Model for Vocal Tract Filter

Sonia Mittal¹, Rakesh Garg², Sonia Saini³
SJPM, Radaur, India¹, KITM, Kurukshetra, India², SJPM ,Radaur, India³
¹Sonia.ece.2010@gmail.com, ²rakesh435@rediffmail.com, ³Sonial886@gmail.com

Abstract: In this paper, LPC is used to modelled the transfer function of speech signal. Lpc use for estimating speech parameters like pitch, formants, and spectrum. The principle behind the use of lpc is to minimize the sum of square difference between original speech signal and the estimated speech signal over a finite duration. This could be used to give a unique set of predictor coefficients. The two most commonly used method to compute the coefficients are covariance and correlation. For our implementation we are using autocorrelation. The reason is that this method is superior in the sense that roots of the polynomial in the denominator of the above equation is always guaranteed to be inside the unit circle, guaranteeing the stability of the system .Levinson Durbin recursion will be utilized to compute the required parameters for autocorrelation.

Keywords: vocal chord, vocal tract, glottal pulse; voiced sound; unvoiced sound; lpc;framing.

I. INTRODUCTION

Vocal tract filter is modeled to estimate the source of speech using lpc technique. The parametric technique lpc is used to express the estimated glottal excitation in numerical form. The source-filter theory of speech production provides the theoretical background for the inverse filtering. By calculating the lpc coefficients using levison Durbin algorithms, the source of speech is identified.

Measurements of the vibrating vocal folds is difficult, as the oscillation can neither be examined directly due to the hidden position of the vocal folds nor be observed without aids because of the high speed of the oscillations. Visual analysis is widely used especially in clinical investigation of voice production. Several techniques, such as video stroboscopy, digital high-speed stroboscopy and kymography have been developed, and many of them are currently used in daily practices in voice clinics. In most GIF methods, the estimation of the vocal tract and lip radiation can be computed solely from the acoustical speech pressure signal, which makes the analysis fully non-invasive. GIF[1] is typically used in association with parameterization of the estimated glottal excitation in order to describe voice production quantitatively.

II. RELATED MODEL

A pulse train produced by the glottis which provides the frequency information of the signal[2-3]. The vocal tract, which acts as a filter to change the overall characteristics of the pulse train. The LPC coefficients identify formants in the given frame that represent the harmonics generated by the vocal tract filter. The state of the vocal tract filter is represented by the LPC coefficients.

A. Physical Model

The human voice production mechanism [4] can be roughly divided into three parts: lungs, vocal folds, and vocal tract. The lungs function as a source of air flow and pressure. The vocal tract functions as an acoustic filter that shapes the spectrum of the sound. Finally, sound is radiated to the surrounding air at the lips and nostrils. Lungs are equivalent to source of sound & vocal tract equivalent to filter called source filter model [5] for sound production.

Fig 1.1 Human Voice Production System

Working-

Air is pushed from your lung through your vocal tract and out of your mouth comes speech. For certain voiced sound, your vocal cords vibrate (open and close).The rate at which the vocal cords vibrate determines the pitch of your voice. Women and young children tend to have high pitch (fast vibration) while adult males tend to have low pitch (slow vibration).For certain fricatives and plosive (or unvoiced) sound, your vocal cords do not vibrate but remain constantly opened .The shape of your vocal tract determines the sound that you make. As you speak, your vocal tract changes its shape producing different sound.

B. Mathematical Model

A simplified manner to study the functioning of the human speech production mechanism is to categorize speech sounds into three main classes according to
the production mechanism. Voiced sounds[6], which are excited by the fluctuation of the vocal folds. Unvoiced sounds, where the sound excitation is turbulent noise.

![Mathematical Model](image)

\[ S(n) = g(n) * h(n) * l(n) \]

In Frequency Domain:
\[ S(z) = G(z) * H(z) * L(z) \]

\[ G(z) = \frac{S(z)}{L(z) * H(z)} \]

C. Block Diagram
Glottal inverse filtering (GIF) [1] refers to methods of estimating the source of voiced speech, the glottal volume waveform. GIF is based on the idea of inversion, in which the effects of the vocal tract and lip radiation are cancelled from the output of the voice production mechanism, the speech signal.

![Block Diagram Of Speech Production System](image)

The production of a speech pressure signal, denoted by \( S(z) \), as a cascade of three processes: (i) glottal excitation, \( G(z) \) (ii) vocal tract filtering, \( H(z) \) and (iii) lip radiation, \( L(z) \). Estimation of the glottal excitation with GIF corresponds to cancelling the filtering effects of the vocal tract and lip radiation from the speech signal.

In Time Domain:
\[ s(n) = g(n) * h(n) * l(n) \]

In Frequency Domain:
\[ S(z) = G(z) * H(z) * L(z) \]

D. Various Forms of LPC Filter:
(a) Transfer Function:
\[ H(Z) = \frac{1}{1+a_1z^{-1}+a_2z^{-2}+\ldots+a_{10}z^{-10}} \]

(b) Input-Output relationship of the filter is given by the linear difference equation:
\[ s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n) \]

(c) In vector form:
\[ A = (a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, G, V, UV, T) \]

A changes every 20 msec or so. At a sampling rate[7] of 8000 samples/sec, 20 msec is equivalent to 160 samples. The digital speech signal is divided into frames of size 20 msec. There are 50 frames/second.

(d) Relation B/W linear difference equation & vector form:
\[ A = (a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, G, V, UV, T) \]

is equivalent to
\[ S=(s(0), s(1), \ldots, s(159)) \]

Thus the 160 values of \( S \) is compactly represented by the 13 values of \( A \).

III. LPC ANALYSIS
Consider one frame of speech signal:
\[ S=(s(0), s(1), \ldots, s(159)) \]

The signal \( s(n) \) is related to the innovation \( u(n) \) through the linear difference equation[8]:
\[ s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n) \]

The ten LPC parameters \( (a_1, a_2, \ldots, a_{10}) \) are chosen to minimize the energy of the innovation:
\[ f = \sum_{n=0}^{159} u^2(n) \]

Using standard calculus, we take the derivative of \( f \) with respect to \( a_i \) and set it to zero:
\[ \frac{df}{da_i} = 0 \]

We now have 10 linear equations with 10 unknowns:
where

\[ R(k) = \sum_{n=0}^{159-k} s(n)s(n+k) = \text{autocorrelation of } s(n) \]

The above matrix equation could be solved using:
1. The Gaussian elimination method.
2. Any matrix inversion method (MATLAB).
3. The Levinson-Durbin recursion.

A. Levinson-Durbin recursion algorithm

\[ E^{(i)} = R(0) \]

\[ k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E^{(i-1)}} \quad i = 1, 2, 3, \ldots, 10 \]

\[ a_i^{(i)} = k_i \]

\[ a_j^{(i)} = a_j^{(i-1)} - k_i a_j^{(i-2)} \quad j = 1, 2, \ldots, i - 1 \]

\[ E^{(i)} = (1 - k_i^2) E^{(i-1)} \]

Solve the above for \( i = 1, 2, 3, \ldots, 10 \) and then set

\[ a_i = -a_i^{(10)} \]

The Levinson-Durbin (L-D) Algorithm [9] is a recursive algorithm that is considered very computationally efficient since it takes advantage of the properties of \( R \) when determining the filter coefficients. The filter order is denoted with a superscript, \( a_j^{(0)} \) for a \( j \)th order filter, and the average mean squared error of a \( j \)th order filter is denoted \( E_j \) instead of \( E[e_n^2] \).

D. Framing

The speech samples are divided into 30-ms window frames. Each 30 ms window frame consists of 660 samples as illustrated below:

(Sampling Rate)(Frame Length) = Number of Samples in a Frame

(22000 samples/second)(0.030 second) = 660 samples.

E. Preemphasis

The pre-emphasis[10] is a low-order digital filter, which increases the power in the upper frequency bands of the speech signal. The FIR filter is similar to a high pass filter. The Finite Impulse Response (FIR)[11] filters increase the magnitudes of certain frequencies that are characteristic of the vowel sound.

\[ H(z) = 1 - 0.9375 z^{-1} \]

Once signal conversion is complete, the last step of digital post filtering is most often executed using a Finite Impulse Response (FIR) filter given as

\[ H_{pre}(z) = \sum_{k=0}^{N_{pre}} a_{pre}^{(k)} z^{-k} \]

Normally, a one coefficient digital filter known as pre-emphasis filter.

\[ H_{pre}(z) = 1 + a_{pre} z^{-1} \]

A typical range of values for \( a_{pre} \) is [-1.0,-0.4]. The preemphasis filter boosts the signal spectrum approximately 20 dB per decade.

Advantages of preemphasis filter:
1. The voiced sections of speech signal naturally have a negative spectral slope, approximately 20 Db per decade due to physiology of speech production system.
2. The hearing is more sensitive above the 1-kHz region of the spectrum. The preemphasis filter amplifies this area of the spectrum. This assists the spectral analysis algorithm in modelling the perceptually important aspects of speech spectrum.
The equation to form a Hamming window is shown below:

\[ W(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{239} \right), \quad 0 \leq n \leq 599 \]

Equation generates 660 discrete points for the Hamming window. For each 30ms window frame, 660 samples are multiplied point by point with the 660 discrete Hamming window points. The hamming window is used to gradually taper the window frame to zero at its beginning and end boundaries. Therefore, the signal discontinuities at the beginning and end of each frame are minimized. For example, if a signal of finite length is passed through a filter of finite length, the beginning and end of the filtered signal depends on the samples before and after the signal. Since they are not of infinite lengths, the output should be weighted heavily in the middle.

### III. RESULTS

#### Technical Details:
- Length = 37840
- Sampling Rate = 22050 samples/second
- Bit Resolution = 16 bits/sample
- Total Duration = 1.716 sec
- Total Samples = 37837.8
- Duration of 1 Frame = 0.03 ms
- No. of Frames = 57.2
- No. of Samples in One Frame = 660 = 22050 * 0.03
- Original speech signal = 38,567 samples.
- We are analysing 10,000 (20,000 to 30,000) samples out of 38,567 samples.
- 1 frame = 660 samples
- Samples under observation = 10,000
- Total frame to be observed = 10,000 / 660 = 15 frames.
- Corresponding frames are: 20,000 to 20660, 20660 to 21320, 21320 to 21980, 21980 to 22640, ……., 29240 to 29900.

We have calculated lpc coefficients for each frame by applying pre emphasis, windowing, autocorrelation to each frame. Then average lpc coefficients of 15 frames is calculated. Using average lpc coefficients, pole-zero plot is done for transfer function.
LPC coefficients of corresponding frames: 20,000 to 20660, 20660 to 21320, 21320 to 21980, 21980 to 22640, 22640 to 29240 to 29900 are following:

1. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -2.8698 \ 4.3288 \ -4.5575 \ 4.5118 \ -4.8744 \ 4.5118 \ -4.8744 \ 4.6390 \ -2.0386 \ -0.9273
   \]
   
   \[
   0.2773
   \]

2. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -3.1882 \ 5.9740 \ -8.0514 \ 9.0835 \ -9.3095 \ 8.7116 \ -7.3583 \ 5.0244 \ -2.4558
   \]
   
   \[
   0.6437
   \]

3. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -3.1627 \ 5.7900 \ -7.6219 \ 8.5528 \ -8.8327 \ 8.3458 \ -7.1496 \ 5.0005 \ -2.5172 \ 0.6636
   \]

4. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -3.1675 \ 5.7930 \ -7.6043 \ 8.4748 \ -8.6943 \ 8.1377 \ -7.2687 \ 4.7734 \ -2.3977
   \]
   
   \[
   0.6434
   \]

5. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -3.1832 \ 5.9365 \ -7.9892 \ 9.1109 \ -9.4469 \ 8.8304 \ -7.3782 \ 5.0064 \ -2.4616
   \]
   
   \[
   0.6481
   \]

6. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -2.9741 \ 5.3046 \ -6.8192 \ 7.5925 \ -7.7990 \ 7.2687 \ -6.0649 \ 4.0464 \ -1.9535
   \]
   
   \[
   0.4799
   \]

7. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -2.4356 \ 3.8665 \ -4.4408 \ 4.8240 \ -4.8861 \ 4.3770 \ -3.4675 \ 2.0462 \ -0.8776
   \]
   
   \[
   0.1068
   \]

8. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -2.1947 \ 3.0714 \ -3.0824 \ 3.2633 \ -3.3539 \ 2.8696 \ -2.0863 \ 1.0370 \ -0.3855
   \]
   
   \[
   0.0281
   \]

9. \( \text{lpcoefs} = \)
   
   \[
   1.0000 \ -2.6406 \ 4.2089 \ -4.8254 \ 5.1417 \ -5.2844 \ 4.9508 \ -4.1624 \ 2.7205 \ -1.2770
   \]
   
   \[
   0.2462
   \]
(10) lpcoefs =
1.0000 -2.5266  3.8117 -4.1156  4.2494 -  
4.4301  4.2494 -3.6098  2.3040 -1.0331  
0.1825
(11) lpcoefs =
1.0000 -2.7813  4.2390 -4.5416  4.6831 -  
4.9087  4.7649 -3.9630  2.6143 -1.3037  
0.3520
(12) lpcoefs =
1.0000 -2.5464  3.5936 -3.4301  3.1858 -  
3.4529  3.4383 -2.7227  1.4249 -0.4960  
0.0557
(13) lpcoefs =
1.0000 -2.5685  2.7264 -1.1135  0.2070 -  
1.4382  2.5356 -2.1781  1.1796 -0.5431  
0.2310
(14) lpcoefs =
1.0000 -1.5618  0.5740  0.4833  0.0974 -  
1.4409  0.8896  0.2904 -0.6292  0.1065 -  
0.2057
(15) lpcoefs =
1.0000 -1.7718  1.6540 -0.7044  0.3859 -  
0.9437  1.0880 -0.8740  0.2009 -0.0117 -  
0.0229

Avg lpc:
1.0000 -2.638  4.058 -4.625  4.883 -5.272 -  
5.002 -4.113  2.669 -1.24  0.312

\[ H(z) = \frac{1}{1+2.638z^{-1}-4.058z^{-2}+4.625z^{-3}-4.883z^{-4}} \]
+5.272z^{-5}-5.002z^{-6}+4.113z^{-7}-2.669z^{-8}+1.24z^{-9}-0.312z^{-10} 

REFERENCES