Spectrum Compensation Method for Speech Signals Based on Prediction Error Filtering

MD ARIFOUR RAHMAN
Graduate School of
Science and Engineering
Saitama University
255 Shimo-Okubo
Saitama, 338-8570
JAPAN
arifour@sie.ics.saitama-u.ac.jp

YOSUKE SUGIURA
Graduate School of
Science and Engineering
Saitama University
255 Shimo-Okubo
Saitama, 338-8570
JAPAN
ysugiura@mail.saitama-u.ac.jp

TETSUYA SHIMAMURA
Information Technology Center
Saitama University
255 Shimo-Okubo
Saitama, 338-8570
JAPAN
shima@sie.ics.saitama-u.ac.jp

Abstract: This paper proposes a technique for improving the performance of linear prediction (LP) by utilizing the prediction error filter (PEF) as a pre-processor. Problems often occur in estimating the power spectrum of the input speech signal using LP due to the large spectral dynamic range of speech which makes the autocorrelation matrix ill-conditioned. In the proposed method, the LP based power spectrum estimation is compensated by the spectrum characteristics of the designed PEF. The accuracy of formant frequency estimation is verified on synthetic speech. The validity of the proposed method is also illustrated by inspecting real air conducted and bone conducted speeches. Through the experiments, we show that the proposed method can estimate the power spectrum more accurately than the conventional direct and pre-emphasis LP methods.

Key–Words: Linear prediction, prediction error filter, formant frequency estimation, spectrum compensation, air conducted speech, bone conducted speech.

1 Introduction

Linear prediction (LP) is one of the most powerful methods that have been extensively used in variety of signal processing applications [1], [2]. Especially in speech processing, LP has received a considerable attention because it has a close connection with the production model of speech [3]. Two mainly used methods for LP are the autocorrelation method [4] and covariance method [5]. They are sometimes referred to as the stationary method and non-stationary method, respectively [6]. In this paper, the autocorrelation method in which less computation is required is considered.

It is known that the Levinson-Durbin algorithm utilized in the autocorrelation method guarantees the resulting autoregressive model to be stable [1], [3]. The performance is, however, degraded in an ill-conditioned environment [1] where the input signal has a wide spread of power spectrum dynamic range. A number of techniques have been mentioned to mitigate the problem of ill-conditioning [3] [7]. A pre-emphasis filter is very often used [3] [7]. In this case, however, the resulting power spectrum estimate of the input signal is inherently biased. Although ill-conditioning is mitigated by adding a small positive value to the diagonal elements of the correlation matrix, the resulting power spectrum estimate is again biased [7]. In this paper, we present a spectrum compensation (SC) method for LP to deal with the ill-conditioning problem. To obtain an accurate representation of speech power spectrum, the prediction error filter (PEF) is used as a pre-processor. The followed LP provides an estimate of power spectrum. Unlike the conventional approaches, however, the resulting power spectrum is compensated by the spectrum characteristics the PEF possesses. As a result, an unbiased and accurate power spectrum estimate is obtained. The performance of the proposed SC method is investigated through experiments on synthetic and real speeches.

This paper is organized as follows. We describe the conventional LP methods in Section 2 briefly and derive the SC method in Section 3. Section 4 is devoted to experimental results on synthetic speech. In Section 5, the SC method is applied to real speeches recorded by two different types of microphones, which are for air conduction and for bone conduction. Finally, in Section 6 a conclusion is drawn.

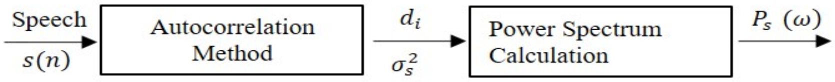

Figure 1: Power spectrum estimation using direct method

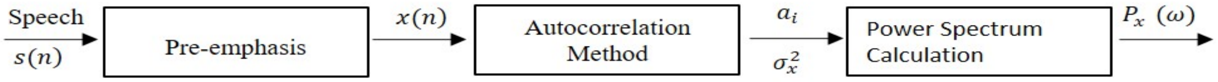

Figure 2: Power spectrum estimation using pre-emphasis method

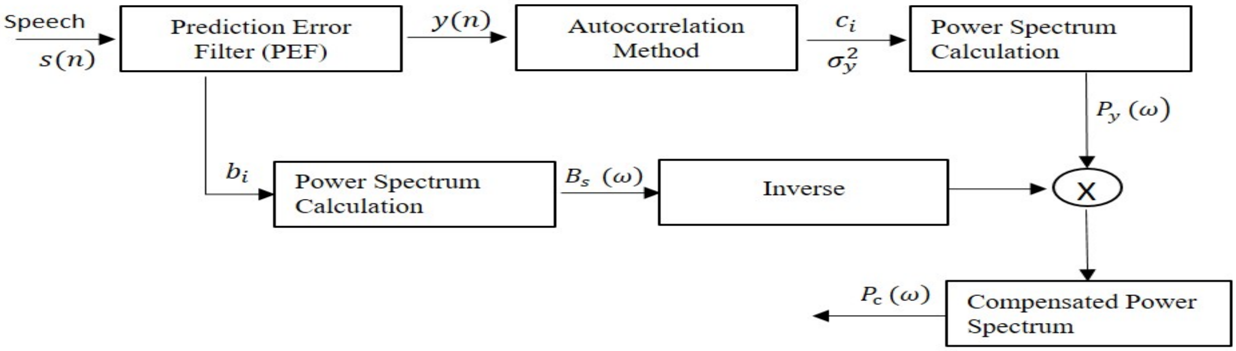

Figure 3: Power spectrum estimation using SC method

2 Conventional LP Methods

Let us assume that the input speech signal is represented by $s(n)$ where n denotes a discrete time. The sampling period and sampling frequency are T and f_s, respectively, that is $T = 1/f_s$. When LP is directly used to $s(n)$, we usually use a procedure shown in Figure 1. In the autocorrelation method of LP, the Levinson-Durbin algorithm is applied to determine the prediction error power, σ_s^2, and the prediction coefficients, d_i, $(i = 1, 2...M_s)$, where M_s is the prediction order. The power spectrum is estimated as

$$P_s(\omega) = \frac{\sigma_s^2}{|1 + \sum_{i=1}^{M_s} d_i e^{-ji\omega T}|^2} \tag{1}$$

where ω is the angular frequency and j is defined by $j = \sqrt{-1}$. However, since $s(n)$ has a certain spread of power spectrum, in many cases a pre-emphasis filter whose transfer function is represented by

$$H_{PE}(z) = 1 - \eta z^{-1} \tag{2}$$

is used as shown in Figure 2. The parameter η in (2) is called pre-emphasis coefficient. The output of the pre-emphasis filter is given by

$$x(n) = s(n) - \eta s(n-1). \tag{3}$$

The coefficient η is often set to between 0.9 and 1, reflecting the degree of pre-emphasis. Basically, high frequency components of the input signal $s(n)$ are emphasized through the pre-emphasis filter. Since the pre-emphasis filter produces another signal $x(n)$ from the input signal $s(n)$, the resulting power spectrum from the autocorrelation method of LP is described as

$$P_x(\omega) = \frac{\sigma_x^2}{|1 + \sum_{i=1}^{M_x} a_i e^{-ji\omega T}|^2} \tag{4}$$

where a_i are the prediction coefficients, σ_x^2 is the prediction error power and M_x corresponds to the predictor order in this case. To the input signal $s(n)$, the resulting power spectrum $P_x(\omega)$ will provide more accurate peaks than the direct LP method shown in Figure 1 does. However, $P_x(\omega)$ does not provide the original power spectrum of the input signal $s(n)$. Although the pre-emphasis method shown in Figure 2 is very often used for the purpose of formant frequency estimation and pitch detection, its application is restricted in practice.

3 Proposed Method

In this section, a method to estimate accurately the original power spectrum of the input speech signal, SC method, is derived. In the SC method, the PEF works as a pre-processor as shown in Figure 3. The PEF filter is realized as a finite impulse response (FIR) filter. This filter type is the same as that of the pre-emphasis filter in Figure 2. The filter realization of the PEF is, however, more flexible. The transfer function of the PEF is described as

$$H_{PEF}(z) = 1 + \sum_{i=1}^{L} b_i z^{-i} \tag{5}$$

where b_i are the prediction coefficients and L is the prediction order. The pre-emphasis filter in Figure 2 corresponds to the case where $L = 1$ and $b_1 = -\eta$. For the pre-emphasis filter, the coefficients η is fixed and used for implementation. On the other hand, for the PEF, the filter order L is increased more and the filter coefficients b_i are determined depending on the input signal $s(n)$.

The output of the PEF, $y(n)$, will have a relatively flat power spectrum compared to that of the input signal $s(n)$. The output signal $y(n)$ is followed by the autocorrelation method of LP and the power spectrum, $P_y(\omega)$, is calculated as

$$P_y(\omega) = \frac{\sigma_y^2}{|1 + \sum\limits_{i=1}^{M_y} c_i e^{-ji\omega T}|^2} \tag{6}$$

where c_i are the prediction coefficients, σ_y^2 is the prediction error power and M_y corresponds to the predictor order. Here, from the PEF used as the pre-processor, we calculate

$$B_s(\omega) = |1 + \sum_{i=1}^{L} b_i e^{-ji\omega T}|^2. \tag{7}$$

Then we compensate for the power spectrum in (6) as

$$P_c(\omega) = \frac{P_y(\omega)}{B_s(\omega)}. \tag{8}$$

Since there exists the following relationship between the input and output through the PEF;

$$P_y(\omega) = |H_{PEF}(e^{j\omega T})|^2 P_s(\omega), \tag{9}$$

from (7) and (9), we can find that $P_c(\omega)$ in (8) provides an estimate of the original power spectrum $P_s(\omega)$. For the SC method, an unbiased estimate of the original power spectrum $P_s(\omega)$ is obtained through (8).

For the SC method, the order of the PEF, L, should be small as $L < M_y$. This is because the autocorrelation method suffers from ill-conditioning of the correlation matrix of the input signal. Let us assume that the correlation matrix of the input signal $s(n)$ is expressed by $\mathbf{R}_s$. The degree of ill-conditioning of $\mathbf{R}_s$ is measured by the magnitude of the condition number defined by

$$C_s = \frac{\lambda_{s,max}}{\lambda_{s,min}} \tag{10}$$

where $\lambda_{s,max}$ and $\lambda_{s,min}$ correspond to the maximum and minimum eigenvalues of $\mathbf{R}_s$. In implementing the autocorrelation method, the condition number C_s severely affects the performance of the autocorrelation method. In many cases of speech processing, C_s is very large. This is the reason why the use of the autocorrelation method shown in Figure 2 is often used. The pre-emphasis filter mitigates the spread of eigenvalues in the correlation matrix, leading to accurate power spectrum estimation. It is known that an increase of the prediction order accelerates the degree of ill-conditioning [8], [9]. Therefore, in the proposed method, the prediction order L of the PEF should be set to a comparatively small one. In this case, the prediction accuracy of $B_s(\omega)$ will be increased. Furthermore, the computational complexity of the autocorrelation method is dominated by that of the Levinson-Durbin algorithm, which is a square order of the prediction order. In the proposed method, when $L \ll M_y$, the computation to obtain $B_s(\omega)$is significantly less than that to do $P_y(\omega)$ and almost negligible.

4 Experiments

To validate the performance of the proposed SC method, we conducted experiments. In this section, synthetic vowels were employed as speech data. We utilized the Liljencrants-Fant (LF) model [10], [11] to generate the synthetic vowels. Table 1 shows the first

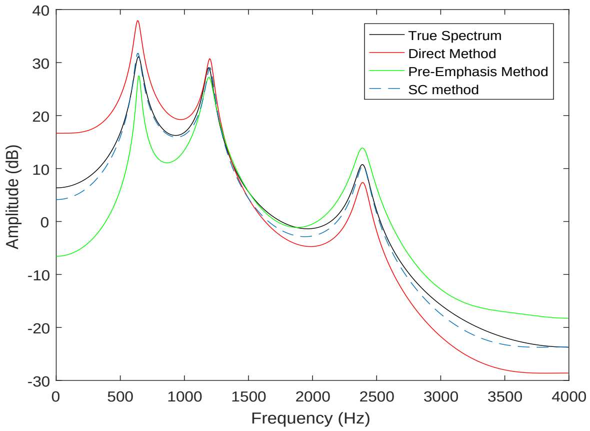

Figure 4: Power spectrum of synthetic vowel /a/

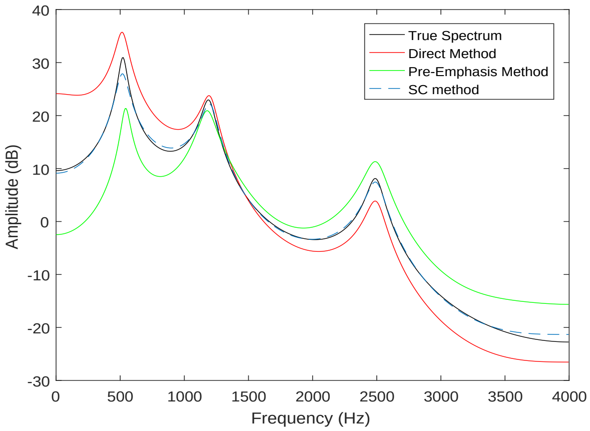

Figure 5: Power spectrum of synthetic vowel /u/

three formants (F1, F2 and F3) and their corresponding bandwidths (B1, B2 and B3) used to generate the synthetic vowels.The fundamental frequency for the source excitation was 120 Hz. The vowels were generated with a sampling rate of 8 kHz. Table 2 shows the experimental conditions. We compared the performance of the SC method with that of the direct method (Figure 1) and the pre-emphasis method (Figure 2) . Comparison was made by the visual inspection as well as by computing the spectral bias defined as

$$B = \frac{1}{\pi f_s} \int_0^{\pi f_s} \frac{|\hat{P}(\omega) - P(\omega)|}{P(\omega)} d\omega \qquad (11)$$

where $P(\omega)$ and $\hat{P}(\omega)$ denote the true power spectrum and estimated power spectrum, respectively. We calculated the estimated power spectrum from the prediction coefficients and prediction error power using fast Fourier transform (FFT). The true power spectrum was obtained by constructing the all pole filter to generate the synthetic vowel and by FFT of the resulting filter coefficients. As example, in Figures 4 and 5 we show the power spectrum of the synthetic vowel signals /a/ and /u/, respectively. It can be seen from Figures 4 and 5 that the power spectrum estimated by the SC method (dotted line) is commonly the closest to the true power spectrum. The average value of the spectral bias B for five vowels was measured and shown in Table 3. We calculated the average for five vowels taking twenty frames from each vowel data. It can be seen from Table 3 that for all the cases, the proposed SC method provides smaller values of B than the other two methods. A smaller value of B indicates that the estimated spectrum is closer to the true spectrum. Therefore, the SC method estimates the power spectrum more accurately than the other two methods.

To investigate further the performance of the SC method, formant frequency estimation was observed. Each of the five vowels was used for twenty frames with half overlapping. The location of each formant was found by peak-picking the power spectrum evaluated, and the formant frequency was detected. We used the cepstral algorithm to obtain the formant locations. Low-time liftering was used for estimating the vocal tract characteristics from the computed cepstrum. FFT was applied to find the log amplitude spectrum of low-time lifted cepstral coefficients. Finally, the formant frequency estimation was achieved by the peak-picking algorithm. Each formant frequency estimate of F1, F2 and F3 was averaged for five vowels. In order to optimize the order L of the PEF, we investigated the formant estimation accuracy for the SC method using different values of L. Table 4 shows the resulting formant estimation errors in percentage to the true formant frequency. From Table 4, it can be seen that the case of $L = 2$ provides the minimum error. This result indicates that the performance of the SC method is degraded with the order $L \geq 3$ of the PEF, while the order $L = 1$ of the PEF is not sufficient for the SC method. Since the condition number of the correlation matrix to implement the PEF is increased as the order L is increased, a small setting for L is desired. The result in Table 4 is matched to this property and suggests that the best order for the PEF in the SC method is $L = 2$.

Table 5 lists the averaged formant estimation errors in percentage for performance comparison where the SC method was implemented with $L = 2$ being the best setting. It can be seen from Table 5 that the estimation errors made by the SC method are smaller than the other two methods again. It is interesting to compare the performance of the pre-emphasis method in Table 5 with the case of $L = 1$ of the SC method in Table 4. From Tables 4 and 5, we see that the SC method provides an improvement relative to the pre-emphasis method even if $L = 1$ is set for the PEF in the SC method. This suggests that the coefficient in the pre-emphasis filter should be time-variant depending on the input speech signal and that the spectrum compensation is useful to lead to more accurate power spectrum estimation.

For the purpose of observing the conditioning of the correlation matrix used in the autocorrelation method for each comparative method, we evaluated the condition number as follows. For the direct method, the correlation matrix R_s of the input speech signal $s(n)$ was constructed and C_s in (10) was used to calculate the condition number. For the pre-emphasis and SC methods, from each output of the pre-filter, $x(n)$ and $y(n)$, the corresponding correlation matrices, R_x and R_y, were constructed and the condition numbers, C_x and C_y, were calculated, respectively. The five vowels were used in twenty frames with half overlapping to calculate the average value of the condition number. Table 6 lists the resulting condition number in power decibels (for example, C_s(in dB)$= 10\log_{10} C_s$ for the direct method). It can be seen from Table 6 that the condition number evaluated from the SC method is smaller than the other two methods. This indicates that for the SC method, the spectral dynamic range is reduced, resulting in that the problem of ill-conditioning for the correlation matrix is mitigated. The degree of ill-conditioning is directly associated with the performance of the autocorrelation method of LP [3]. Therefore, Table 6 could validate that the SC method provides excellent performances in power spectrum estimation.

Table 1: First three formants and their corresponding bandwidths used to generate synthetic vowels

Vowel	F1	F2	F3	B1	B2	B3
/a/	640	1190	2390	60	60	100
/i/	390	1990	2550	50	100	140
/u/	520	1190	2490	65	110	140
/e/	270	2311	3010	70	100	200
/o/	730	1090	2440	80	50	130

5 Application to Real Speeches

In this section, the application of the SC method to two types of real speeches; air-conducted (AC) speech and bone-conducted (BC) speech, is discussed. Continuous AC and BC speech signals of two male and two female speakers, which were simultaneously recorded in a sound-isolated room

Table 2: Experimental conditions

LP Order L: 2
LP Order (M_x, M_y and M_s): 12
FFT Points: 1024
Frame Length: 25ms
Frame Shift: 12.5ms
Window Type: Hamming
Signal Length: 2 sec.

Table 3: Spectral bias of five vowels for different methods

Vowel	Direct	Pre-emphasis	SC
/a/	0.54	0.52	0.16
/i/	0.45	0.46	0.20
/u/	0.56	0.54	0.18
/e/	0.35	0.24	0.14
/o/	0.36	0.31	0.11
Average	0.45	0.41	0.16

Table 4: Formant estimation errors of SC method for different order PEF

Formants	L=1	L=2	L=3	L=4
F1	2.46	2.41	2.47	2.48
F2	0.56	0.47	0.68	0.44
F3	0.41	0.39	0.38	0.62

Table 5: Formant estimation errors in percentage for different methods

Formants	Direct	Pre-emphasis	SC
F1	2.61	2.52	2.41
F2	0.92	0.66	0.47
F3	0.52	0.48	0.39

Table 6: Condition number in decibels for different methods

Vowel	Direct	Pre-emphasis	SC
/a/	83.20	65.46	40.35
/i/	73.69	59.81	47.52
/u/	95.73	68.96	65.77
/e/	73.01	60.65	40.40
/o/	107.06	85.28	71.86
Average	86.54	68.02	53.18

Table 7: Condition number in decibels for different methods on AC and BC speeches

Speech	Direct	Pre-emphasis	SC
AC	71.54	64.96	47.44
BC	108.32	82.68	62.01

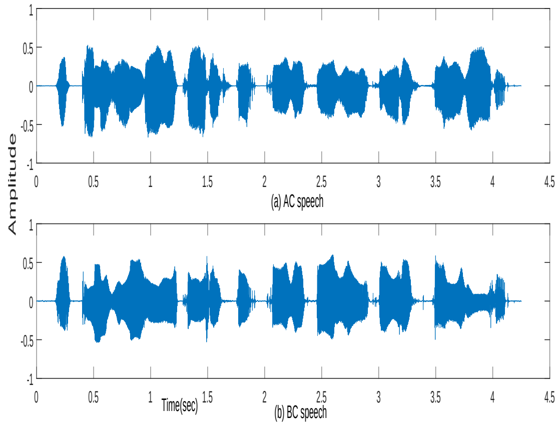

Figure 6: Waveforms of real speech signals

were utilized with a sampling frequency of 8 kHz. Sony ECM-J3M and Temco HG-17 were used as the microphone for recording the AC and BC speeches, respectively. One speaker uttered two different Japanese sentences, each length of which was about 4 seconds. Figure 6 shows an example of the AC and BC speech waveforms used in the experiments, where the sentence “Bukka no hendou wo kouryoshite kyufusuijun wo kimeru hitsuyouga aru” was spoken by a female speaker. In the following experiments, twenty frames of the voiced parts from one sentence were randomly selected. For eight sentences from two male and two female speakers, totally 160 frames were randomly selected and used.

Although there are some research works related with the BC speech [12] [13], the properties of the BC speech is still not well known. One known property of the BC speech signal is that in noiseless environment, the speech quality is worse than that of the corresponding AC speech. This may be caused by the fact that the high frequency components in the BC speech deteriorate, resulting a larger power spectral dynamic range compared to that of the AC speech. If this is true, this phenomenon of the BC speech may lead to more severe ill-conditioning of the correlation matrix for the autocorrelation method. Clearly, this is not suitable for LP analysis. However, as far as we know, this topic for the BC speech has never been discussed up to now. In this section, we set out to investigate the conditioning of a correlation matrix constructed by the BC speech signal.

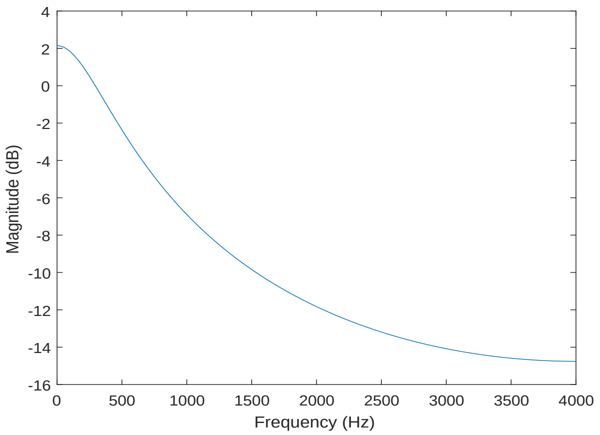

Figure 7: Amplitude response of the IIR filter

At first, to know the spectral dynamic range of the BC speech, we derive a filter model to transform the AC speech signal into the corresponding BC speech signal. An infinite impulse response (IIR) filter model is assumed as

$$v(n) = \mu v(n-1) + \varphi u(n) \tag{12}$$

where $u(n)$ and $v(n)$ are the AC speech signal and BC speech signal, respectively, and μ and φ are the filter coefficients. Adjusting the coefficients μ and φ in (12), for each frame among the total 160 frames, we found the coefficients pair for each frame by which the power spectrum of the filtered AC speech signal is best matched to that of the corresponding BC speech signal. And, observing all the results, we found a typical one for the coefficients pair of the IIR filter as $\mu = 0.85$ and $\varphi = 0.32$. For the above all cases, the coefficient μ was variated a range from 0.80 to 0.95, and the coefficient φ was done in a range from 0.25 to 0.35. Figure 7 shows the amplitude response of the resulting typical IIR filter. The resulting filter indicates clearly a low pass type, which provides about 17 dB attenuation in the high frequency region relative to the amplitude at the zero frequency. Figure 8 shows an example of the amplitude spectrum of the AC speech signal filtered by the IIR filter shown in Figure 7. In Figure 8, we can observe that the amplitude spectrum of the BC speech signal is almost overlapped with that of the filtered AC speech signal. This validates that the IIR filter shown in Figure 7 is a good transformation filter from the AC speech signal to the BC speech signal.

From Figure 7, it is obvious that the spectral dynamic range of the BC speech is more expanded than that of the corresponding AC speech. It is

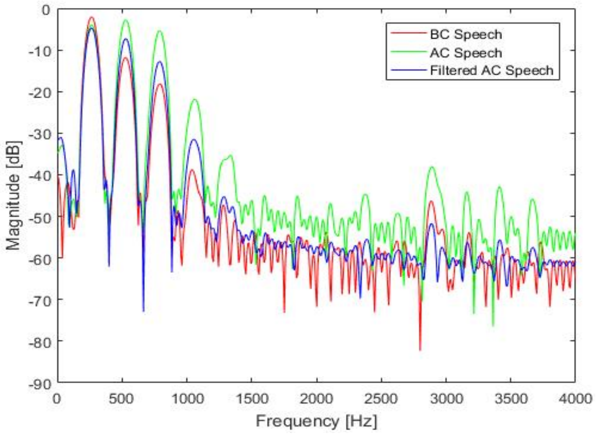

Figure 8: Amplitude spectrum of the filtered AC speech

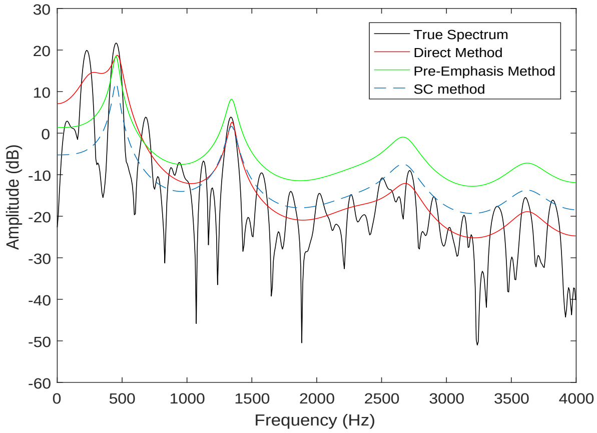

Figure 9: Power spectrum of AC speech from female speaker

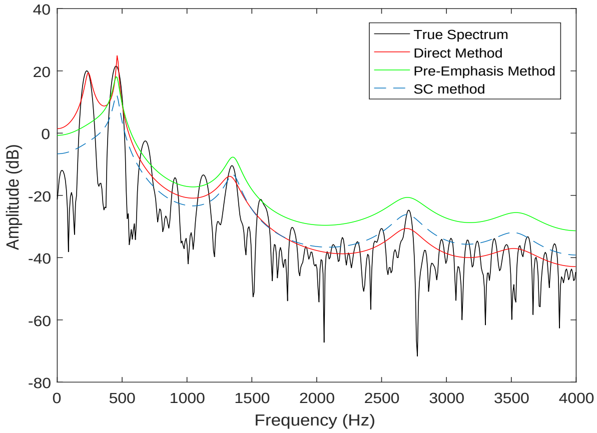

Figure 10: Power spectrum of BC speech from female speaker

known that the condition number of the correlation matrix of an input speech signal is proportional to the spectral dynamic range of the input speech signal

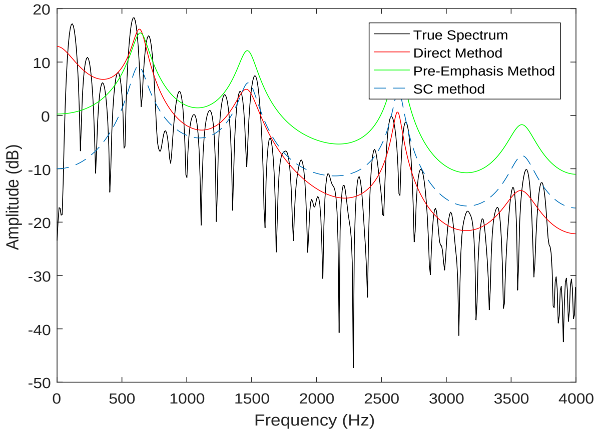

Figure 11: Power spectrum of AC speech from male speaker

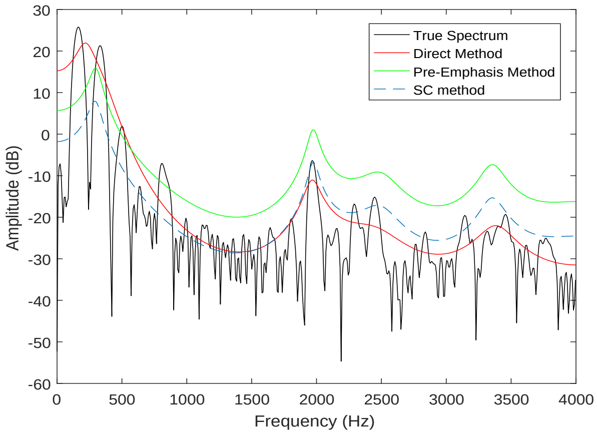

Figure 12: Power spectrum of BC speech from male speaker

[3]. To confirm if this is true, we applied the direct, pre-emphasis and SC methods to the prepared AC and BC speech data sets. Table 7 shows the condition number for different methods on the AC and BC speeches. Table 7 shows that the condition number is comparatively increased on the BC speech. In Table 7, the SC method provides lower values of the condition number compared to the other two methods again. However, it should be noted that the condition number for the direct method on the BC speech is very large, which is not suitable for LP analysis. This is a new observation on the BC speech. On the other hand, when the SC method is used on the BC speech, the condition number 62.01 is smaller than the cases of the direct and pre-emphasis methods on the AC speech; 71.54 and 64.96. This could mean that the SC method on the BC speech provides more reliable results for power spectrum estimation than the conventional LP based approaches on the AC speech.

Figures 9 and 10 show an example of the power spectra obtained by the three methods from a frame of female speaker. Figures 11 and 12 do the counter parts from a frame of male speaker. Through these Figures, it is observed that the curve by the SC method is commonly more close to that of the true power spectrum obtained by FFT from the original speech signal, especially in the [1000-4000] Hz region. This result validates that the SC method is useful on the real speeches as well. The excellent performance of the SC method is anticipated from the condition number in Table 7.

Although the BC speech produces severe ill-conditioning, Figures 10 and 12 visualize that the SC method provides accurate power spectrum estimation of the BC speech signal.

6 Conclusion

In this paper, we have proposed the SC method to estimate accurately the original power spectrum of the input speech signal. In the proposed method, the PEF is designed more flexibly than the pre-emphasis filter and utilized to compensate for the resulting power spectrum. Experiments through synthetic and real speeches have demonstrated that the SC method provides more accurate power spectrum estimation than the direct and pre-emphasis methods for LP. The BC speech is inherently more ill-conditioned, but the SC method is useful for LP analysis of the BC speech as well.

References:

[1] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 2002.

[2] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, Wiley, 2009.

[3] J. Makhoul, Linear Prediction: A Tutorial Review, *Proc. IEEE,* Vol. 63, No. 4, 1975, pp. 561-580.

[4] J. D. Markel, Digital Inverse Filtering- a New Tool for Format Trajectory Estimation, *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, No. 2, 1972, pp. 129-137.

[5] B. S. Atal and S. Hanauer, Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, *J. Acoust. Soc. Amer.*, Vol. 50, No. 2, 1974, pp. 637-655.

[6] S. Chandra and W. C. Lin, Experimental Comparison Between Stationary and Nonstationary Formulations of Linear Prediction Applied to Voiced Speech Analysis, *IEEE Trans. on Acoust., Speech, Signal Processing*, Vol. ASSP-22, No. 6, 1974, pp.403-415.

[7] P. Kabal, Ill-Conditioning and Bandwidth Expansion in Linear Prediction of Speech, *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing Acoust*, 2003, pp. 824-827.

[8] S. V. Parter, On the Extreme Eigenvalues of Truncated Toeplitz Matrices. *Bulletin of Amer. Math. Soc.*, Vol. 67, 1961, pp. 191-196.

[9] H. Kesten, On the Extreme Eigenvalues of Translation Kernels and Toeplitz Matrices, *J. d'Analyse Math.*, Vol. 10, 1962, pp. 117-138.

[10] G. Fant, J. Liljencrants and Q. G. Lin, A Four Parameter Model of Glottal Flow, *Quart. Progress and Status Rep., Speech Transmission Lab*, Royal Inst. Technol., 1985, pp. 1-13.

[11] H. Strik, Automatic Parameterization of Differentiated Glottal Flow: Comparing Methods by Means of Synthetic Flow Pulses, *J. Acoust. Soc. Amer.*, Vol. 103, No. 5, 1998, pp. 2659-2669.

[12] S. Stenfelt and R. Goode, Transmission Properties of Bone Conducted Sound: Measurements in Cadaver Heads, *J. Acoust. Soc. Amer.*, Vol. 118, No. 4, 2005, pp. 2373-2391.

[13] M. McBride, P. Tran, T. Letowski and R. Patric, The Effect of Bone Conduction Microphone Locations on Speech Intelligibility and Sound Quality, *Applied Ergonomics*, Vol. 42, No. 3, 2011, pp. 495-502.