L7: Linear prediction of speech

Introduction
Linear prediction
Finding the linear prediction coefficients
Alternative representations

This lecture is based on [Dutoit and Marques, 2009, ch1; Taylor, 2009, ch. 12; Rabiner and Schaefer, 2007, ch. 6]
Introduction

Review of speech production

– Speech is produced by an excitation signal generated in the throat, which is modified by resonances due to the shape of the vocal, nasal and pharyngeal tracts

– The excitation signal can be

  • Glottal pulses created by periodic opening and closing of the vocal folds (voiced speech)
    – These periodic components are characterized by their fundamental frequency ($F_0$), whose perceptual correlate is the pitch
  • Continuous air flow pushed by the lungs (unvoiced speech)
  • A combination of the two

– Resonances in the vocal, nasal and pharyngeal tracts are called formants
– On a spectral plot for a speech frame
  
  • Pitch appears as narrow peaks for fundamental and harmonics
  • Formants appear as wide peaks in the spectral envelope

[Dutoit and Marques, 2009]
Linear prediction

The source-filter model

– Originally proposed by Gunnar Fant in 1960 as a linear model of speech production in which glottis and vocal tract are fully uncoupled

– According to the model, the speech signal is the output $y[n]$ of an all-pole filter $1/A(z)$ excited by $x[n]$

\[
Y(z) = X(z) \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = X(z) \frac{1}{A_p(z)}
\]

  • where $Y(z)$ and $X(z)$ are the z transforms of the speech and excitation signals, respectively, and $p$ is the prediction order

– The filter $1/A_p(z)$ is known as the synthesis filter, and $A_p(z)$ is called the inverse filter

– As discussed before, the excitation signal is either

  • A sequence of regularly spaced pulses, whose period $T_0$ and amplitude $\sigma$ can be adjusted, or
  • White Gaussian noise, whose variance $\sigma^2$ can be adjusted
[Dutoit and Marques, 2009]
The above equation implicitly introduces the concept of linear predictability, which gives name to the model.

Taking the inverse $z$-transform, the speech signal can be expressed as

$$y[n] = x[n] + \sum_{k=1}^{p} a_k y[n - k]$$

which states that the speech sample can be modeled as a weighted sum of the $p$ previous samples plus some excitation contribution.

In linear prediction, the term $x[n]$ is usually referred to as the error (or residual) and is often written as $e[n]$ to reflect this.
Inverse filter

- For a given speech signal $x[n]$, and given the LP parameters $\{a_i\}$, the residual $e[n]$ can be estimated as

$$e[n] = y[n] - \sum_{k=1}^{p} a_k y[n - k]$$

- which is simply the output of the inverse filter excited by the speech signal (see figure below)

- Hence, the LP model also allows us to obtain an estimate of the excitation signal that led to the speech signal

  - One will then expect that $e[n]$ will approximate a sequence of pulses (for voiced speech) or white Gaussian noise (for unvoiced speech)

[Dutoit and Marques, 2009]
Finding the LP coefficients

How do we estimate the LP parameters?

- We seek to estimate model parameters \( \{a_i\} \) that minimize the expectation of the residual energy \( e^2(n) \)

\[
\{a_i\}_{\text{opt}} = \arg \min [e^2(n)]
\]

- Two closely related techniques are commonly used
  - the covariance method
  - the autocorrelation method
The covariance method

– Using the term $E$ to denote the sum squared error, we can state

$$E = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} \left( y[n] - \sum_{k=1}^{p} a_k y[n - k] \right)^2$$

– We can then find the minimum of $E$ by differentiating with respect to each coefficient $a_i$ and setting to zero

$$\frac{\partial E}{\partial a_j} = 0 \Rightarrow \sum_{n=0}^{N-1} \left( 2 \left( y[n] - \sum_{k=1}^{p} a_k y[n - k] \right) y[n - j] \right) =$$

$$= -2 \sum_{n=0}^{N-1} y[n]y[n - j] + 2 \sum_{n=0}^{N-1} \sum_{k=1}^{p} a_k y[n - k]y[n - j] = 0$$

∀ $j = 1, 2, ... p$

– which gives

$$\sum_{n=0}^{N-1} y[n]y[n - j] = 2 \sum_{k=1}^{p} a_k \sum_{n=0}^{N-1} y[n - k]y[n - j]$$
– Defining $\phi(j, k)$ as

$$\phi(j, k) = \sum_{n=0}^{N-1} y[n - j]y[n - k]$$

– This expression can be written more succinctly as

$$\phi(j, 0) = \sum_{k=1}^{p} \phi(j, k)a_k$$

– Or in matrix notation as

$$\begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \phi(p,0) \end{bmatrix} = \begin{bmatrix} \phi(1,1) & \phi(1,2) & \phi(1,p) \\ \phi(2,1) & \phi(2,2) & \phi(2,p) \\ \phi(p,1) & \phi(p,2) & \phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_p \end{bmatrix}$$

– or even more compactly as $\Phi = \Psi a$

– Since $\Phi$ is symmetric, this system of equations can be solved efficiently using Cholesky decomposition in $O(p^3)$
- **NOTES**
  
  - This method is known as the covariance method (for unclear reasons)
  
  - The method calculates the error in the region $0 \leq n < N - 1$, but to do so uses speech samples in the region $-p \leq n < N - 1$
    - Note that to estimate the error at $y[0]$, one needs samples up to $y[-p]$
  
  - No special windowing functions are needed for this method
  
  - If the signal follows an all-pole model, the covariance matrix can produce an exact solution
    - In contrast, the method we will see next is suboptimal, but leads to more efficient and stable solutions
The autocorrelation method

- The autocorrelation function of a signal can be defined as
  \[ R(n) = \sum_{m=-\infty}^{\infty} y[m]y[n - m] \]

- This expression is similar to that of \( \phi(j, k) \) in the covariance method but extends over to \( \pm \infty \) rather than to the range \( 0 \leq n < N \)
  \[ \phi(j, k) = \sum_{n=j}^{\infty} y[n]y[n - k] \]

- To perform the calculation over \( \pm \infty \), we window the speech signal (i.e., Hann), which sets to zero all values outside \( 0 \leq n < N \)

- Thus, all errors \( e[n] \) will be zero before the window and \( p \) samples after the window, and the calculation of the error over \( \pm \infty \) can be rewritten as
  \[ \phi(j, k) = \sum_{n=0}^{N-1+p} y[n]y[n - k] \]

- which in turn can be rewritten as
  \[ \phi(j, k) = \sum_{n=0}^{N-1-(j-k)} y[n]y[n + j - k] \]
– thus, $\phi(j, k) = R(j - k)$

– which allows us to write $\phi(j, 0) = \sum_{k=1}^{p} \phi(j, k) a_k$ as

$$R(j) = \sum_{k=1}^{p} R(j - k) a_k$$

– The resulting matrix

$$\begin{bmatrix}
    R(1) \\
    R(2) \\
    \vdots \\
    R(p)
\end{bmatrix} =
\begin{bmatrix}
    R(0) & R(1) & R(p - 1) \\
    R(1) & R(0) & R(p - 2) \\
    \vdots & \vdots & \vdots \\
    R(p-1) & R(p-2) & R(0)
\end{bmatrix}\begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
a_p
\end{bmatrix}$$

– is now a Toeplitz matrix (symmetric, with all elements on each diagonal being identical), which is significantly easier to invert

• In particular, the Levinson-Durbin recursion provides a solution in $O(p^2)$
Speech spectral envelope and the LP filter

- The frequency response of the LP filter can be found by evaluating the transfer function on the unit circle at angles $2\pi f / f_s$, that is

$$|H(e^{j2\pi f / f_s})|^2 = \left| \frac{G}{1 - \sum_{k=1}^{p} a_k e^{-j2\pi k f / f_s}} \right|^2$$

- Remember that this all-pole filter models the resonances of the vocal tract and that the glottal excitation is captured in the residual $e[n]$

- Therefore, the frequency response of $1/A_p(z)$ will be smooth and free of pitch harmonics

- This response is generally referred to as the spectral envelope
How many LP parameters should be used?

– The next slide shows the spectral envelope for $p = \{12, 40\}$, and the reduction in mean-squared error over a range of values

  • At $p = 12$ the spectral envelope captures the broad spectral peaks (i.e. the harmonics), whereas at $p = 40$ the spectral peaks also capture the harmonic structure

  • Notice also that the MSE curve flattens out above about $p = 12$ and then decreases modestly after

– Also consider the various factors that contribute to the speech spectra

  • Resonance structure comprising about one resonance per 1Khz, each resonance needing one complex pole pair

  • A low-pass glottal pulse spectrum, and a high-pass filter due to radiation at the lips, which can be modeled by 1-2 complex pole pairs

  • This leads to a rule of thumb of $p = 4 + f_s/1000$, or about 10-12 LP coefficients for a sampling rate of $f_s = 8kHz$
(a) Comparison of STFT with $H(e^{j2\pi f/f_s})$

![Graph showing comparison of STFT with $H(e^{j2\pi f/f_s})$]

- Log magnitude (dB)
- Frequency in kHz ($f$)
- STFT, $p = 12$, $p = 40$

(b) Normalized Mean-Squared Prediction Error

![Graph showing normalized mean-squared prediction error]

$V(p) = \frac{E(\phi)}{\phi[0]}$

$n$

[Ref: Rabiner and Schafer, 2007]
Examples

ex7p1.m

- Computing linear predictive coefficients
- Estimating spectral envelope as a function of the number of LPC coefficients
- Inverse filtering with LPC filters
- Speech synthesis with simple excitation models (white noise and pulse trains)

ex7p2.m

- Repeat the above at the sentence level
Alternative representations

A variety of different equivalent representations can be obtained from the parameters of the LP model

- This is important because the LP coefficients \( \{a_i\} \) are hard to interpret and also too sensitive to numerical precision
- Here we review some of these alternative representations and how they can be derived from the LP model
  - Root pairs
  - Line spectrum frequencies
  - Reflection coefficients
  - Log-area ratio coefficients
- Additional representations (i.e., cepstrum, perceptual linear prediction) will be discussed in a different lecture
Root pairs

- The polynomial can be factored into complex pairs, each of which represents a resonance in the model
  - These roots (poles of the LP transfer function) are relatively stable and are numerically well behaved
- The example in the next slide shows the roots (marked with a ×) of a 12-th order model
  - Eight of the roots (4 pairs) are close to the unit circle, which indicates they model formant frequencies
  - The remaining four roots lie well within the unit circle, which means they only provide for the overall spectral shaping due to glottal and radiation influences
Fig. 1 LPC spectral speech frame with LSPs overlaid

[Rabiner and Schafer, 2007]

[McLoughlin and Chance, 1997]
Line spectral frequencies (LSF)

- A more desirable alternative to quantization of the roots of $A_p(z)$ is based on the so-called line spectrum pair polynomials
  
  \[
  P(z) = A(z) + z^{-(p+1)}A(z^{-1})
  
  Q(z) = A(z) - z^{-(p+1)}A(z^{-1})
  \]

  - which, when added up, yield the original $A_p(z)$

- The roots of $P(z)$, $Q(z)$ and $A_p(z)$ are shown in the previous slide

  - All the roots of $P(z)$ and $Q(z)$ are on the unit circle and their frequencies (angles in the z-plane) are known as the line spectral frequencies

  - The LSFs are close together when the roots of $A_p(z)$ are close to the unit circle; in other words, presence of two close LSFs is indicative of a strong resonance (see previous slide)

  - LSFs are not overly sensitive to quantization noise and are also stable, so they are widely used for quantizing LP filters
Reflection coefficients

– The reflection coefficients represent the fraction of energy reflected at each section of a non-uniform tube model of the vocal tract

– They are a popular choice of LP representation for various reasons
  • They are easily computed as a by-product of the Levinson-Durbin iteration
  • They are robust to quantization error
  • They have a physical interpretation, making them amenable to interpolation

– Reflection coefficients may be obtained from the predictor coefficients through the following backward recursion

\[
\begin{align*}
    r_i &= a_i^i \quad \forall i = p, \ldots, 1 \\
    a_j^{i-1} &= \frac{a_j^i + a_i^i a_{i-j}^i}{1 - r_i^2} \quad 1 \leq j < i
\end{align*}
\]

• where we initialize \( a_i^p = a_i \)
**Log-area ratios**

– Log-area ratio coefficients are the natural logarithm of the ratio of the areas of adjacent sections of a lossless tube equivalent to the vocal tract (i.e., both having the same transfer function)
  
  • While it is possible to estimate the ratio of adjacent sections, it is not possible to find the absolute values of those areas

– Log-area ratios can be found from the reflection coefficients as

\[ A_k = \ln \left( \frac{1 - r_k}{1 + r_k} \right) \]

• where \( g_k \) is the LAR and \( r_k \) is the corresponding reflection coefficient