Accurate Power Spectrum Estimation of Speech with Spectrum Compensation 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 linear prediction (LP) method to estimate accurately the original power spectrum of the input speech signal. A prediction error filter (PEF) is used as a pre-processor, and the LP based power spectrum estimation is compensated by the frequency characteristics of the designed PEF. Through experiments on synthetic vowels, we show that the proposed spectrum compensation method can estimate the power spectrum more accurately than the direct and pre-emphasis LP methods.

Key–Words: Linear prediction, prediction error filter, formant frequency, spectrum compensation.

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 auto-regressive model to be stable [1], [3]. The performance is, however, degraded in an ill-conditioned environment 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]. In this paper, we present a spectrum compensation (SC) method for LP to deal with the problem. To obtain an accurate representation of speech power spectrum, a prediction error filter (PEF) is used as a pre-processor. The followed LP provides an estimate of power spectrum. Unlike the conventional LP, however, the resulting power spectrum is compensated by the spectrum characteristics the PEF possesses. Through experiments, the performance of the proposed SC method is investigated.

This paper is organized as follows. We describe the conventional LP methods in Section 2 and derive the SC method in Section 3. Section 4 shows experimental results obtained using synthetic speeches. Finally, in Section 5 a conclusion is drawn.

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, $\sigma^2_s$, and the prediction coefficients $d_i$, ($i = 1, 2...M_s$), where $M_s$ is the prediction order. The power spectrum is estimated as

$$P_s(\omega) = \frac{\sigma^2_s}{1 + \sum_{i=1}^{M_s} d_i e^{-j\omega T}}$$  \hspace{1cm} (1)$$

where $\omega$ is the angular frequency. However, since $s(n)$ has a certain spread of power spectrum, in many cases a pre-emphasis filter whose transfer function is represented by

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

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

\[
x(n) = s(n) - \eta s(n-1).
\]

The coefficient \( \eta \) 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 of \( \text{LP} \) is described as

\[
P_x(\omega) = \frac{\sigma_x^2}{1 + \sum_{i=1}^{M_s} a_i e^{-j\omega T_i}}
\]

where \( a_i \) are the prediction coefficients, \( \sigma_x^2 \) is the prediction error power and \( M_s \) 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-procesor. The PEF filter is realized as an 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

\[
H_{PEF}(z) = 1 + \sum_{i=1}^{L} b_i z^{-i}
\]

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 \( \eta \) 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,
Amplitude (dB)

-30
-20
-10
0
10
20
30
40

Frequency (Hz)

0 500 1000 1500 2000 2500 3000 3500 4000

Figure 4: Spectrum of the synthetic vowel

\[ P_y(\omega) = \frac{\sigma_y^2}{1 + \sum_{i=1}^{M_y} c_i e^{-j\omega T}} \]  \hfill (6)

where \( c_i \) are the prediction coefficient, \( \sigma_y^2 \) is the prediction error power and \( M_y \) the predictor order. Here, from the PEF used as the pre-processor, we calculate

\[ B_s(\omega) = |1 + \sum_{i=1}^{L} b_i e^{-j\omega T}|^2. \]  \hfill (7)

Then the power spectrum in (6) is compensated as

\[ P_e(\omega) = \frac{P_y(\omega)}{B_s(\omega)}. \]  \hfill (8)

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

\[ P_y(\omega) = |H_{PEF}(e^{j\omega T})|^2 P_s(\omega), \]  \hfill (9)

from (7) and (9), we can find that \( P_e(\omega) \) 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 by (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 \( R_s \). The degree of ill-conditioning of \( R_s \) is measured by the magnitude of the condition number defined by

\[ C_s = \frac{\lambda_{s,max}}{\lambda_{s,min}} \]  \hfill (10)

where \( \lambda_{s,max} \) and \( \lambda_{s,min} \) correspond to the maximum and minimum eigenvalues of \( R_s \). 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, the computation to obtain \( B_s(\omega) \) is significantly less than that to do \( P_y(\omega) \).

4 Experiments

To validate the performance of the proposed SC method, we conducted experiments. 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 three formants (F1, F2 and F3) and their corresponding bandwidth (B1, B2 and B3) used to generate the synthetic vowels. The sampling frequency was 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

\[ B = \frac{1}{\pi f_s} \int_0^{\pi f_s} [P(\omega) - \hat{P}(\omega)]/P(\omega) d\omega. \]  \hfill (11)

where \( P(\omega) \) and \( \hat{P}(\omega) \) denote the true power spectrum and estimated power spectrum, respectively. We calculated the power spectrum using fast Fourier transform. As an example, we show in Figure 4 the power spectrum of a synthetic vowel signal (vowel /a/). It can be seen from Figure 4 that the power spectrum estimated by the SC method (dotted line) is 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 estimation accuracy was observed. The five vowels were used in twenty different window positions. 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 frequencies from the spectrum. First three formants frequency picking were achieved by localizing the spectrum maxima from the envelope. Each formant frequency of $F_1$, $F_2$ and $F_3$ was averaged for five vowels. Table 4 lists the averaged formant estimation errors in percentage for the methods to be compared. It can be seen from Table 4 that the estimation error made by the SC method are smaller in magnitude than the other two methods. Table 4 shows that the SC method provides an improvement relative to the pre-emphasis method. Especially, this indicates that a first order filter for the pre-emphasis is not enough for obtaining accurate formant frequency estimates.

## 5 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 vowels have demonstrated that the SC method provides more accurate power spectrum than the direct and pre-emphasis methods for LP.

### References:


