1 Lecture Outline

Reading: Chapter 5.

- Analysis and synthesis of pole-zero speech models (introduction)
- All-pole modeling of voiced speech (deterministic signals)

2 Introduction

We have developed a CT and DT transfer function for the relation between the acoustic pressure at the lips output and the volume velocity at the glottis–this is one of the major components of our speech production model. For the idealized voiced speech case, the transfer function *contains poles* that correspond to the resonances of the vocal tract cavity. More generally, the transfer function from the glottis to the lips also *includes zeros* that represent the energy-absorbing anti-resonances resulting from the

1. back cavity during unvoiced plosives or fricatives
2. oral passage during nasal consonants
3. nasal passage during nasalized vowels

We seek to develop methods for estimating the parameters of an *all-pole* system function for both deterministic (periodic or impulsive sources) and stochastic (noise sources) sound classes. The approach is referred to as *linear prediction analysis* and the basic idea is that each speech sample is approximated as a linear combination of past speech samples.

Referring back to the source-filter model for speech production, we have

$$S(z) = A[G(z)U_g(z) \text{ or } U_o(z)]V(z)R(z) \tag{1}$$

where $S(z)$ is the z-transform of a short segment of speech or a phoneme,

- source or input is periodic $G(z)U_g(z)$ or noisy $U_o(z)$
  - $G(z)$ denotes the glottal waveform over one cycle. Multiplication with a harmonic spectrum $U_g(z)$ ($u_g[n]$ is an impulse train), yields spectrum of glottal airflow waveform (Loizou Fig. 3.5)
  - The subscript "o" in $U_o(z)$ denotes the turbulence (noise) which can occur anywhere in the oral tract and not necessarily at the glottis
- $V(z)$ spectrally shapes the input (vocal tract)
- $R(z)$ (high pass filter) models acoustic radiation (lips)
- A is a gain factor that takes into account intensity or volume

For voiced speech, we can rewrite the model as

$$\begin{aligned} S(z) &= AG(z)V(z)R(z)U_g(z) \\ S(z) &= H(z)U_g(z) \end{aligned} \tag{2}$$

For unvoiced speech, we can rewrite the model as

$$\begin{aligned} S(z) &= AV(z)R(z)U_o(z) \\ S(z) &= H(z)U_o(z) \end{aligned} \tag{3}$$

In either case, we have a source $U_g(z)$ (harmonic spectrum) or $U_o(z)$ (noise spectrum) and a filter $H(z)$. $H(z)$ models the spectral envelope of the speech spectrum

We wish to efficiently estimate the spectral envelope $H(z)$ from the speech samples. Why? Knowledge of the spectral envelope $H(z)$ allows us to:

- Analyze the speech signal in terms of formants
- Code the speech signal [parameters of $H(z)$]
- Synthesize speech

3 All-Pole Modeling of Deterministic Signals

Consider the transfer function from the glottis to the lips output for deterministic signals. The transfer function consists of glottal flow, $G(z)$; vocal tract, $V(z)$; and radiation load, $R(z)$ contributions to yield

$$H(z) = AG(z)V(z)R(z). \tag{4}$$

Figure 1: Source/filter model

It can be shown that $G(z)$ (glottal waveform) can be modeled as (Quatieri)

$$G(z) = \frac{1}{(1-\beta z)^2}. \tag{5}$$

We assume the vocal tract can be modeled as a resonator so that $V(z)$ is an all-pole transfer function (with poles at the formant frequencies)

$$V(z) = \frac{1}{1-\sum_k v_k z^{-k}}. \tag{6}$$

In Chapter 4, we represented $R(z)$ with a single-zero transfer function (model). Of course, a single-zero transfer function is mathematically equivalent to an all-pole transfer function provided we have an infinite number of poles. To see this note that for $|a| < 1$

$$\left(1-az^{-1}\right)\left(1+az^{-1}+a^2z^{-1}+a^2z^{-1}+\cdots\right) = 1 \tag{7}$$

Thus

$$\begin{aligned} R(z) &= 1-az^{-1} \\ &= \frac{1}{\sum_{k=0}^{\infty} a^k z^{-k}}. \end{aligned} \tag{8}$$

To further simplify, we approximate $R(z)$ with a *finite* number of poles

$$R(z) = \frac{1}{1-\sum_k r_k z^{-k}}. \tag{9}$$

Combining these terms together, we approximate $H(z)$ as an *all-pole* transfer function

$$\begin{aligned} H(z) &= AG(z)V(z)R(z) \\ &= A\left(\frac{1}{(1-\beta z)^2}\right)\left(\frac{1}{1-\sum_k v_k z^{-k}}\right)\left(\frac{1}{1-\sum_k r_k z^{-k}}\right) \\ &= \frac{A}{1-\sum_{k=1}^{p} a_k z^{-k}} \end{aligned} \tag{10}$$

The $\{a_k\}$ are referred to as the model coefficients. Our goal is to estimate the filter coefficients, a_k for a specific order p and gain A. In doing so, we can establish a complete transfer function from the glottis to the lips output for periodic or impulsive source signals.

3.1 Linear Prediction Analysis: Voice Speech

Consider the z-transform of the vocal tract input $u_g[n]$, $U_g(z)$, with gain A and let $S(z)$ denote the z-transform of its output. We have as the I/O relation

$$S(z) = U_g(z)H(z). \tag{11}$$

Relating (10) to (11) we have

$$\frac{S(z)}{U_g(z)} = \frac{A}{1-\sum_{k=1}^{p} a_k z^{-k}}. \tag{12}$$

or with cross-multiplication and rearranging, we have

$$S(z)\left[1-\sum_{k=1}^{p} a_k z^{-k}\right] = AU_g(z) \tag{13}$$

or

$$S(z) = \sum_{k=1}^{p} a_k S(z) z^{-k} + AU_g(z). \tag{14}$$

In the time domain, this is written as

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Au_g[n]. \tag{15}$$

The above equation is referred to as an autoregressive (AR) model. The coefficients a_k are referred to as the *linear prediction coefficients* and their estimation is termed *linear prediction analysis*. Quantization of these coefficients is called *linear prediction coding* (LPC).

Recall that in voicing we have lumped the glottal airflow into the system function so that $u_g[n]$ is a train of unit samples. Therefore, except for the times at which $u_g[n]$ is nonzero (every pitch period), we can think of

Figure 2: Periodic or impulsive sources

$s[n]$ as a *linear combination* only of past values of $s[n]$, i.e. each speech sample is approximated as a *linear combination of past speech samples*

$$s[n] = \sum_{k=1}^{p} a_k s[n-k], \text{when } u_g[n] = 0. \tag{16}$$

Equation (16) predicts $s[n]$ from a linear combination of the past p samples $s[n-1], ..., s[n-p]$.

A *linear predictor* of order p is defined as

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k]. \tag{17}$$

The sequence $\tilde{s}[n]$ is the one-step prediction of $s[n]$ by the sum of p past weighted samples of $s[n]$.

Suppose now that $s[n]$ is the ideal voiced speech sample. If the predictor coefficients equal the model coefficients, i.e. $\alpha_k = a_k$, we can write the prediction error as

$$\begin{aligned} e[n] &= s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \\ &= \sum_{k=1}^{p} a_k s[n-k] + Au_g[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \\ &= Au_g[n]. \end{aligned} \tag{18}$$

Thus $e[n]$ is an impulse train and therefore equal to zero most of the time except at the periodic impulses (pitch times). Likewise, when the speech waveform is a single impulse, as for an idealized plosive sound, then $e[n]$ is zero except at the time of the impulsive input.

Rearranging, we have

$$Au_g[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] = \tag{19}$$

or

$$s[n] = Au_g[n] + \sum_{k=1}^{p} \alpha_k s[n-k] \tag{20}$$

This suggests a method for coding these speech signals: represent $s[n]$ with $\{\alpha_k\}$ and enough information to generate $Au_g[n]$, i.e. pitch period. Then the $\{\alpha_k\}$ can be used to generate $\tilde{s}[n]$ so that the original speech signal can be recreated as $s[n] = \tilde{s}[n] + e[n]$.

The approach of linear prediction analysis is to find a set of prediction coefficients that minimizes the mean-squared (prediction) error (MSE) over a short segment of the speech waveform.

4 Error Minimization (Autocorrelation Method)

One method for estimating the parameters of the all-pole model, referred to as the *autocorrelation method*, assumes zero-valued samples outside the speech window (this is different than in the covariance method). The method does not result in an exact solution for the parameters of the all-pole model, however, this suboptimal method is very efficient computationally.

For a window of N samples of voiced speech whose leading edge is at time n, define the prediction errors within the window as

$$\begin{aligned} e_n[m] &= \text{actual} - \text{prediction} \\ &\equiv s[n-m] - \sum_{k=1}^{p} \alpha_k s[n-m-k], \quad 0 \le m \le N-1 \\ &= s_n[m] - \sum_{k=1}^{p} \alpha_k s_n[m-k]. \end{aligned} \tag{21}$$

We formulate our problem as follows. Find $\{\alpha_k\}$ that minimizes the total squared-error in the window

$$\begin{aligned} J_n &= \sum_{m=0}^{N-1} e_n[m]^2 \\ &= \sum_{m=0}^{N-1} \left(s_n[m] - \alpha^T \mathbf{s}_n[m-1] \right)^2. \end{aligned} \tag{22}$$

Notes:

- $s_n[m]$ denotes the m-th sample in a length N window whose leading edge is at time n
- Any samples in the regressor which are outside the window are assumed to be zero
- Ex. Suppose $N = 10$ and $p = 3$. Clearly $s[1]$ is inside the window but $s[1-3]$ is not.

We expand

$$J_n = \sum_{m=0}^{N-1} \left(s_n[m] - \alpha^T \mathbf{s}_n[m-1] \right) \left(s_n[m] - \mathbf{s}_n^T[m-1]\alpha \right)$$

$$J_n = \sum_{m=0}^{N-1} s_n[m]^2 - 2\alpha^T \sum_{m=0}^{N-1} \mathbf{s}_n[m-1] s_n[m] + \alpha^T \sum_{m=0}^{N-1} \mathbf{s}_n[m-1] \mathbf{s}_n^T[m-1] \alpha$$

$$J_n = \sum_{m=0}^{N-1} s_n[m]^2 - 2\alpha^T \mathbf{r} + \alpha^T \mathbf{R} \alpha \tag{23}$$

where the short-time cross-correlation vector

$$\mathbf{r}_n \equiv \sum_{m=0}^{N-1} \mathbf{s}_n[m-1] s_n[m] \tag{24}$$

and the short-time autocorrelation matrix

$$\mathbf{R}_n \equiv \sum_{m=0}^{N-1} \mathbf{s}_n[m-1] \mathbf{s}_n^T[m-1]. \tag{25}$$

Define the k-th short-time autocorrelation as

$$r_n[k] \equiv \sum_{m=0}^{N-1} s_n[m-k]s_n[m] \tag{26}$$

The short-time cross-correlation vector

$$\begin{aligned} \mathbf{r}_n &\equiv \sum_{m=0}^{N-1} \mathbf{s}_n[m-1]s_n[m] \\ &= [r_n[1], r_n[2], ..., r_n[p]]^T \end{aligned} \tag{27}$$

The short-time autocorrelation matrix is then a Toeplitz matrix

$$\begin{aligned} \mathbf{R}_n &\equiv \sum_{m=0}^{N-1} \mathbf{s}_n[m-1]\mathbf{s}_n^T[m-1] \\ &= \begin{bmatrix} r_n[0] & r_n[1] & \cdots & r_n[p-1] \\ r_n[1] & r_n[0] & \cdots & r_n[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_n[p-1] & r_n[p-2] & \cdots & r_n[0] \end{bmatrix} \end{aligned} \tag{28}$$

Any regressor samples outside the window are assumed zero.

In order to minimize the squared prediction error

$$J_n = \sum_{m=0}^{N-1} s_n[m]^2 - 2\alpha^T \mathbf{r}_n + \alpha^T \mathbf{R}_n \alpha \tag{29}$$

we require the gradient vector to be zero

$$\begin{aligned} \mathbf{0} &= \partial J_n / \partial \alpha \\ &= -2\mathbf{r}_n + 2\mathbf{R}_n \alpha. \end{aligned} \tag{30}$$

We thus have the normal equations

$$\mathbf{R}_n \alpha = \mathbf{r}_n \tag{31}$$

and the solution yields the LPCs

$$\alpha = \mathbf{R}_n^{-1} \mathbf{r}_n \tag{32}$$

where $\mathbf{R}_n$ is given by (28) and $\mathbf{r}_n$ is given by (27).

5 Recap

For short segments of voiced speech, we can compute the parameters (LPCs)

$$\alpha = \mathbf{R}^{-1}\mathbf{r} \tag{33}$$

of an all-pole filter which models the spectral envelope

$$\hat{H}(z) = \frac{A}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}. \tag{34}$$

Since $\mathbf{R}$ is a positive definite matrix, $\mathbf{R}^{-1}$ exists. Furthermore, since $\mathbf{R}$ is Toeplitz there is an efficient algorithm known as the Levinson recursion for solving the normal equations. We still also need to compute the gain, A.