Linear predictive speech coding

Extracting poles from the waveform

Jont Allen

ECE-437
Basic Assumptions

- Speech is generated by glottal pulses of air . . .
- Spectrally shaped by the vocal tract transfer function
- The vocal tract is a transmission line, with an output at the mouth
- When the input and output of a tube are at the ends, the spectrum is “all-pole”
- The voiced speech transfer function is an all-pole response
Basic model of speech generation:
Atal (1971): ... traditional Fourier analysis methods require a relatively long speech segment to provide adequate spectral resolution. As a result, rapidly changing speech events cannot be accurately followed.

This is very true, but a heavy price is paid, when the speech is noisy.

Atal (1971): Although pitch-synchronous analysis by synthesis techniques can provide a partial solution . . . , such techniques are extremely cumbersome and time consuming even for modern digital computers and are therefore unsuitable for automatic processing of large amounts of speech data.

Today pitch-synchronous methods are widely used, due to their much higher quality.
Basic covariance method

- The linear prediction estimate is defined in terms of unknown coefficients $a_k$

$$\hat{s}_n \equiv \sum_{k=1}^{p} a_k s_{n-k} \equiv S_n \cdot A$$

- The prediction error is defined as

$$e_n \equiv s_n - \hat{s}_n$$

- and the total error $E_T$ is minimized to find $A$

$$E_T \equiv \sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} \left( s_n - \sum_{k=1}^{p} a_k s_{n-k} \right)^2$$
Finding the prediction coefficients

- The *normal equations* for prediction coefficients $A$

\[
- \frac{1}{2} \frac{\partial E_t}{\partial a_i}igg|_{i=1,\ldots,p} = \sum_{n=1}^{N} s_n s_{n-i} - \sum_{k=1}^{p} a_k \sum_{n=1}^{N} s_{n-k} s_{n-i} = 0,
\]

\[
\begin{bmatrix}
R_{11} & R_{12} & \cdots & R_{1p} \\
R_{21} & & & \\
\vdots & \ddots & \cdots & \\
R_{p1} & & & R_{pp}
\end{bmatrix}
\begin{bmatrix}
r_1 \\
r_2 \\
\vdots \\
r_p
\end{bmatrix}
\]
Once $A(t)$ is determined, every 5-10 ms, the error $e_n$ is computed.

The error is computed by convolution of $s_n$ with $[1, -A]$

$$e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} = [1, -A] \ast S_n$$

$e_n$ contains everything that was not captured by $A(t)$.
Model of speech generation Atal 71

- The speech may be regenerated exactly from $e_n$

- Or, the speech may be regenerated from the encoded error
Other definitions

- Normal-equations may also be defined in terms of *windowed speech* samples
- or *windowed error*
- All these options, and many more, have been extensively explored, and in great depth
Finding the formants

- The roots of $[1, -A] \equiv [1, -a_1, -a_2, \ldots, -a_p]$ are estimates of the formants of the speech.
- Ideally speaking, the formants are the poles of the vocal tract.
- In practice, these estimates are not very robust.