Optimization of Input Parameters for Estimation of LP Coefficients for Isolated Word Recognition

Mohammad Saleem
Research Scholar
Department of Computer Science,
Jamia Millia Islamia,
New Delhi-110 025 INDIA.
mdsaleem@gmail.com

Abdul Mobin
Scientist
Central Electronics Engineering Research Institute, Delhi Centre,
CSIR Complex, New Delhi-110 012.
amobin_ansari@yahoo.co.in

Khurram Mustafa
Reader
Department of Computer Science, Jamia Millia Islamia,
NewDelhi-110 025 INDIA
kmfarooki@yahoo.com

ABSTRACT
Linear Prediction Coding is most popular technique which provides compact representation of vocal tract configuration. The performance of the LPC based spoken word recognition systems are closely related to the input parameters; LP order, frame size, analysis window and windowing technique for the estimation of LP coefficients. In this paper an attempt has been made to optimize the input parameters by performing number of experiments considering the variation of input parameters. The 10^{th} order linear prediction coefficient is found as an optimal order for the isolated word recognizer using the Dynamic Time Warping (DTW). The frame size and analysis window is also optimized for the estimation of LP Coefficients. The use of windowing technique like; Hanning window and Hamming window is also experimented. A speech database of 10 English digits spoken by 10 speakers was used to perform the whole speech recognition experiment.

Keywords
Analysis Window, Hamming Window, Prediction order, Dynamic Time Warping.

1. Introduction
Linear prediction (LP) forms an integral part of almost modern day speech coding algorithms. The fundamental idea is that a speech sample can be approximated as a linear combination of past samples. Within a frame, the weights used to compute the linear combination are found by minimizing the mean-squared prediction error; the resultant weights, or linear prediction coefficients (LPC), are used to represent the particular frame [5]. Linear Prediction can also be viewed as a redundancy removal procedure where information repeated in an event is eliminated, therefore, there is no need for a data if it can be predicted.

Paper Identification Number : CI-1.8
This peer-reviewed paper has been published by the Pentagram Research Centre (P) Limited. Responsibility of contents of this paper rests upon the authors and not upon Pentagram Research Centre (P) Limited. Copies can be obtained from the company for a cost.

By displacing the redundancy in a signal, the amount of bits required to carry the information is lowered, therefore achieving the purpose of compression and processing.

2. Data Acquisition & Database
The speech signal captured by the microphone is pre-emphasized, amplified and passed through the anti-aliasing filter etc. The database is created in *.wav format using a multimedia PC. To perform the recognition experiments, we have created a database of 10 English digits (zero through nine) spoken by several male speakers.

3. Linear Prediction Coefficients
Linear prediction is described as a system identification problem, where the parameters of the model are estimated from the signal itself [2]. A linear prediction is used to predict signal s[n] based on the M past samples; this is done with

$$s[n] = -\sum_{i=1}^{M} a_i * s[n-1],$$

Where a_i are the estimate of the model parameters and are referred to as the linear prediction coefficients (LPC), M is the prediction order. Therefore, prediction is based on a linear combination of the M past samples of the signals, and hence the prediction is linear [3]. The prediction error is equal to

$$e[n] = s[n] - \hat{s}[n]$$

That is, it is the difference between the actual sample and the predicted one [11].

4. Choice of Parameters
In order to perform linear prediction analysis, some basic parameters must be chosen. The variations of these parameters results in varying performance. The following set of parameters is used for the estimation of LP Coefficients.

4.1 Filter Order
It is necessary to find the minimum order of the LP analysis required for modeling the significance of speech. When the speech spectrum is modeled, the vocal tract resonances or the

formants are important. To model the vocal tract resonances the memory of filter A(z) must be at least the twice the time required for the sound wave to travel from glottis to lips. This time interval is 2L/c, where L is the length of the vocal tract (usually 17cm) and c is the velocity of the sound wave (340 m/s). So, the memory should be at least 1 ms. when the sampling frequency is 8 kHz, 1 ms memory means using 8 previous samples. Thus, the order of the filter should be at least 8. It was found from experimental results that if the sampling frequency (fs) is expressed in kHz then the number of poles should be fs plus 4 or 5, this agrees with the simulation results [4]. Since the sampling frequency is 8 kHz, a very high prediction gain is found with a 12th order or a 13th order LP analysis. The choice of the order is a compromise among the spectral accuracy or quality of sound, computational time, memory of filter and transmission bandwidth.

4.2 Frame Size

The choice of the frame length basically depends on whether the analysis is done on a transient speech segment or quasi-periodic speech segment. The analysis should be done in an interval where the vocal tract movement is negligible. Usually, for most vowels, a 15-20 ms analysis frame is sufficient, but some may have significant movement in that time period. In the case of unvoiced speech, the length of the interval should be smaller than 15-20 ms [8]. For example, a burst associated with the release of an unvoiced stop consonant in the initial position exists only for few ms. In order to accommodate that change, a smaller interval like 10 ms is needed. The frame length may be expressed in terms of the number of samples by multiplying the sampling frequency fs by the time interval. The LP analysis is done by varying the frame length. Usually the speech signal is stationary in short interval, such as 20 ms. Consequently; we have analyzed the spoken word recognizer.

4.3 Analysis Window

In practice, windows have finite length and by shifting that finite length window, different regions of the speech signal can be examined. According to the choice of window size:

- The length of windows should be short enough so that the speech properties of interest change minimally within the windows.
- The windows length should be long enough to allow the calculation of the desired parameters.

When the analysis is periodically repeated, successive windows should not be so short that the solution of S(n) is omitted. It implies that the window length must be greater than or equal to the frame length, otherwise some parts of the signal will not be analyzed. Usually the frame length is about $2/3^{rd}$ of the window length, so that the successive windows overlap by 33%, which is logical, especially when w(n) has a shape that de-emphasizes speech samples near its edges. Typically w(n) is smooth, because its values are the weighting factor of s(n), and priori all samples are equally relevant. Because of the windowing distortion, the LP window should include at least two pitch periods for accurate spectral estimate. Typically a 20-30 ms window includes two periods even at low F0 (fundamental frequency). The major difficulty with short windows arises from unpredictability of the speech excitation signals u(n).

The prediction gain increase after the window length is increase above the frame length and it reaches a fairly high value when the window length overlapped approximately 33%.

4.4 Window offset

In the LPC analysis, the optimization comes to picture when windows are overlapped by 33% [6]. The window offset is the parameters, which define the position of the last samples of the window relative to the position of the last samples of the speech frame. This parameter is used to align the window with respect to the frame. In one of particular case, a window offset takes the value 125 (window end 125 samples after the last samples of the speech frame) while frame length is 250 samples and window length is 375 samples.

4.5 Windowing

Windowing means multiplying the speech signal s(n) by a window w(n), which allows us to weight the speech samples in different ways. A rectangular window has an abrupt discontinuity at the edge in the time domain. As a result there are large side lobes and undesirable ringing effects in the frequency domain representation of the rectangular window. To discard the large oscillations, we should use a window without abrupt discontinuities in the time domain known as Hamming Window, which corresponds to low side lobes of the window in the frequency domain. It is actually a raised cosine function:

$$W(n) = \begin{cases} 0.54\text{-}0.46\cos\left(\frac{2\Pi n}{N-1}\right), & 0<n<N-1 \\ 0, & \text{otherwise} \end{cases}$$

There are other types of tapered windows, such as the Hanning window, which used the function as

$$W(n) = \begin{cases} 0.5(1-\cos\left(\frac{2\Pi n}{N-1}\right)), & 0<n<N-1 \\ 0, & \text{otherwise} \end{cases}$$

5. Spoken Word Recognizer

In this study, the dynamic time warping based recognizer is used to perform the speech recognition experiment. The dynamic programming approach has been used in order to minimize the speaking rate variation effectively as follows: [1]

Let $A=a_1, a_2, ..., a_i, ..., a_I$ and
$B=b_1, b_2, ..., b_j, ..., b_J$

be time-sampled speech patterns, where a_i and b_j are time-sampled feature vectors of A and B respectively, and $d(i, j) = a_i - b_j$ be distance between a_i and b_j. By calculating recurrence relation

$$g(i, j) = d(i, j) + \text{Min} \begin{cases} g(i, j-1) \\ g(i-1, j) \\ g(i-1, j-1) \end{cases}$$

up to (I, J) with initial condition

$$g(1, 1) = d(1, 1)$$

the similarity between A and B is obtained as

$$S(A, B) = g(I, J)/(I+J)$$

Actually, the combination of a and b in the above mentioned recurrence relation is restricted within the domain called adjustment window defined by

$$W= \{I(i, j)I Ii-jI < = r\}$$

Where a positive integer r is appropriately chosen so that the timing variation is practically limited within the window. Introduction of the restriction window reduces the amount of computation considerably. Also further reduction of computation was achieved when a rejection threshold was used to reject the words whose cumulative distances exceed the specified threshold level.

Pattern Classification

The last major step in the pattern-recognition model of speech recognition is the decision rule, which chooses which (reference) pattern (or patterns) most closely matches the unknown test pattern [7]. It is based on the cumulative distances obtained from the recurrence relation of the reference and test patterns as follows:

$$\text{distance} = \sum_{q=1}^{F} S(A_q, B_q)$$

Where, F is the order of LPC and $S(A_q, B_q)$ is the similarity between A_q and B_q (reference and test) time-sampled speech patterns with respect to the q^{th} LP coefficient.

6. Performance Evaluation

The recognition performance of the spoken word recognizer was evaluated under well defined test procedures based on its different input parameters such as filter order, window size, and window types etc. However, overall recognition score varies from 92% to 97% for a chosen values of filter order, window size and analysis window. Some cases are described here in short in the form of tables.

CASE I

In this case we are considering the analysis window of 20ms with a 5ms of overlapping. The testing were performed by using hamming window.

Table 1: Optimization of LP order based on accuracy and computation per recognized word.

Filter Order	Recognition Score in %	Computation per Word
8	92	7200
9	95	8100
10	**97**	**9000**
12	96	10800
14	95	12600
16	94	14400

CASE II

This case considers the optimized filter order and analyses the variation in the frame size and overlapping window. The 10^{th} order filter is used to perform the experiments using hamming window.

Table 2: Optimization of Analysis Window

Frame size	Analysis Window	Recognition Score in %
15	15	95
15	**20**	**97**
20	20	93
20	25	96
25	25	94
25	30	95

CASE III

The final case deals with windowing technique (hamming /hanning window) used for the removal of side lobes in the signal due to rectangular window. Recognition score on these two different windowing techniques has been analyzed. The 10^{th} order filter and analysis window of 20ms has been chosen.

7. Conclusion

A progressive enhancement in recognition accuracy has been achieved for a DTW based spoken word recognizer. The 10^{th} order linear prediction coefficient is found to be very much suitable for this purpose with the analysis window of 20ms having overlapped with 5ms. The implementation effect of the hamming and hanning window on the experiment performed in this study does not show any significant improvement on the recognition accuracy of the recognizer.

8. References

[1] Abdul Mobin, S.S. Agrawal, “Role of filter-bank features in DTW based word recognition for saving the computational time and rejecting the non-vocabulary word”, Int. Conf. on Systemics Cybernetics and Informatics (SCI 2001.ISAS 2001), Volume XIII, 2001.

[2] Einarsson T. & Thorsteinsson J. “Speech Linear Prediction and Synthesis”, University of Maryland, Project 2 for ENEE624 – Fall 2002.

[3] Islam T., “Interpolation of Linear Prediction Coefficients for Speech Coding”, ME Thesis, McGill University, Montreal, Canada, 2000.

[4] J.D. Markel and AH Gray, “Lianear Prediction of Speech”, Berlin, Germany: Springer-Verlag, 1976.

[5] Jakimovski B., Gligoroski D., “Methods for Digital Signal Processing of Speech Signal“, 2nd Int. Conf. CiiT, Molika, 189-196, Dec 2001.

[6] Johnson M.H. and Alwan A., "Speech Coding: Fundamentals & Applications", to appear as a chapter in the Encyclopedia of Telecomms, Wiley, December 2002.

[7] L.R Rabiner, A.E Rosenberg and S.E Levinson, "Considerations in Dynamic Time Warping Algorithm for Discrete Word Recognition", IEEE Trans. ASSP, Vol. 26, 1978, pp. 562-575.

[8] Picone J, “Signal Modeling Techniques in Speech Recognition”, Proceedings IEEE, 1993.

[9] Rabiner L. and Schafer W., "Digital Processing of Speech Signals", Prentice- Hall, Englewood Cliffs, NJ, 1978.

[10] Rabiner L, Juang B., “Fundamentals of Speech Recognition”, Prentice Hall Signal processing Series, 1993.

[11] Wai, C. Chu, “Speech Coding Algorithms: Foundation and Evolution of Standardized Coders”, John Willey & Sons, 2003.