all pass filter $\tilde{z}^{-k}$.

In the spectral domain, (8) can be rewritten as

$$\hat{S}_w(e^{j\lambda}) = \tilde{H}_w(e^{j\tilde{\lambda}})X(e^{j\lambda}) \quad (9)$$

Let $\hat{\tilde{S}}_w(e^{j\tilde{\lambda}})$ be the spectrum of the bilinear transformed signal of $\hat{s}_w[n]$. Since $\hat{S}_w(e^{j\lambda}) = \hat{\tilde{S}}_w(e^{j\tilde{\lambda}})$ from the definition of frequency warped signal as in (1), we have the following relation

$$\hat{\tilde{S}}_w(e^{j\tilde{\lambda}}) = \tilde{H}_w(e^{j\tilde{\lambda}})\tilde{X}(e^{j\tilde{\lambda}}) \quad (10)$$

This equation shows that $\tilde{H}_w(e^{j\tilde{\lambda}})$ is a linear filter to estimate the spectrum $\hat{\tilde{S}}_w(e^{j\tilde{\lambda}})$ from the input spectrum $\tilde{X}(e^{j\tilde{\lambda}})$ on the warped-frequency domain.

Now, the sum of the square error is given by

$$\xi\{\tilde{\mathbf{h}}_\mathbf{w}\} = \sum_{n=0}^{\infty}(s[n] - \hat{s}_w[n])^2 \quad (11)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\tilde{S}(e^{j\tilde{\lambda}}) - \tilde{H}_w(e^{j\tilde{\lambda}})\tilde{X}(e^{j\tilde{\lambda}})\right|^2 \cdot \left|\tilde{W}(e^{j\tilde{\lambda}})\right|^2 d\tilde{\lambda} \quad (12)$$

where $\frac{d\lambda}{d\tilde{\lambda}} = \left|\tilde{W}(e^{j\tilde{\lambda}})\right|^2$ with $\tilde{W}(\tilde{z}) = \frac{\sqrt{1-\alpha^2}}{1+\alpha\cdot\tilde{z}^{-1}}$.

Unlike (6), (12) shows that the error signal energy on the warped frequency domain is weighted by $\tilde{W}(e^{j\tilde{\lambda}})$.

The minimization of (11) with respect to $\{\tilde{h}_w[k]\}$ gives the following normal equations

$$\sum_{k=0}^{p-1} \tilde{\phi}_{xx}(m,k)\tilde{h}_w(k) = \tilde{\phi}_{sx}(0,m) \quad (m = 0,..,p-1), \quad (13)$$

where

$$\tilde{\phi}_{xx}(m,k) = \sum_{n=0}^{\infty} x_m[n]x_k[n] \quad (14)$$

and

$$\tilde{\phi}_{sx}(m,k) = \sum_{n=0}^{\infty} s_m[n]x_k[n] \quad (15)$$

In the warped frequency domain (14) and (15) can be rewritten as

$$\tilde{\phi}_{xx}(m,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\tilde{X}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}})\right|^2 \cdot e^{j(m-k)\tilde{\lambda}}\, d\tilde{\lambda} \quad (16)$$

and

$$\tilde{\phi}_{sx}(m,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{S}(e^{j\tilde{\lambda}})\tilde{X}^*(e^{j\tilde{\lambda}})\left|\tilde{W}(e^{j\tilde{\lambda}})\right|^2 \cdot e^{j(m-k)\tilde{\lambda}}\, d\tilde{\lambda} \quad (17)$$

Therefore, $\tilde{\phi}_{xx}(m,k)$ is the autocorrelation function of the signal $\tilde{x}_w[n]$ whose Fourier transform is equal to the frequency warped and frequency weighted spectrum $\tilde{X}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}})$. Similarly, $\tilde{\phi}_{sx}(m,k)$ is the crosscorrelation function between $\tilde{x}_{w,m}[n]$ and $\tilde{s}_{w,k}[n]$ whose Fourier transform is $\tilde{S}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}})$. We call $\tilde{\phi}_{xx}(m,k)$ and $\tilde{\phi}_{sx}(m,k)$ as the "generalized" autocorrelation and crosscorrelation functions, respectively. Fig. 1 defines the generalized crosscorrelation function.

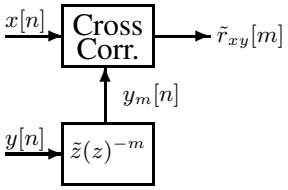


Figure 1: *Generalized crosscorrelation function.*

From (16) and (17), it should be noted that each of $\tilde{\phi}_{xx}(m,k)$ and $\tilde{\phi}_{sx}(m,k)$ is a function of the difference $(k-m)$. Thus, both functions are calculated from the sum of finite terms as

$$\tilde{\phi}_{xx}(m,k) = \tilde{r}_{xx}[k-m] = \sum_{n=0}^{N-1} x[n]x_{|k-m|}[n] \quad (18)$$

and

$$\tilde{\phi}_{sx}(0,k) = \tilde{r}_{sx}[k] = \sum_{n=0}^{N-1} s[n]x_k[n] \quad (19)$$

### 2.4. Approximation of crosscorrelation function $\tilde{\mathbf{r}}_{\mathbf{sx}}$

In practical situation, since both the speech and noise are unobservable, the crosscorrelation function between clean and noisy speech is approximated as

$$\tilde{r}_{sx}[m,t] \approx \begin{cases} \tilde{r}_{xx}[m,t] - s\cdot\hat{\tilde{r}}_{nn}[m,t]; & \text{if } lr[t] \geq v_2 \\ \gamma(\tilde{r}_{xx}[m,t] - s'\cdot\hat{\tilde{r}}_{nn}[m,t]) & \\ \quad + \tilde{r}_{fx}[m]; & \text{if } lr[t] < v_2 \end{cases} \quad (20)$$

where $s$ is a scaling factor, given by

$$s = \begin{cases} 1; & \text{if } \tilde{r}_{xx}[0] > \tilde{r}_{nn}[0] \\ 0.9\tilde{r}_{xx}[0]/\hat{\tilde{r}}_{nn}[0]; & \text{if } \tilde{r}_{xx}[0] \leq \hat{\tilde{r}}_{nn}[0] \end{cases} \quad (21)$$

$$\gamma = (\tilde{\sigma}_f/\tilde{\sigma}_x)^{(lr[t]-v_2)/(v_1-v_2)} \quad (22)$$

$\tilde{r}_{fx}[m]$ is the floored crosscorrelation function between $x[n]$ and a windowed random sequence whose rms value $\tilde{\sigma}_f$ is set to -30 dB from the maximum rms value of the input speech, $\tilde{\sigma}_x$ is the rms value of current frame, $lr[t]$ is the likelihood ratio given in (24), $s'$ is the ratio between residual energy of current frame to the residual energy of noise, which compensates the noise level between noise model and current noise frame, $v_1 = 0.1$ and $v_2 = 0.145$ are two experimentally tunable constants. The parameter $\gamma$ prevents the abrupt transition between non-speech and speech segment.

# 3. System Overview

### 3.1. Voice Activity Detection

Fig. 2 shows the block diagram of the proposed technique. The generalized autocorrelation of the current frame and the estimated generalized noise autocorrelation from the corresponding speech/silence decision of the VAD block are used in the Mel-Weiner filter design block to estimate the filter coefficients.

The voice activity detector (VAD) is based on Itakura-Saito distortion measure between autoregressive (AR) model [14] of noise and input speech signal. From initial 20 frames a noise model is created, i.e., the model is assumed to be $M$th order autoregressive with coefficients $\tilde{\mathbf{b}}^{\mathbf{t}} = [\tilde{b}_0\tilde{b}_1 \ldots \tilde{b}_M]$, where $\tilde{b}_0 = 1$. For the input frame $t$, $\tilde{r}_{xx}[m,t]$ is calculated to estimate Itakura-Saito distortion $d_{IS}[t]$ and likelihood ratio $lr[t]$ as follows:

$$d_{IS}[t] = \frac{1}{\sigma^2_{e\tilde{n}}}\delta(\tilde{x};\tilde{\mathbf{b}}) + \log\frac{\sigma^2_{e\tilde{n}}}{\sigma^2_{e\tilde{x}}} - 1 \quad (23)$$

and

$$lr[t] = \delta(\tilde{x};\tilde{\mathbf{b}}) - 1 \quad (24)$$

where

$$\delta(\tilde{x};\tilde{\mathbf{b}}) = R_{\tilde{b}}[0]\tilde{r}_{xx}[0] + 2\sum_{i=1}^{M} R_{\tilde{b}}[i]\tilde{r}_{xx}[i] \quad (25)$$

$\sigma^2_{e\tilde{n}}$ and $\sigma^2_{e\tilde{x}}$ are the residual energies of the estimated noise and current frame $t$, respectively, and $R_{\tilde{b}}[i]$ is the autocorrelation function of the AR coefficients.

Finally, $d_{IS}[t]$ is compared with a threshold value $\eta$, which is calculated as follows:

$$\eta = mean(d_{IS}[t]) + N_{th} \cdot std(d_{IS}[t]) \quad (26)$$

where $mean(d_{IS}[t])$ is the exponentially weighted average of $d_{IS}[t]$ and $std(d_{IS}[t])$ is the standard deviation of $d_{IS}[t]$ over all previous noise frames, and $N_{th}$ is a threshold factor with value of 0.01. For $d_{IS}[t] < \eta$, the frame $t$ is detected as noise, otherwise, speech frame.

### 3.2. Noise Estimation

If frame $t$ is detected as noise, a lag window of length 50 is applied on $\tilde{r}_{xx}[m]$. Now, the estimated generalized autocorrelation function of noise $\hat{\tilde{r}}_{nn}[m]$ is updated by accumulating $\tilde{r}_{xx}[m,t]$ as follows:

$$\hat{\tilde{r}}_{nn}[m,t] = \begin{cases} \beta\hat{\tilde{r}}_{nn}[m,t_p] + (1-\beta)\tilde{r}_{xx}[m,t]; \\ \qquad\qquad \text{if frame } t \text{ is silence} \\ \hat{\tilde{r}}_{nn}[m,t_p]; \quad \text{if frame } t \text{ is speech} \end{cases} \quad (27)$$

where $t_p$ is the previous noise frame and $\beta$ is the forgetting factor with value of 0.96.

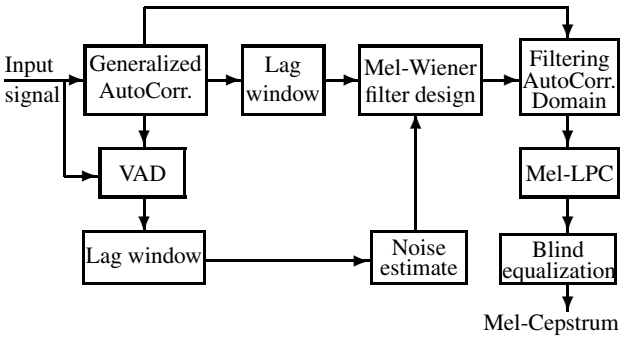


Figure 2: *Mel-LPC analysis with the proposed Wiener filter.*

### 3.3. Filtering in Autocorrelation Domain

Filtering is done in the autocorrelation domain to estimate the generalized autocorrelation function of the filtered speech $\hat{s}_w[n]$ as follows:

$$\hat{\tilde{r}}_{ss}[m] = \sum_{n=0}^{\infty} \hat{s}_w[n]\hat{s}_{w,m}[n] \quad (28)$$

$$= \sum_{k=-p+1}^{p-1} r_{\tilde{h}\tilde{h}}[k]\tilde{r}_{xx}[m-k] \quad (29)$$

where $\tilde{r}_{xx}[m]$ is the generalized autocorrelation function of the noisy speech, and $r_{\tilde{h}\tilde{h}}[m]$ is the autocorrelation function of $\tilde{h}_w[m]$. Finally, the Mel-prediction coefficients are obtained by Durbin's algorithm from $\tilde{r}_{ss}[m]$. Although the estimated model (3) includes the frequency weighting $\tilde{W}(e^{j\tilde{\lambda}})$, this is easily removed by inverse filtering in the generalized autocorrelation domain using $\{\tilde{W}(\tilde{z})\tilde{W}(\tilde{z}^{-1})\}^{-1}$.

### 3.4. Blind Equalization

Blind equalization is applied on the cepstral coefficients in order to minimize the channel effects. This technique is based on the least mean square algorithm, which minimizes the mean square error computed as a difference between the current and reference cepstrum [13]. We use the same algorithm as that implemented in AFE [11], and the long-term cepstrum of training clean speech is used as reference cepstrum.

# 4. Evaluation on Aurora 2

### 4.1. Experimental Setup

The proposed system was evaluated on Aurora 2 database. In this experiment, 12 Mel-LPC order was used. The speech signal without preemphasis was windowed using Hamming window of length 25 ms with 10 ms frame period. The frequency warping factor was set to 0.4. The HMM was trained on clean condition with 16 states per word and a mixture of 6 Gaussians per state. As front-end, 14 cepstral coefficients and their delta coefficients including $0^{th}$ terms were used.

### 4.2. Recognition Results

From Fig. 3, it has been shown that the highest word accuracy is attained at the order of 5. This optimum order is much lower than that of the all-pole model in Mel-LPC. This result from the fact that fine spectrum is not required in estimation of all-pole model. Consequently, the order of filter is set to 5.

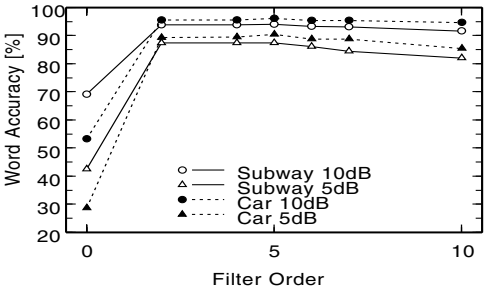


Figure 3: *Recognition accuracy as a function of filter order.*

The recognition accuracy for set A without and with proposed filter are given in Table 1 and Table 2, respectively. Table 3 shows the average accuracy for sets B and C. From tables, it is observed that Mel-Wiener filter improves the word accuracy except under match condition (clean). The average recognition accuracy over SNRs 20 to 0 dB are 87.53%, 85.99% and 84.11% for sets A, B and C, respectively.

Finally, we compare our current result with ETSI AFE, which is shown in Table 4. The recognition result for AFE is obtained from [12]. As compared to the result of AFE, our system slightly outperforms AFE. As shown in Table 4, the overall word accuracy for our proposed system is 86.23%, while the accuracy for AFE is 86.04%. In our previous implementation, we got 78.76% overall word accuracy [10]. So, the current result shows that substantial improvement has been achieved with the present technique. The computational cost of the proposed filter is 1.72 times of Mel-LPC analysis for addition/subtraction and 0.46 times for multiplication/division operations.

Table 1: *Recognition accuracy without Wiener filter for set A.*

| Noise | cln | 20dB | 15dB | 10dB | 5dB | 0dB | -5dB |
|---|---|---|---|---|---|---|---|
| Subway | 99.05 | 95.30 | 86.83 | 69.11 | 42.46 | 20.05 | 10.80 |
| Babble | 98.73 | 87.09 | 68.59 | 46.28 | 23.31 | 9.70 | 5.05 |
| Car | 98.78 | 93.20 | 78.94 | 53.00 | 28.36 | 12.08 | 7.46 |
| Exhib | 98.98 | 95.37 | 88.12 | 69.58 | 37.46 | 17.86 | 10.55 |
| Average | 98.89 | 92.74 | 80.62 | 59.50 | 32.90 | 14.93 | 8.48 |

Table 2: *Recognition accuracy with proposed filter for set A.*

| Noise | cln | 20dB | 15dB | 10dB | 5dB | 0dB | -5dB |
|---|---|---|---|---|---|---|---|
| Subway | 98.43 | 97.85 | 96.62 | 94.04 | 87.32 | 65.06 | 30.40 |
| Babble | 98.19 | 97.34 | 96.13 | 93.47 | 83.49 | 55.99 | 22.64 |
| Car | 98.63 | 97.94 | 97.64 | 95.91 | 90.07 | 68.03 | 25.20 |
| Exhib | 98.73 | 97.25 | 96.45 | 93.52 | 83.65 | 62.67 | 29.44 |
| Average | 98.50 | 97.60 | 96.71 | 94.24 | 86.14 | 62.94 | 26.92 |

## 5. Conclusion

An improved Mel-Wiener filter has been presented, which is directly estimated from the input signal on the linear frequency scale and effectively implemented in the autocorrelation domain incorporating with the Mel-LPC based spectral analysis. It has been shown that our proposed Mel-Wiener filter incorporating with the Mel-LPC can be used as front-end to improve the recognition accuracy and it slightly outperforms the ETSI AFE for DSR. The proposed filter is computationally efficient because it does not require any time-frequency conversion of signal, which saves a large amount of computational cost. As a result of recognition experiments, it is found that the optimum filter order is 5, which is smaller than that of Mel-LPC analysis, which further reduces the computational cost. The overall accuracy obtained by the proposed system is about 86.23%.

Table 3: *Average recognition accuracy for sets B and C.*

| Test Set | cln | 20dB | 15dB | 10dB | 5dB | 0dB | -5dB |
|---|---|---|---|---|---|---|---|
| B(w/ WF) | 98.50 | 97.52 | 96.32 | 93.04 | 83.42 | 59.61 | 25.71 |
| B(w/o WF) | 98.89 | 91.18 | 77.08 | 54.39 | 29.27 | 13.72 | 7.76 |
| C(w/ WF) | 98.13 | 96.99 | 95.60 | 91.41 | 80.47 | 56.08 | 25.34 |
| C(w/o WF) | 98.94 | 91.72 | 84.10 | 69.37 | 45.63 | 21.36 | 10.12 |

Table 4: *Comparative result for proposed system and ETSI AFE.*

| | Set A | Set B | Set C | Overall |
|---|---|---|---|---|
| **Proposed** | **87.53** | **85.99** | **84.11** | **86.23** |
| ETSI AFE | 87.18 | 86.29 | 83.25 | 86.04 |

## 6. Acknowledgment

This research has been supported by The Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid No.15500106.

## 7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust.,Speech Signal Processing* vol. 27, no. 2, pp. 113-120, 1979.

[2] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (nss), hidden Markov models and the projection or robust speech recognition in cars," *Speech Commun.*, vol. 11, no. 2- 3, pp. 215 -228, 1992.

[3] A. Agarwal and Y.M. Cheng, "Two-stage Mel-warped Wiener filter for robust speech recognition," *Proc. of ASRU'99*, 1999.

[4] D. Macho, et al., "Evaluation of a noise-robust DSR front- end on AURORA databases," *Proc. of ICSLP 2002*, pp. 17-20, 2002.

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No. 4, pp. 357-366, 1980.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 17-29, 1987.

[7] H.W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Am.*, vol. 68, no. 4, pp. 1071-1076, 1980.

[8] H. Matsumoto, Y. Nakatoh and Y. Furuhata, "An efficient Mel-LPC analysis method for speech recognition," *Proc. of ICSLP'98*, pp. 1051-1054, 1998.

[9] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *IEEE Proceedings*, vol. 60, no. 6, pp. 681-691, 1972.

[10] M. B. Islam, H. Matsumoto and K. Yamamoto, "Evaluation of Mel-Wiener filter for Mel-LPC based speech recognition," *Proc. SPECOM'05*, pp. 531-534, 2005.

[11] ETSI standard doc. "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm", ETSI ES 202 050 V1.1.1 (2002-10).

[12] Jin-Yu Li, et al., "A complexity reduction of ETSI advanced front-end for DSR", *Proc. ICASSP 2004*, vol. I, pp. 61-64, 2004.

[13] L. Mauuary, "Blind equalization in the cepstral domain for robust telephone speech recognition", *Proc. EUSPICO'98*, vol. 1, pp. 359-363, 1998.

[14] B. Juang, "On the hidden Markov model and dynamic time warping for speech recognition - a unified view," *AT&T Bell Laboratories Technical Journal*, vol.63, no.7, pp. 1213–1243, 1984.