CONTENTS

1. Analysis and Synthesis of Pole-Zero Speech Models
2. Short-time Fourier Transform analysis and synthesis
3. Filter-Bank Analysis/Synthesis
4. Sinusoidal Analysis/Synthesis

1. Analysis and Synthesis of Pole-Zero Speech Models

- For idealized voiced speech, the transfer function relating the acoustic pressure at the lips output and the volume velocity at the glottis contains both poles and zeros:

  - poles corresponds to the resonances of the vocal tract cavity,
  - two poles (outside the unit circle) to model the glottal airflow,
  - one zero inside the unit cycle for the radiation load at the lips.

- other additional zeros to represent the energy-absorbing anti-resonances from the back cavity during unvoiced plosives and fricatives.

Etc.
1.1 All-Pole Modeling of Speech Signals

- We assume that contributions from the glottal flow, vocal tract, and radiation load to the transfer function can be modeled by an all-pole filter.

\[ H(z) = AG(z)V(z)R(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \]  

(1-1)

where \( G(z) \), \( V(z) \) and \( R(z) \) are respectively the glottal flow, vocal tract, and radiation load contributions to the transfer function.

- Although \( R(z) \) is a single zero inside the unit circle, it is approximated by an all-pole model. This is motivated by the following expansion:

\[ 1 - a z^{-1} = \frac{1}{\Sigma_{k=0}^{p} a^k z^{-k}} = \frac{1}{\Pi_{k=0}^{p} (1 - b_k z^{-k})}, \quad |z| > |a| \]  

(1-2)

- Since \( a^k \to 0 \) as \( n \to \infty \), it can be approximated by a finite set of poles.
1.2 Linear Prediction

- The above all-pole model (or autoregressive (AR) model) is closely related to linear prediction analysis (LPA).

- The basic idea of LPA is to predict each speech sample from a linear combination of past speech samples.

- To see this, let $s[n]$ and $u[n]$ be the output and input of $H(z)$. Then

$$H(z) = \frac{S(z)}{U(z)} = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \text{ (AR model)} \quad (1-3)$$

where $S(z)$ and $U(z)$ are the z-transforms of $s[n]$ and $u[n]$, respectively.

- Taking the inverse z-transform, we have the following time-domain result:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Au[n] = \tilde{s}[n] + Au[n]. \quad (1-4)$$
In other words, $s[n]$ can be “predicted” from a linear combination of $s[n-k]$, $k=1,...,p$, and the error or prediction residual is $Au[n]$, that is the source $H(z)$.

- The value $p$ is called the order of the linear predictor $\tilde{s}[n]$. $a_k$ are called the linear prediction coefficients.

- The transfer function between $\tilde{s}[n]$ and $s[n]$ is $P(z)$:

$$\tilde{s}[n] = \sum_{k=1}^{p} a_k s[n-k] \leftrightarrow \tilde{S}(z) = P(z) \cdot S(z)$$

(1-5)

where $P(z) = \sum_{k=1}^{p} a_k z^{-k}$.

- The transfer function between the error and $s[n]$ is $A(z)=1-P(z)$

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k] \leftrightarrow E(z) = (1 - P(z)) \cdot S(z).$$

(1-6)

$A(z)$ is called the inverse filter because it recovers the input $u[n]$. 
This is illustrated as follows:

![Diagram of linear prediction](image)

**Fig. 5.1** Filtering view of linear prediction.

**Example:** For \( s[n] = A\delta[n] \ast h[n] = Aa^n u[n] \), i.e. a unit impulse passing through a 1\(^{st}\) order all-pole filter with pole \( a \).

\[
e[n] = s[n] - as[n - 1] = A(a^n u[n] - aa^{n-1} u[n-1]) = A\delta[n],
\]

which is exactly the input.
1.3 Least Squares (LS) Estimation

- We now estimate the linear prediction coefficients $a_k$ by minimizing the Least Squares criterion:

$$E_n = \sum_{m=-\infty}^{\infty} (s_n[m] - \tilde{s}_n[m])^2 = \sum_{m=-\infty}^{\infty} e_n^2[m],$$

over a short segment of the speech waveform, where

$$e_n[m] = s_n[m] - \sum_{j=1}^{P} a_k s_n[m-k], \quad n-M \leq m \leq n+M$$

and zero elsewhere.

$s_n[m]$ is defined in the vicinity of time $n$ over the interval $[n-M-p,n+M]$.

- This leads to a set of linear equations in $a_k$ and the gain $A$. 


(a) prediction at time $n_0$.

(b) samples in the vicinity of time $n$, i.e., samples over the interval $[n - M, n + M]$.

(c) samples required for the prediction of samples in the interval $[n - M, n + M]$.

Fig. 5.2 Short-time sequence used in linear prediction analysis.
To minimize $E_n$ over $a_k$, we set the derivatives of $E_n$ with respect to each variable $a_k$ to zero:

$$\frac{\partial E_n}{\partial a_i} = 0, \quad i = 1, 2, 3, \ldots, p.$$  \hfill (1-9)

This gives

$$\frac{\partial E_n}{\partial a_i} = \frac{\partial}{\partial a_i} \sum_{m=-\infty}^{\infty} \left( s_n[m] - \sum_{k=1}^{p} a_k s_n[m-k] \right)^2$$

$$= 2 \sum_{m=-\infty}^{\infty} \left( s_n[m] - \sum_{k=1}^{p} a_k s_n[m-k] \right) \left( -\frac{\partial}{\partial a_i} \sum_{k=1}^{p} a_k s_n[m-k] \right).$$  \hfill (1-10)

Since

$$\left( -\frac{\partial}{\partial a_i} \sum_{k=1}^{p} a_k s_n[m-k] \right) = -s_n[m-i],$$  \hfill (1-11)

we then have for $\partial E_n / \partial a_i = 0$ the following:

$$0 = 2 \sum_{m=-\infty}^{\infty} \left( s_n[m] - \sum_{k=1}^{p} a_k s_n[m-k] \right) (-s_n[m-i]).$$  \hfill (1-12)
Multiplying through gives

\[ \sum_{m=\infty}^{\infty} s_n[m]s_n[m-i] = \sum_{k=1}^{p} a_k \sum_{m=\infty}^{\infty} s_n[m-i]s_n[m-k], \quad 1 \leq i \leq p. \quad (1-13) \]

Define the function

\[ \Phi_n[i,k] = \sum_{m=\infty}^{\infty} s_n[m-k]s_n[m-i], \quad 1 \leq i \leq p, \quad 1 \leq k \leq p. \quad (1-14) \]

Then, (1-13) can be rewritten as

\[ \Phi_n[i,0] = \sum_{k=1}^{p} a_k \Phi_n[i,k], \quad i=1,2,3,\ldots, p. \quad (1-15) \]

This is a system of linear equations, sometimes referred to as the normal equations, in the unknown \( a_k \), and can be written in matrix form as:

\[ \Phi \alpha = b, \quad (1-16) \]

where \( b = [\Phi_n[1,0],\ldots,\Phi_n[p,0]]^T \) and \( [\Phi]_{i,k} = \Phi_n[i,k] \).
Substituting the optimal values \( a_k \) into (1-7), one gets the minimum squares error as

\[
E_{n,\text{min}} = \Phi_n[0,0] - \sum_{k=1}^{p} a_k \Phi_n[0,k].
\]

(1-17)

- In the covariance method, the samples in the interval \([n-M-p,n+M]\) are used to predict the samples in the interval \([n-M,n+M]\).
- In the autocorrelation method, the samples in the interval \([n-M-p,n-M-1]\) are treated as zero during the prediction. Hence, it only involves the samples in the interval \([n-M,n+M]\). To avoid the shape discontinuity of the prediction residual at the boundaries, the speech samples are usually windowed by multiplying \( s[n] \) by a window \( w[n] \), e.g. Hamming or Hanning windows, located in the interval \([n-M,n+M]\).
The waveform $s[m]$ [panel (a)] is shifted by $n$ samples [panel (b)] and then windowed by an $N_w$-point rectangular sequence $w[m]$ [panel (c)].

Fig. 5.3 Formulation of the short-time sequence in the autocorrelation method.
Fig. 5.4 Example of a third-order predictor in the autocorrelation method of linear prediction: (a) sliding predictor filter; (b) prediction error. Prediction error is largest at the beginning and the end of the interval \([0, N_w + p - 1]\).

The normal equation of the autocorrelation method is

\[
\sum_{k=1}^{p} a_k \Phi_n[i, k] = \Phi_n[i,0], \quad i = 1, 2, 3, \ldots, p. \tag{1-18}
\]

where

\[
\Phi_n[i, k] = \sum_{m=0}^{N_w + p - 1} s_n[m - i]s_n[m - k], \quad 1 \leq i \leq p, \quad 0 \leq k \leq p. \tag{1-19}
\]
After slight manipulation, it can be shown that

\[
\Phi_n[i, k] = \sum_{m=0}^{N_w-1-(i-k)} s_n[m] s_n[m + (i - k)], \quad 1 \leq i \leq p, \quad 0 \leq k \leq p.
\]  
(1-20)

which is a function of only the difference i-k and so we denote it as

\[
\Phi_n[i, k] = r_n[i - k], \quad 1 \leq i \leq p, \quad 0 \leq k \leq p.
\]  
(1-21)

Letting \( \tau = i - k \), called the correlation lag, we see that

\[
r_n[\tau] = \Phi_n[i, k] = \sum_{m=0}^{N_w-1-\tau} s_n[m] s_n[m + \tau] = s_n[\tau] * s_n[-\tau],
\]  
(1-22)

which is the short-time sequence \( s_n[m] \) convolved with itself flipped in time.

\( r_n[\tau] \) is called the autocorrelation function and it is a measure of the “self-similarity” of the signal at different lags \( \tau \).
Hence, (1-18) can be written as

\[ \sum_{k=1}^{p} a_k r_n[i, k] = r_n[i, 0], \quad i = 1, 2, 3, \ldots, p. \quad (1-18) \]

And in matrix form

\[ R_n \alpha = r_n, \quad (1-18) \]

where \( R_n \) is called the autocorrelation matrix and \( r_n \) is called the autocorrelation vector.

- An advantage of the autocorrelation method is that the all-pole filter obtained is always stable. However, due to windowing, the solution is biased even if the input is generated by an all-pole filter.

- Theoretically, the covariance method can give exact estimation in certain cases. However, the stability of the all-pole filter obtained is not guaranteed.
Fig. 5.6 Illustration of autocorrelation functions of speech: (a) vowel /o/ in “pop”; (b) unvoiced plosive /k/ in “baker”; (c) unvoiced fricative /f/ in “father”; (d) voiced plosive /g/ in “go”.
The large prediction errors suggest that all-pole filter alone cannot satisfactorily describe all the speech sound signals. In speech compression or coding, we also need to represent the excitation to the time-varying all-pole filter.
As the order $p$ increases, the harmonic structure of the speech spectrum $S(w)$ is revealed in the all-pole spectrum. On the other hand, too small a value for $p$ will remove the fine resonant structure of the underlying all-pole system. Normally, $p$ is chosen as 10-14 in speech analysis and coding.
1.4 Synthesis Based on All-pole Modeling

Fig. 5.16 Overlap-add synthesis using all-pole model. The waveform is generated frame by frame by convolutional synthesis and the filter output on each frame is overlapped and added with adjacent frame outputs.
Fig. 5.18 Speech reconstruction based on linear prediction analysis using the autocorrelation method: (a) original; (b) synthesized. The reconstruction was performed with overlap-add synthesis and using a 14$^{th}$-order all-pole model.