Linear Prediction Method Evaluation in Speech Formant Frequencies Estimation

Med Ali KAMMOUN, Dorra GARGOURI, Mondher FRIKHA and Ahmed BEN HAMIDA

Equipe Traitement du Signal et Electronique Medicale
Laboratoire d’Electronique et de la Technologie de l’Information ‘LETI’
ENIS, Universite de Sfax – Tunisie

E-mail: kammoun_med_ali@yahoo.fr, tgdorra@yahoo.fr

Abstract

In this paper an improved method is presented to estimate the first four formant frequencies from LPC analysis. The presumed method which computes prediction coefficients has been implemented with Matlab and was applied to the problem of accurate measurement of formant frequencies.

The conceived algorithm estimate formant frequencies from the all pole model of the vocal tract transfer function. The approach relies on the source – filter model supposing that the speech signal can be considered to be the output of a linear system. In fact, the vocal tract shape is considered as the “filter” that filters the excitation to produce the speech signal. The frequency response of the filter has different spectral characteristics depending on the shape of the vocal tract. The spectral peaks in the spectrum are the resonances of the vocal tract and are commonly referred to as formants. The linear prediction analysis is the traditional method used to compute the model of the vocal tract. The obtained result, i.e. prediction coefficients, was then used to estimate formant frequencies.

Results showed that there is a narrow range in the estimated values of formant frequencies for male and female speakers. Such LP method evaluation validates the use of this technique for the accuracy estimation of formant frequencies.

Keywords: Speech production, Vocal tract, Formant estimation, Linear Prediction Coefficients.

I. Introduction

In LP model of speech, each complex pole pair corresponds to a second order resonator. The resonance frequency of each pole is associated with a peak in spectral energy or a formant candidate. Although in the long run automatic formant analysis of speech has received considerable attention and a variety of approaches have been developed, the calculation of accurate formant features from the speech signal is still considered a non-trivial problem. The accuracy of formant tracking using the conventional frame-based LPC analysis is affected by following factors [1]:

1. The number of LPC model coefficients.
2. The influence of pitch (glottal formant) on the first formant.
3. Formant merging.
4. Rapid formant variation that may occur in consonant vowel transitions or diphthongs.
5. Source-vocal tract interaction (pre-condition in LPC analysis).
6. Effects of lips radiation and internal loss on formant bandwidth and frequency.

The purpose of the present article is to evaluate the LP method for the estimation of the first four formant frequencies for both male and female speakers. This research could be hence useful for specific application regarding precision for formant frequencies.

The outline of this paper is as follow. In the next section, the employed LP method used for formant speech features estimation is discussed. In section 3, the experimental results are described. And finally, in section 4, a conclusion of this study is stated.

II. Vocal Tract Filtering and formant frequencies

The vocal tract is made up of the oral cavity from the larynx to the lips and the nasal passage coupled with the oral passage (through the vellum). The oral tract can take on various different configurations depending upon the shape and movement of the tongue, mouth, teeth, lips, jaw, etc. The vocal tract provides frequency shaping to the output from the larynx and also generates new sources for sound production (impulsive source).

The vocal tract can be modeled as a linear filter with resonances. The resonance frequencies of the vocal tract are called formant frequencies. Graphically, the peaks of the vocal tract response correspond roughly to its formant frequencies. Therefore, if the vocal tract is modeled as a time-invariant, all-pole linear system, then each of the conjugate pole pairs corresponds to a formant frequency (resonance frequency).
Figure 1 shows a complete discrete-time speech production model for periodic, noisy and plosive speech. \( G(z) \) is the Z-transform of the glottal flow input, \( R_g(z) \) is the radiation impedance modeled by a single zero and \( 1/H(z) \) is the stable all-pole vocal tract transfer function. \( A_v, A_n, \) and \( A_i \) are the gains that controls the loudness of the sound for periodic, noisy and plosive sources respectively. \( R_l(z) \) models the radiation impedance of the lips.

The vocal tract transfer function \( 1/H(z) \) varies with the type of sound produced and also depends on the speakers and their speaking style. The formant frequencies vary with different vocal tract configurations and therefore, formant frequencies vary in speech with time as the vocal tract changes its shape.

The peaks of the vocal tract response in each configuration correspond roughly to its formant frequencies \([1,2,3,9]\). The first resonance of the vocal tract is called the first formant frequency (or \( F_1 \)), the second resonance of the vocal tract is called the second formant frequency (or \( F_2 \)), and so on. For perfectly voiced and periodic speech (as in sustained vowels) the vocal tract can be accurately modeled by the stable all-pole model for \( 1/H(z) \). However, in order to model other types of sounds, zeros are also added to \( 1/H(z) \) in order to model the nasal cavity of the vocal tract. The resonances or peaks of the vocal tract transfer function (poles of the \( 1/H(z) \) transfer function) correspond roughly to the formant frequencies of a particular sound.

To solve the transfer function of the vocal tract, parametric modeling, and so linear prediction, has been invented. This class of technique attempt to optimally model the spectrum as an autoregressive process \([1,2]\), and then discuss the implications in formant frequencies estimation.

The aim of linear prediction is to estimate the transfer function of the vocal tract from the speech. The signal model can be defined as:

\[
s(n) = -\sum_{i=1}^{N_{LP}} a_{LP}(i) \cdot s(n-i) + e(n) \quad (1)
\]

Where \( N_{LP}, a_{LP} \) and \( e(n) \) represent, respectively, the number of coefficients in the model the linear prediction coefficients and the error in the model.

Equation (1) can be written in Z-transform notation as a linear filtering operation:

\[
E(z) = H_{LP}(z) \cdot S(z) \quad (2)
\]

\( E(z) \) and \( S(z) \) represent, respectively, the Z-transform of the error signal and the speech signal, and \( H_{LP}(z) \) is defined as the linear prediction inverse filter:

\[
H_{LP}(z) = 1 + \sum_{i=1}^{N_{LP}} a_{LP}(i) z^{-i}
\]

or,

\[
H_{LP}(z) = \sum_{i=0}^{N_{LP}} a_{LP}(i) z^{-i} \quad (3)
\]

We would like to minimise the mean squared error so the coefficients of equation (3) can be obtained from the following matrix equation:

\[
\hat{a}_{LP} = \Phi^{-1} \hat{\phi} \quad (4)
\]

Where,

\[
\hat{a}_{LP} = [a_{LP}(1), \ldots, a_{LP}(N_{LP})]^T \quad (5)
\]
Predictor coefficients can be computed on following two basic ways: covariance methods based on the covariance matrix and the autocorrelation methods based on the autocorrelation function.

In speech analysis, the autocorrelation method is almost exclusively used for the reason that this method always produces a predictor filter whose zeros lie inside the unit circle in the z-plane.

In the autocorrelation method, we define $\phi_n$ as:

$$\phi_n(j, k) = \phi_n(0, |j - k|)$$  (8)

Or, the autocorrelation method as:

$$R_n(k) = \frac{1}{N_n} \sum_{m=0}^{N-1-k} s(n + m) \cdot s(n + m - k)$$  (9)

This simplification results by constraining the evaluation interval to the range $[0,N-1]$, and assuming values outside this range are zero [2].

The simplification below, allows the predictor coefficients to be computed efficiently using the Levinson-Durbin algorithm (1947) [1,2].

The transfer function $1/HLP(z)$ is a FIR whitening filter for the speech. The frequency response for this can be computed as the FT of the filter coefficients, then inverted to give the frequency response of $HLP(Z)$ (see Figure 2).

Formant frequencies can be estimated from the LP smoothed spectrum. From this spectrum, local maxima are found and those of small bandwidths are related to formants [3].

Peak-picking can then be used to estimate formants, but this method provides a significant improvement over the accuracy that would be expected from an attempt to pick peaks from the unprocessed speech spectrum.

However, we will use another way to estimate formant frequencies based on the relationship between formant and poles of the vocal tract filter [11].

The denominator of the transfer function may be factored:

$$1 + \sum_{i=1}^{N_{LP}} a_{LP}(i) z^{-i} = \prod_{k=1}^{N_{LP}} (1 - c_k \cdot z^{-1})$$  (10)

Where $C_k$ are a set of complex numbers, with each complex conjugate pair of poles representing a resonance at frequency:

$$\hat{R}_k = \left( \frac{F_{ \pi } \cdot c_k}{2 \pi} \right) \tan^{-1} \left( \frac{\text{Im}(c_k)}{\text{Re}(c_k)} \right)$$  (11)

And bandwidth:

$$\hat{B}_k = -\left( \frac{F_{ \pi } \cdot c_k}{\pi} \right) \ln |C_k|$$  (12)

If the pole is close to the unit circle then the root represents a formant:

$$r_k = \sqrt{\text{Im}(c_k)^2 + \text{Re}(c_k)^2} \geq 0.7$$  (13)

III. Experiments and results

The speech data (16 kHz sampling frequency) used in this study pertains to the TIMIT speech corpus.

The choice of the TIMIT data base, for testing speech analysis algorithms, is justified by the fact that it contains labeled and segmented speech from a great number of speakers.

For our experiment, we used ten different subjects from each sex. All speakers read the same text (sa1.wav). From the 22 different vowels and diphthongs that are present in the TIMIT phoneme database we have selected six vowels. These vowels are [ih, ix, aa, ux, iy, y].

Formant frequencies can be estimated from the LP smoothed spectrum. From this spectrum, local maxima are found and those of small bandwidths are related to formants [3].

Peak-picking can then be used to estimate formants, but this method provides a significant improvement over the accuracy that would be expected from an attempt to pick peaks from the unprocessed speech spectrum.

However, we will use another way to estimate formant frequencies based on the relationship between formant and poles of the vocal tract filter [11].

The denominator of the transfer function may be factored:

$$1 + \sum_{i=1}^{N_{LP}} a_{LP}(i) z^{-i} = \prod_{k=1}^{N_{LP}} (1 - c_k \cdot z^{-1})$$  (10)

Where $C_k$ are a set of complex numbers, with each complex conjugate pair of poles representing a resonance at frequency:

$$\hat{R}_k = \left( \frac{F_{ \pi } \cdot c_k}{2 \pi} \right) \tan^{-1} \left( \frac{\text{Im}(c_k)}{\text{Re}(c_k)} \right)$$  (11)

And bandwidth:

$$\hat{B}_k = -\left( \frac{F_{ \pi } \cdot c_k}{\pi} \right) \ln |C_k|$$  (12)

If the pole is close to the unit circle then the root represents a formant:

$$r_k = \sqrt{\text{Im}(c_k)^2 + \text{Re}(c_k)^2} \geq 0.7$$  (13)

III. Experiments and results

The speech data (16 kHz sampling frequency) used in this study pertains to the TIMIT speech corpus.

The choice of the TIMIT data base, for testing speech analysis algorithms, is justified by the fact that it contains labeled and segmented speech from a great number of speakers.

For our experiment, we used ten different subjects from each sex. All speakers read the same text (sa1.wav). From the 22 different vowels and diphthongs that are present in the TIMIT phoneme database we have selected six vowels. These vowels are [ih, ix, aa, ux, iy, y].
Table 1. Mean values of formant frequencies (in Hz) for English vowels by male and female speakers

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ih</td>
<td>412.33</td>
<td>466.54</td>
<td>1835.39</td>
<td>2329.84</td>
</tr>
<tr>
<td>ix</td>
<td>648.09</td>
<td>732.29</td>
<td>1894.44</td>
<td>2070.54</td>
</tr>
<tr>
<td>aa</td>
<td>700.92</td>
<td>724.81</td>
<td>1493.98</td>
<td>1985.36</td>
</tr>
<tr>
<td>ux</td>
<td>366.23</td>
<td>417.65</td>
<td>1439.98</td>
<td>1985.36</td>
</tr>
<tr>
<td>iy</td>
<td>336.87</td>
<td>407.80</td>
<td>1984.94</td>
<td>2313.68</td>
</tr>
<tr>
<td>y</td>
<td>253.39</td>
<td>356.26</td>
<td>2205.18</td>
<td>2478.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3206.33</td>
<td>3375.80</td>
</tr>
</tbody>
</table>

Table 2. Coefficient of deviation between male and female formant

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ih</td>
<td>0.8838</td>
<td>0.7878</td>
<td>0.9420</td>
<td>1.0002</td>
</tr>
<tr>
<td>ix</td>
<td>0.8850</td>
<td>0.9149</td>
<td>0.9013</td>
<td>0.9383</td>
</tr>
<tr>
<td>aa</td>
<td>0.9670</td>
<td>0.8771</td>
<td>0.8905</td>
<td>0.8967</td>
</tr>
<tr>
<td>ux</td>
<td>0.8769</td>
<td>0.7525</td>
<td>0.9260</td>
<td>0.8929</td>
</tr>
<tr>
<td>iy</td>
<td>0.8261</td>
<td>0.8579</td>
<td>0.9421</td>
<td>0.8869</td>
</tr>
<tr>
<td>y</td>
<td>0.7113</td>
<td>0.8898</td>
<td>0.9498</td>
<td>1.0071</td>
</tr>
</tbody>
</table>

Table 3. Standard deviation of formant frequencies for English vowels by male and female speakers

All LP coefficients (12 coefficients) have been computed from pre-emphasized speech signal using 512 points Hamming windowed speech frames. Formant frequency candidates are calculated by solving the prediction polynomial using Levinson-Durbin algorithm. Only poles agreeing with equation 13 are considered as formant candidates.

Various experiments have been carried out on a set of wav files selected from the TIMIT corpus. We tested the formant estimation algorithm on different male and female subjects. For each vowel pronounced by each speaker, we extracted the first four formant frequencies. The mean values of formant frequencies for each speaker are summarized in Table 1.

As can be seen there is considerable subject to subject variability in the measurements of formant frequencies.

The coefficient of deviation $k_i = \frac{F_{i,\text{male}}}{F_{i,\text{female}}}$ between the corresponding male and female formant is indicated in Table 2.

Formant frequencies of male speakers should be lower than those of female ($k_i<1$). Reference to Table 2 shows that 91.66% of cases justify this result. Generally, as the length of the vocal tract increases the formant frequencies decrease, so the formant frequencies of adult males are somewhat lower than those of adult females, for the same sound [8].

In comparison with other algorithm using Cesptrum smoothed spectrum, the formant estimation algorithm, based on LP coefficients, proves more accuracy in the measurement of formants F3 and F4 ($k_i>1$).
In order to confirm the accuracy of our algorithm, the standard deviation has been computed. The results are summarized in Table 3.

After examining results of Table 3, we can deduce that there is a narrow range in the estimated values of formant frequencies. However, we remark that the standard deviation increases with the order of the formant.

The previous results allow us to collect an important explanation about the LPC method; in comparison with the Cepstral method, the LP algorithm is a practical way to estimate formant of the speech signal especially at high frequencies.

IV. Conclusion

We presented in this paper LP method evaluation in speech formant frequencies estimation. The LP model computed was generated using the autocorrelation method based on the Levinson-Durbin recursion.

Vowel data were collected for 10 speakers from each sex, and were analyzed using the LP method which is believed to reflect vocal tract resonances. Significant variations among the speakers were observed for all the acoustic measures. The data collected for each gender were compared. In agreement with predictions based on theoretical models and previous reports of physiological, we found that formant frequencies for male speakers were lower than those of female.

V. References


