SPEECH ENHANCEMENT BASED ON ITERATIVE WIENER FILTER USING COMPLEX SPEECH ANALYSIS

Keiichi Funaki

Computing & Networking Center, Univ. of the Ryukyus
Senbaru 1, Nishihara, Okinawa, 903-0213, Japan
phone: +(81)98-895-8946, fax: +(81)98-895-8963, email: funaki@cc.u-ryukyu.ac.jp

ABSTRACT

Recently, applications of speech coding and speech recognition have been exploding; for example, cellular phones and car navigation systems in an automobile. Since these are commonly used in noisy environment, noise reduction method, viz., speech enhancement is required as a pre-processor for speech coding and recognition. Iterative Wiener filter (IWF) method has been adopted as the speech enhancement that estimates speech and noise power spectra using LPC analysis iteratively. In this paper, we propose an improved method for Wiener filter algorithm by introducing the complex LPC speech analysis instead of the conventional LPC analysis. The complex speech analysis can estimate more accurate spectrum in low frequencies, thus it is expected that it can perform better for the IWF especially for babble noise or car internal noise that contains much energy in low frequencies. The objective evaluation has been performed for speech signal corrupted by white Gaussian, pink noise, babble noise or car internal noise by means of spectral distance. The results demonstrate that the proposed method can perform better for babble or car internal noise than the conventional real-valued method.

1. INTRODUCTION

In these days, the speech enhancement plays an important roll to improve the performance of the speech coding and speech recognition as cellular phone or the car navigation system are being widely used more and more. These systems are often used in noisy environment. Therefore, the quality of speech coding or the performance of speech recognition is deteriorated due to the surrounding noise. In order to avoid the deterioration, technology that removes noise from the noisy speech viz., speech enhancement is strongly desired. Especially, speech enhancement is an important factor for speech coding to keep the quality even under noisy environment. 3GPP (The 3rd Generation Partnership Project) thus provides the minimum performance requirement and the evaluation procedure for AMR-NB[1]. Several speech enhancement methods such as [2] have already satisfied the requirement. Moreover, speech enhancement is being sincerely demanded for wide band speech coding such as [3] since the additive noise in wide band speech can be percepted. Traditional approaches for speech enhancement have been proposed from the end of 1970's to 1980's [4],[5],[6]. Spectrum subtraction (SS) method [4] is widely adopted since it can be implemented easily and it can realize some degree of effect. However, the SS generates unpleasant artificial sound called musical noise so that it is not suitable for speech coding. Wiener filter method has been proposed by J.S.Lim[6], and the method designs the optimal filter minimizing the mean squared error (MSE) in the frequency domain. The musical noise is reduced by the Wiener filter method than the SS method. If accurate power spectrum for clean speech and accurate power spectrum of additive noise can be estimated, the Wiener filter can be designed accurately. However, the power spectrum of clean speech cannot be observed directly. Therefore, the iterative Wiener filter (IWF) method is adopted to estimate the power spectrum more accurately. First, the power spectrum for noise is estimated in silent segment of speech and the speech power spectrum is estimated by LPC analysis for noisy speech. Next, the Wiener filter is designed by using the estimated spectra and speech enhancement is carried out by filtering the noisy speech with the Wiener filter to obtain the enhanced speech. Next, LPC analysis is operated for the enhanced speech and the Wiener filter is designed again and the filter is operated for the noisy speech to obtain enhanced speech. These procedures are repeated to obtain more accurate speech power spectrum and to design more optimal Wiener filter. However, it is known that the spectrum of the enhanced speech is distorted after several iterations and the optimal number of iteration cannot be determined[7],[8].

On the other hand, the complex LPC speech analysis methods have already been proposed for an analytic signal [9],[10],[11]. An analytic signal is a complex signal having an observed signal in real part and a Hilbert transformed signal for the observed signal in imaginary part. Since the analytic signal provides the spectrum only on positive frequencies, the signals can be decimated by a factor of 2 with no degradation. As a result, the complex speech analysis offers attractive features, for example, more accurate spectral estimation in low frequencies. The remarkable feature is feasible to design more appropriate Wiener filter in the IWF and it may lead to higher performance of speech enhancement especially for the additive noise whose energy is concentrated in low frequencies, for example, babble noise or car internal noise.

In this paper, we propose an improved IWF method by adopting the MMSE based time-varying complex AR (TV-CAR) speech analysis[11] instead of LPC analysis. The TV-CAR speech analysis introduces the TV-CAR speech model, in which the AR model parameters are represented by complex basis expansion.

The reminder of this paper is organized as follows. We will explain the TV-CAR speech analysis in Section 2 and will explain the iterative Wiener filter method and the proposed algorithm in Section 3. The benefit of complex speech analysis will be explained in Section 4. We will explain the experiments evaluating the performance for additive white Gaussian, pink, babble, or car internal noise in Section 5.

2. TV-CAR SPEECH ANALYSIS

2.1 Analytic speech signal

Target signal of the time-varying complex AR (TV-CAR) method is an analytic signal that is complex-valued signal defined by

$$y^c(t) = \frac{y(2t) + j \cdot y_H(2t)}{\sqrt{2}} \tag{1}$$

where $y^c(t)$, $y(t)$, and $y_H(t)$ denote an analytic signal at time t, an observed signal at time t, and a Hilbert transformed signal for the observed signal, respectively. Since analytic signals provide the spectra only over the range of $(0, \pi)$, analytic signals can be decimated by a factor of two. The term of $1/\sqrt{2}$ is multiplied in order to adjust the power of an analytic signal with that of the observed one. Note that superscript c denotes complex value in this paper.

2.2 Time-Varying Complex AR (TV-CAR) model

Conventional LPC model is defined as

$$Y_{LPC}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I} a_i z^{-i}} \tag{2}$$

where a_i and I are i-th order LPC coefficient and LPC order, respectively. Since the conventional LPC model cannot express the time-varying spectrum, LPC analysis cannot extract the time-varying spectral features from speech signal. In order to represent the time-varying features, the TV-CAR model employs a complex basis expansion shown as

$$a_i^c(t) = \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) \tag{3}$$

where $a_i^c(t)$,I,L,$g_{i,l}^c$ and $f_l^c(t)$ are taken to be i-th complex AR coefficient at time t, AR order, finite order of complex basis expansion, complex parameter, and a complex-valued basis function, respectively. By substituting Eq.(3) into Eq.(2), one can obtain the following transfer function.

$$\begin{aligned} Y_{TVCAR}(z^{-1}) &= \frac{1}{1 + \sum_{i=1}^{I} a_i^c(t) z^{-i}} \\ &= \frac{1}{1 + \sum_{i=1}^{I} \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) z^{-i}} \end{aligned} \tag{4}$$

The input-output relation is defined as

$$\begin{aligned} y^c(t) &= -\sum_{i=1}^{I} a_i^c(t) y^c(t-i) + u^c(t) \\ &= -\sum_{i=1}^{I} \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) y^c(t-i) + u^c(t) \end{aligned} \tag{5}$$

where $u^c(t)$ and $y^c(t)$ are taken to be complex-valued input and analytic speech signal, respectively. In the TV-CAR model, the complex AR coefficient is modeled by a finite number of arbitrary complex basis. Note that Eq.(3) parameterizes the AR coefficient trajectories that continuously change as a function of time so that the time-varying analysis is feasible to estimate continuous time-varying speech spectrum. In addition, as mentioned above, the complex-valued analysis facilitates accurate spectral estimation in the low frequencies, as a result, the TV-CAR analysis allows for more accurate spectral estimation in low frequencies and since more optimal Wiener filter can be designed, it assigns better performance on speech enhancement.

Eq.(5) can be represented by vector-matrix notation as

$$\begin{aligned} \bar{y}_f &= -\bar{\Phi}_f \bar{\theta} + \bar{u}_f \\ \bar{\theta}^T &= [\bar{g}_0^T, \bar{g}_1^T, \cdots, \bar{g}_l^T, \cdots, \bar{g}_{L-1}^T] \\ \bar{g}_l^T &= [g_{1,l}^c, g_{2,l}^c, \cdots, g_{i,l}^c, \cdots, g_{I,l}^c] \\ \bar{y}_f^T &= [y^c(I), y^c(I+1), y^c(I+2), \cdots, y^c(N-1)] \\ \bar{u}_f^T &= [u^c(I), u^c(I+1), u^c(I+2), \cdots, u^c(N-1)] \\ \bar{\Phi}_f &= [\bar{D}_0^f, \bar{D}_1^f, \cdots, \bar{D}_l^f, \cdots, \bar{D}_{L-1}^f] \\ \bar{D}_l^f &= [\bar{d}_{1,l}^f, \cdots, \bar{d}_{i,l}^f, \cdots, \bar{d}_{I,l}^f] \\ \bar{d}_{i,l}^f &= [y^c(I-i) f_l^c(I), y^c(I+1-i) f_l^c(I+1), \\ &\qquad \cdots, y^c(N-1-i) f_l^c(N-1)]^T \end{aligned} \tag{6}$$

where N is analysis interval, $\bar{y}_f$ is $(N-I, 1)$ column vector whose elements are analytic speech signal, $\bar{\theta}$ is $(L \cdot I, 1)$ column vector whose elements are complex parameters, $\bar{\Phi}_f$ is $(N-I, L \cdot I)$ matrix whose elements are weighted analytic speech signal by the complex basis. Superscript T denotes transposition.

2.3 MMSE-based algorithm[11]

MSE criterion is defined as

$$\begin{aligned} \bar{r}_f &= [r^c(I), r^c(I+1), \cdots, r^c(N-1)]^T \\ &= \bar{y}_f + \bar{\Phi}_f \hat{\bar{\theta}} \end{aligned} \tag{7}$$

$$r^c(t) = y^c(t) + \sum_{i=1}^{I} \sum_{l=0}^{L-1} \hat{g}_{i,l}^c f_l^c(t) y^c(t-i) \tag{8}$$

$$E = \bar{r}_f^H \bar{r}_f = (\bar{y}_f + \bar{\Phi} \hat{\bar{\theta}})^H (\bar{y}_f + \bar{\Phi} \hat{\bar{\theta}}) \tag{9}$$

where $\hat{g}_{i,l}^c$ is the estimated complex parameter, $r^c(t)$ is an equation error, or complex AR residual and E is Mean Squared Error (MSE) for the equation error. To obtain optimal complex AR coefficients, we minimize the MSE criterion. Minimizing the MSE criterion of Eq.(9) with respect to the complex parameter leads to the following MMSE algorithm.

$$(\bar{\Phi}_f^H \bar{\Phi}_f) \hat{\bar{\theta}} = -\bar{\Phi}_f^H \bar{y}_f \tag{10}$$

Superscript H denotes Hermitian transposition. After solving the linear equation of Eq.(10), we can get the complex AR parameter at time t ($a^c(t)$) by calculating the Eq.(3) with the estimated complex parameter $\hat{g}_{i,l}^c$.

3. WIENER FILTER ALGORITHM

3.1 Wiener filter

Assuming that the clean speech $s(t)$ degraded by an additive noise $w(t)$, the noisy speech $x(t)$ is defined by

$$x(t) = s(t) + w(t). \tag{11}$$

Wiener filter is an optimal filter that minimizes the Mean Squared Error (MSE) criterion. In the case of Eq.(11), the filter can be defined by

$$S(\omega) = H(\omega)X(\omega) \tag{12}$$

where ω is the frequency index, $S(\omega), X(\omega)$, and $H(\omega)$ are the discrete Fourier transform of the clean speech, the transform of noisy speech and transfer function of Wiener filter, respectively. The MSE can be defined as follows. The Wiener filter can be derived by

$$H(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{ww}(\omega)} \tag{13}$$

where $P_{ss}(\omega)$ and $P_{ww}(\omega)$ are power spectrum of speech, $s(t)$ and that for noise, $w(t)$, respectively.

From Eq.(12) and (13), the enhanced speech is estimated in the frequency domain by

$$S(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{ww}(\omega)} X(\omega) \tag{14}$$

The enhanced speech can be obtained by inverse FFT for $S(\omega)$ and OLA (OverLap Add) procedure is carried out in the time domain between adjacent frames to avoid click sound.

3.2 Iterative Wiener Filter (IWF) algorithm[6][8]

The performance of the Wiener filter depends on the accuracy of speech power spectral estimation, $P_{ss}(\omega)$. It is possible to make the estimated spectrum close to the true one by repeating the Wiener filter processing. Figure 1 shows the block diagram of the iterative Wiener filter algorithm. The two kinds of power spectra can be estimated by LPC analysis as follows. Noise power spectrum, $P_{ww}(\omega)$ are estimated in the first non-speech section. Speech power spectrum, $P_{ss}(\omega)$ is estimated by LPC analysis for input noisy speech, $x(n)$. By the Wiener filtering and Inverse FFT operation, enhanced speech is estimated and then it is analyzed in order to estimate more accurate speech power spectrum, $P_{ss}(\omega)$ by means of LPC analysis and the Wiener filter is operated again. The iterative procedure is repeated to obtain more clean speech.

3.3 Proposed Method

The proposed method employs the estimated two spectra, $P_{ss}(\omega)$ and $P_{ww}(\omega)$ estimated by the TV-CAR speech analysis[11] instead of LPC analysis. The two spectra can be estimated as

$$P_{ss}(\omega) = \frac{G_s^2}{|1 + \sum_{i=1}^{I}\sum_{l=0}^{L-1} \hat{g}^c_{s_{i,l}} f_l^c(t) e^{-ji\omega}|^2} \tag{15}$$

$$P_{ww}(\omega) = \frac{G_w^2}{|1 + \sum_{i=1}^{I}\sum_{l=0}^{L-1} \hat{g}^c_{w_{i,l}} f_l^c(t) e^{-ji\omega}|^2}. \tag{16}$$

$\hat{g}^c_{s_{i,l}}$ is estimated by Eq.(10) from analytic speech for the enhanced speech at previous iteration (input speech at first iteration) and G_s is energy of the corresponding residual. $\hat{g}^c_{w_{i,l}}$ is estimated by Eq.(10) from analytic speech for the input speech and G_w is energy of the corresponding residual. Note that these two power spectra provide only one side of spectrum, thus, mirroring is operated to apply to Eq.(13).

As mentioned above, complex speech analysis can estimate more accurate speech spectrum in low frequencies. It is expected that the feature leads to higher performance of the IWF algorithm. In this paper, time-invariant complex speech analysis ($L = 1$), is equivalent to complex LPC analysis, is adopted.

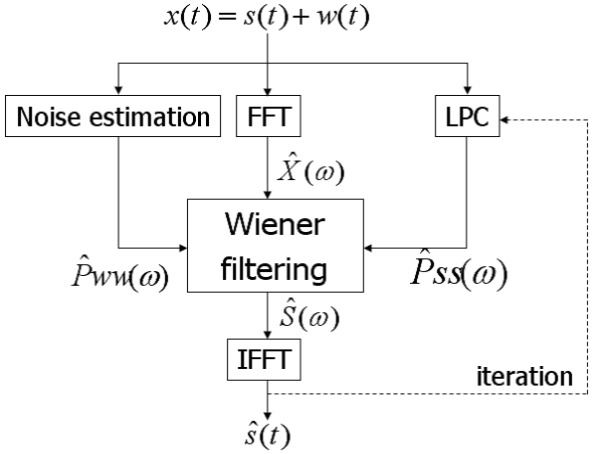


Figure 1: Block diagram of the IWF algorithm

4. BENEFIT OF COMPLEX SPEECH ANALYSIS

Figure 2 shows example of the estimated speech spectra by complex LPC speech analysis for analytic signal[10] and conventional LPC analysis for speech signal.

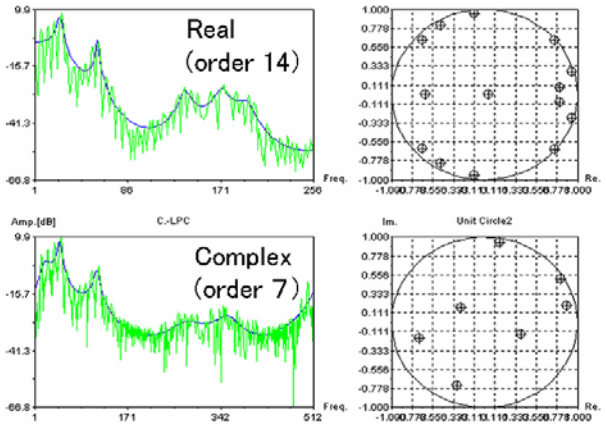


Figure 2: Estimated Spectra with complex and conventional LPC analysis

In Figure 2, left side denote the estimated spectra. Upper is for real-valued LPC analysis. Lower is for complex-valued LPC analysis. Blue line means estimated spectrum by LPC analysis and green line means estimated DFT spectrum. Right side means estimated poles from the estimated AR filter.

One can observe that the complex analysis can estimate more accurate spectrum in low frequencies whereas the estimation accuracy is down in high frequencies. Since speech

spectrum provides much energy in low frequencies, it is expected that the high spectral estimation accuracy in low frequencies makes it possible to improve the performance on the IWF.

5. EXPERIMENTS

We have already carried out the experiments to compare the performance of the proposed method (TV-CAR) for analytic speech with that for the conventional one (LPC method) for observed speech by means of objective evaluation of LPC Cepstral distance (CD). Since the IWF is based on filtering in the frequency domain, spectral distance such as LPC cepstral distance is appropriate measure for objective evaluation. Table 1 shows the experimental conditions. Sampling rates were 16KHz or 8KHz. Additive noises were white Gauss noise, pink noise, babble noise or car internal noise[12]. Noise levels were -5, 0, 5, 10 or 20[dB]. In the TV-CAR speech analysis, L is set to be one, thus the TV-CAR speech analysis is equivalent to non-time varying, complex LPC analysis. Figures 3 and 4 show the experimental results. Figure 3 means the results for 8KHz of speech. Figure 4 means the results for 16KHz of speech. In these figures, (1),(2),(3) and (4) means CDs for additive white Gauss noise, those for additive pink noise, those for for additive babble noise, and those for additive car internal noise, respectively.

In these figures, X-axis means noise level (20, 10, 5, 0, -5 [dB]) and Y-axis means CD. **LPC** denotes the CDs by means of the conventional method with LPC analysis. **CLPC** denotes the CDs by means of the proposed method with complex LPC analysis. The results demonstrate that the proposed method can perform better than the conventional one for additive pink, babble or car internal noise whereas the proposed method does not perform better for additive white Gauss noise. The reason why the proposed method can perform better for additive pink, babble or car internal noise is as follows. The complex speech analysis can estimate more accurate speech spectrum in low frequencies for these noises whose energy is concentrated in low frequencies.

Table 1 Experimental Conditions

Speech data	Male 10 sentences Female 10 sentences ATR database set B
Sampling	8KHz/16bit 16KHz/16bit
Window Length	20m
Shift Length	10ms
FFT	1024 samples
LPC analysis	I=14, L=1(time-invariant)
Pre-emphasis	None
Complex LPC analysis	I=7, L=1(time-invariant)
Pre-emphasis	None
Noise	(1) White Gauss noise (2) Pink noise[12] (3) Babble noise[12] (4) Car internal noise[12]
Noise Level	20,10,5,0,-5[dB]
Cepstral Distance (CD)	LPC Cepstral Distance
Window Length	20[msec]
Shift Length	20[msec]
LPC order	16/32 for 8/16KHz
Cepstral order	16/32 for 8/16KHz

6. CONCLUSIONS

In this paper, we have proposed the improved iterative Wiener filter (IWF) algorithm based on the TV-CAR speech analysis in a single channel system. The performance has already been evaluated by means of LPC cepstral distance (CD) not only for 8KHz but also for 16KHz sampled speech signal corrupted by additive white Gauss, pink, babble or car internal noise. According to an informal listening test and objective evaluation of CD, the proposed method outperforms conventional IWF for additive pink, babble or car internal noise that contains much energy in low frequencies. Future study is as follows.

1. Improve the noise estimation
2. Introduce robust TV-CAR speech analysis based on ELS method[13]
3. Introduce the time-varying speech analysis ($L = 2$).

REFERENCES

[1] "Minimum Performance Requirements for Noise Suppressor Application to the AMR Speech Encoder," 3GPP TS 06.77 V8.1.1,Apr.2001.

[2] M.Kato,et.al., "Noise Suppression with High Speech Quality Based on Weighted Noise Estimation and MMSE STSA," IEICE Trans. Vol.E85-A. No.7, July 2002.

[3] ITU-T Recommendation G.722.2, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)," Jul.,2003.

[4] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans., ASSP-27, pp.113-120,1979.

[5] Y.Ephraim and D.Malah, "Speech enhancement using minimum mean-square error log-spectral amplitude estimator," IEEE Trans., ASSP-33, pp.443-445, 1985.

[6] J.S.Lim and A.V.Oppenheim, "All-pole modeling of degraded speech," IEEE Trans., ASSP-26, pp.197-210, 1978.

[7] H.L.Hansen and M.A.Clements, "Constrained iterative speech enhancement with application to speech recognition," IEEE Trans. Signal Processing, vol.39, pp.795-805, April 1991.

[8] P.C.Loizou,"Speech Enhancement, Theory and Practice," CRC Press, 2007.

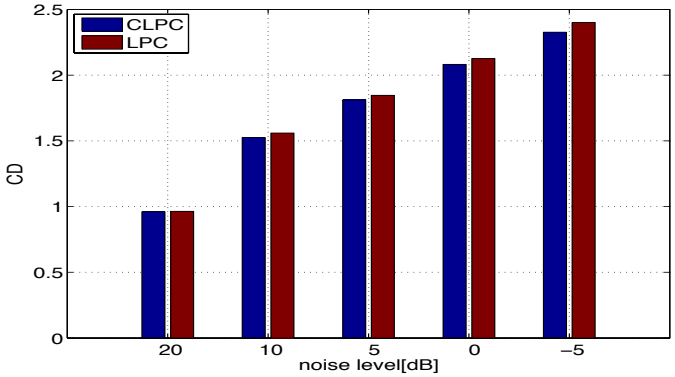
[9] S.M.Kay, "Maximum entropy spectral estimation using the analytic signal," IEEE Trans. ASSP-26, pp.467-469, 1980.

[10] T.Shimamura and S.Takahashi, "Complex linear prediction method based on positive frequency domain," IEICE Trans., Vol.J72-A, pp.1755-1763, 1989. (in Japanese)

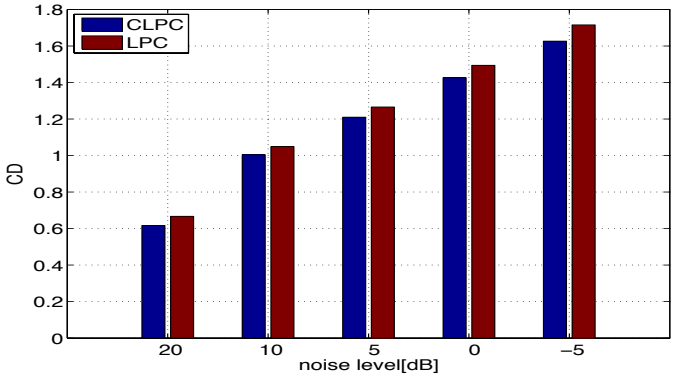
[11] K.Funaki, et.al., "On a time-varying complex speech analysis," Proc. EUSIPCO-98, Rhodes, Greece, Sep. 1998.

[12] NOISE-X92, http://spib.rice.edu/spib/select_noise.html

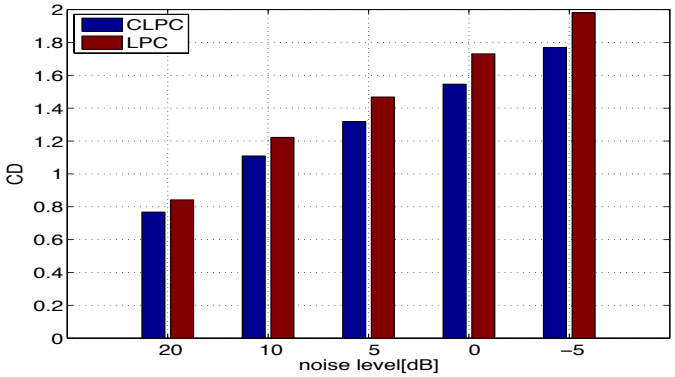
[13] K.Funaki, "A time-varying complex AR speech analysis based on GLS and ELS method," Proc. EUROSPEECH-2001, Alborg, Denmark, Sep. 2001.



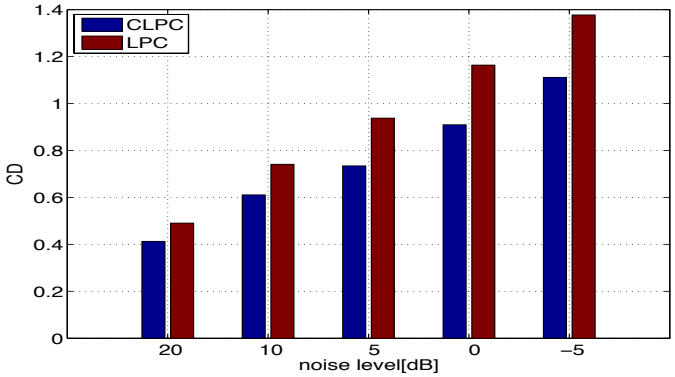
(1) CDs for additive white Gauss noise (8KHz)



(2) CDs for additive Pink noise (8KHz)

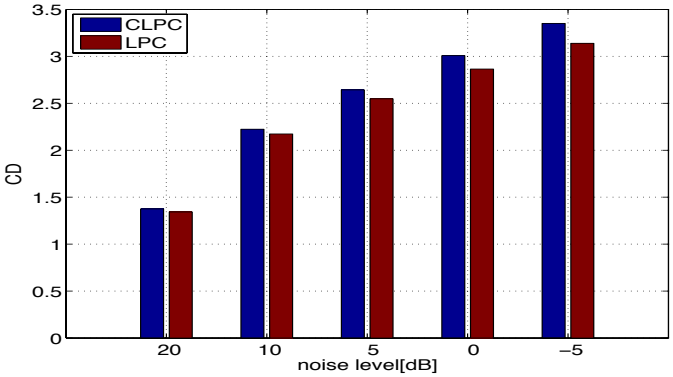


(3) CDs for additive Babble noise (8KHz)

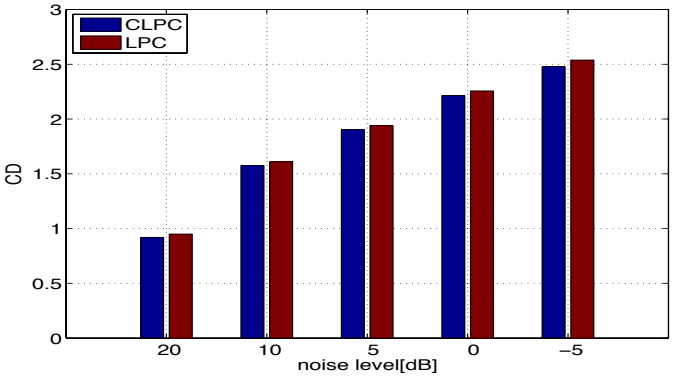


(4) CDs for additive Car Internal noise (8KHz)

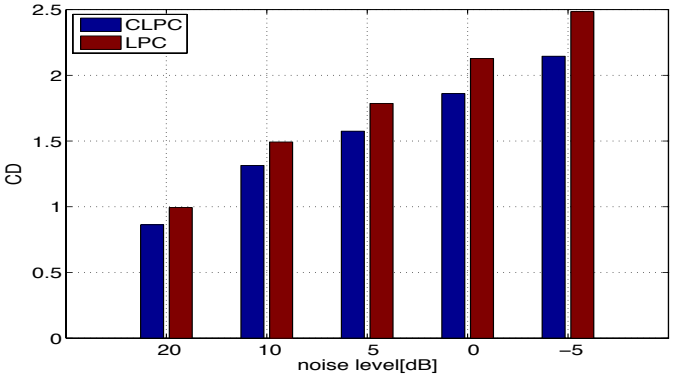
Figure 3: CDs for 8KHz speech



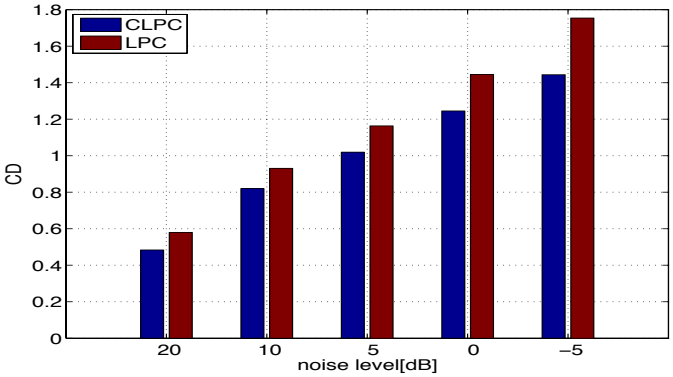
(1) CDs for additive white Gauss noise (16KHz)



(2) CDs for additive Pink noise (16KHz)



(3) CDs for additive Babble noise (16KHz)



(4) CDs for additive Car Internal noise (16KHz)

Figure 4: CDs for 16KHz speech