Lecture 5 - April 16th, 2013

Logistics

Announcements, Assignments, and Reminders

- Visit the URL links that were covered in previous lectures.
Cumulative Outstanding Reading

- Read chapters 1 and 2 in our book (Huang, Acero, Hon, “Spoken Language Processing”).

Outline of today

- Brief overview of Speech Perception theories
- Filter-bank analysis of speech for spectral representations (linear/log scale)
- LPC representations of speech (how to compute LPC coefficients), reflections on an acoustic tube.
Good books (for today)

- our book (Huang, Acero, Hon, “Spoken Language Processing”)
- Goldstein, “Sensation and Perception”
- Moore, “An Intro to the psychology of hearing”
- Pickles, “An Intro to the physiology of hearing”
- Clark & Yallop, “An Intro to Phonetics and Phonology”
- Deller et. al. “Discrete-time Processing of speech signals”
- Ladefoged “A Course in Phonetics”
- Lieberman & Blumstein “Speech physiology, speech perception, and acoustic phonetics”
- K. Stevens, “Acoustic Phonetics”
- Malmberg, “Manual of phonetics”
- Rossing, “The Science of Sound”
- Linguistics 001, University of Pennsylvania

Speech Perception

- Much higher level in auditory cortex or brain - not well understood
- Goal: find invariant acoustic cues for different speech sounds
  - phonemes, phones, syllables, words (monosyllabic, polysyllabic), phrases, sentences, etc. Is there an ideal realization?
- Difficult since acoustic cues for an object change depending on context, other perceptual modalities (McGurk effect), prior beliefs, etc.
Cues depend on context

Real "di" "du" example and spectrogram. file://didu.wav
Cues depend on context

- cues for different “d” sounds are different depending on context
  - F2 totally different
  - This is one of the reasons formants are unreliable to do phone classification
- So even considering acoustics, percept is dependent on more than just acoustic cues at the current time.
- Hence, must be more than just formants that are used to determine words. E.g., temporal patterns, long context, adaptation (time dependence, dependence on what you’ve previously been exposed too), etc.

Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
- Humans interact and influence each other in many different ways.
- Acoustic/Phonetic level: group convergence to common rate-of-speech, pitch range, style of turn timing, loudness quality, prosodic pattern (J. Local, 2003) – phonetic or suprasegmental entrainment.
- Linguistic level: conversations are “easier” to ascertain than monologues (cell phones) because of alignment processes such as entrainment (Brennan & Clark, Pickering & Garrod).
- Behavioral level: gestures, mannerisms (e.g., gait), other social features also eventually converge (Dijksterhuis & Bargh, Wyatt et. al.).
- Could percept change as a function of recent social context (i.e., at signal level, two identical cues might be interpreted entirely different depending on recent social context)?
- Current ASR systems do not fully exploit these phenomena (c.f. adaptation methods later in the course).
Spectral Regions of Speech Perception

- 200-5.5kHz most important — by filtering out spectral regions and measuring intelligibility
- Hence, ISDN: 4kHz BW, early digital speech channel, prior to coding and compression being ubiquitous
- formal perceptual theories to determine intelligibility. E.g., the “Articulation Index” by Steeneken and Houtgast

Ex: filter out < 1kHz, then voicing and manner of articulation discrimination decreases (/p/ vs. /b/ vs. /v/)
- ex: filter out > 1.2kHz, place of articulation discrimination drops (/p/ vs. /t/)
- Telephone bandwidth 200-4kHz (good enough for most intelligibility)
- Particularly bad are the infamous “E-set” /p/, /d/, /e/, /g/, /c/, etc. vowel energy
Human Speech Perception

- we are remarkably good at perceiving speech
- Recall the sine-wave speech, and the 1-bit speech from a few lectures ago.
- face recognition, when do we see a face?
- therefore, hard to identify most important cues, since they all could potentially be used depending on context

- Speech contains much redundant information, much can be removed w/o impacting intelligibility – E.g., checkerboard speech
- Spectral transitions, derivatives might be key?
- Gaussian Scaled Speech (example)
- Apparently, no particular location in time/frequency that contains the crucial information (we can infer what is missing from the other bits)
- Redundancy allows us to perceive speech in many different acoustic situations — e.g., background noise, cocktail party effect
Goal: Produce a representation of speech that contains the “message” of what was said, stripping out what is irrelevant.

Vocal tract: carrier signal (glottal signal), fundamental has F0 from 50-250Hz (males) or 120-500Hz (females), but its harmonics go up to at least 5kHz.

Speech information is contained in relatively slow moving time-varying vocal tract response function. Recall Dudley’s quote from several lectures ago.

So, goal: extract, store, represent and/or process the vocal tract response (and remove spectral response).

Why do this? Useful in all areas of speech processing – some form of filter bank processing is almost always the initial processing steps taken before speech recognition, compression, coding, enhancement, etc.

The vocal tract is a complicated physical system.

Recall Homer Dudley’s Voder system (on the right).
Filter Bank Approach

- Produces useful compressed speech representation.
- Each BPF can be seen as convolution.

\[ s_i(n) = s(n) * h_i(n) = \sum_{m=0}^{M-1} h_i(m)s(n-m) \]  \hspace{1cm} (5.1)

where \( h_i(n) \) is impulse response of \( i^{th} \) filter.

Filter Bank - Aspects

- Filter bandwidths, center frequencies, and filter spacings (spectral resolution)
- Bandpass filter design
- (nonlinear) rectifier
- Low-pass filter and design
- Down sampler
- Amplitude compressor/transformer
Filter Bank - BPF

- BPF separates speech into separate spectral components
- Alternate strategy: Windowed FFT which we will talk about later.

Filter Bank - why rectification?

- There are various forms of rectification.
- Full-wave rectification (essentially absolute value)
  \[ f(s(n)) = |s(n)| = \begin{cases} 
  s(n) & \text{if } s(n) \geq 0 \\
  -s(n) & \text{if } s(n) < 0 
\end{cases} \quad (5.2) \]
- Half-wave rectification
  \[ f(s(n)) = |s(n)| = \begin{cases} 
  s(n) & \text{if } s(n) \geq 0 \\
  0 & \text{if } s(n) < 0 
\end{cases} \quad (5.3) \]
- Square rectification
  \[ f(s(n)) = (s(n))^2 \quad (5.4) \]
- Our analysis will be for the full-wave version, but others have a similar analysis and effect.
How bad distortion is rectification?

Let's try ...

Fs = 1600;
y = audiorecorder(Fs,16,1); % creates the record object.
recordblocking(y,4); % records 4 seconds of speech
s = getaudiodata(y);
sound(s,Fs); % original
sound( abs(s) ); % full-wave rect.
sound( s .* (s >= 0) ); % half-wave rect.
sound( s.^2 ); % square rectification

but of course, we're not rectifying the original speech, we're doing it in sub-bands.

Full-wave analysis

We have that

\[ v(n) = f(s(n)) = s(n)w(n) \]  \hspace{1cm} (5.5)

where

\[ w(n) = \begin{cases} +1 & \text{if } s(n) \geq 0 \\ -1 & \text{if } s(n) < 0 \end{cases} \]  \hspace{1cm} (5.6)

Hence, via the DTFT (mult. in time is conv. in frequency):

\[ V(e^{j\omega}) = S(e^{j\omega}) * W(e^{j\omega}) = \int_{-\pi}^{\pi} S(e^{j\omega})W(e^{j(\omega-\theta)})d\theta \]  \hspace{1cm} (5.7)

if \( s(n) = \alpha \sin(\omega n) \) (i.e., a very narrowband signal), then \( w(n) \) is a square wave (recall, odd-harmonics only).
Full-wave analysis

Sine wave
Square wave
Rectified sine
Spectrum of sine
Spectrum of square
(rectifier signal)
Convolution in frequency

Full-wave analysis - real signal

Original Signal

After BPF

Rectifier Signal

After Rectification

After Low-Pass Filter

After downsampler, use full bandwidth at sr
Filter Spacing

- Filter spacing could be uniform, constant Q, or perceptually inspired.

\[ f_i = \frac{F_s}{N} i \]  

(5.8)

where \( f_i \) is center frequency of \( i \)th filter, \( F_s \) is sampling frequency, \( i = 1, \ldots, Q = N/2 \), so there are \( Q \) filters.

- Difference between ideal and more real looking filter bank.

Alternative: constant Q like

- Bandwidths \( b_i \) and center frequencies \( f_i \) of filters

\[ b_1 = C \]  

(5.9)

\[ b_i = \alpha b_{i-1}, 2 \leq i \leq Q \]  

(5.10)

\[ f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{b_i - b_{i-1}}{2} \]  

(5.11)

where \( C \) is some starting bandwidth.

- Common values are \( \alpha = 2 \) (octave spaced filters) or \( \alpha = 4/3 \) (1/3 octave spaced) filters.

- Can also use critical-band center frequencies (as measured on humans) as \( b_i \).
**Filter bank spacing examples**

\[ \alpha = 2 \text{ so octave spaced } (f_1 = 300, f_2 = 600, f_3 = 1200, \ldots), \]
non-overlapping, 200-3200Hz (requiring \( F_s = 6.67 \text{kHz} \)), \( C = 200 \). This is constant \( Q = \frac{c}{b} \).

\[ \alpha = \frac{4}{3} \text{ so } \frac{1}{3}\text{-octave, 12-band, 200-3200Hz, } C=50, \text{ so} \]

**Alternative view of filtering**

- A band-pass filter can be seen as a frequency modulated standalone low-frequency filter, constructed as follows:

\[ h_i(n) = w(n)e^{j\omega_in} \quad (5.12) \]

where \( w(n) \) is the underlying low-pass filter of length \( L \), evenly centered at zero.

- \( w(n) \) is modulated by the complex sinusoid to become the BPF.
Alternative view of filtering

- Consider BPF \( h_i(n) \), source signal \( s(n) \), and band-limited (filtered) signal \( x_i(n) \).
- Then we have the relationship:

\[
x_i(n) = s(n) * h_i(n) \tag{5.13}
\]

\[
= \sum_{m=0}^{L-1} h_i(m)s(n - m) \tag{5.14}
\]

\[
= \sum_{m} w(m)e^{j\omega_i m} s(n - m) \tag{5.15}
\]

\[
= \sum_{m} s(m)w(n - m)e^{j\omega_i(n-m)} \tag{5.16}
\]

\[
= e^{j\omega_in} \sum_{m} s(m)w(n - m)e^{-j\omega_im} \tag{5.17}
\]

\[
= e^{j\omega_in}\text{DTFT}[s(m)w(n - m)] \tag{5.18}
\]

- \( w(n) \) is a window that looks like:

\[
\begin{array}{c}
\text{w(n)} \\
\hline \\
n
\end{array}
\]

- So, the band-limited signal \( x_i(n) \) is a scaled and modulated DTFT of the speech signal at a window with center at time location \( n \).

\[
x_i(n) = e^{j\omega_in}\text{DTFT}[s(m)w(n - m)] \tag{5.19}
\]

- The time window function \( w(n) \) is the underlying low-pass filter of length \( L \), evenly centered at zero.
Alternative view of filtering

So we have

$$x_i(n) = e^{j\omega_i n}S_n(e^{j\omega_i})$$

(5.20)

where $e^{j\omega_i n}$ is what modulates to center frequency $\omega_i$ and $S_n(e^{j\omega_i})$ is the slower-frequency amplitude/phase at time $n$ for spectral region $\omega_i$.

Hence, we can use FFTs to implement filterbank, and have fast filterbank implementations. This is now ubiquitous!

FFT gives us outputs for linear spaced frequency bins.

$L$, length of $w(n)$ window determines spectral/temporal resolution tradeoff.

$L$ long $\Rightarrow$ good spectral resolution (can get good pitch harmonics, bad temporal resolution, can’t see overall spectral envelope or desired vocal tract (VT) response)

$L$ short $\Rightarrow$ good temporal resolution, bad spectral resolution, good overall spectral envelope shape (VT response), but can’t see pitch harmonics

Simple Matlab Demo

Given signal $s$, we can adjust the window size $L$ and look at the effect of the short-time Fourier transform.

```matlab
Fs = 1600;
y = audiorecorder(Fs,16,1); % creates the record object.
recordblocking(y,4); % records 4 seconds of speech
s = getaudiodata(y);
L=800; % adjust this from small (say 25) to large (say 1024).
[S,F,T,P] = spectrogram(s,L,round(L - L/8),1024,Fs);
surf(T,F,10*log10(P),'edgecolor','none');
axis tight; view(0,90);
xlabel('Time (seconds)');
ylabel('Hz');
```
**Mel/Bark scale warping**

- Mel-scale (normal frequency)
  \[ \text{Mel}(f) = 1127 \log(1 + \frac{f}{700}) \]  
  \[ (5.21) \]

- Mel/Bark scale generalization:
  \[ \hat{\omega} = \tan^{-1} \frac{(1 - \alpha^2) \sin \omega}{(1 + \alpha^2) \cos \omega - 2\alpha} \]  
  \[ (5.22) \]

- \( \approx \) Mel with \( \alpha = 0.42 \), and \( \approx \) Bark with \( \alpha = 0.55 \) at 16kHz

Ultimate goal: better approximate peripheral auditory system and cochlear filter bank. What it thinks is important is probably also what is important for speech recognition.

---

**Constant-Q or perceptual output from FFT**

- Simple: just bin outputs of FFT accordingly for const-Q or critical band frequency.
- Example Mel-scale spectral filters (sum the weighted FFT outputs accordingly).
### Mel Filter Bank

- We get the Mel spectrum as follows:

\[
\tilde{S}(\ell) = \sum_{k=1}^{N/2} S(k) M_\ell(k) \tag{5.23}
\]

where \( \tilde{S}(\ell) \) is the Mel spectrum, \( S(k) \) is the original spectrum at frequency \( k \) (where \( k \) corresponds to \( k f_s/N \) Hz), and \( M_\ell(k) \) is a window function.

- The total number of triangular Mel weighting filters might be around 20 in practice.

### Windowed speech

- This gives us filter-bank processing in a window of speech, so window slides over speech and we get VT estimate for each window (in form of feature vector)

![Waveform with windowed speech](image)

- Typical: 25ms window widths, 15ms overlap (100Hz sample rate)
Recall from lecture 3

- Next two slides are from lecture 3.

Human Vocal-Tract Model

- But human vocal tract is not uniform tube.

- rather than go through analysis for continuous curved tube, it is much easier to use our analysis for piece-wise constant tube model. $A_i$’s are cross sectional areas for tube $i$, $\ell_i$ are lengths (but assume equal), number ($N$) is typically 8-12, but more is better (more accurate).
It is well known that the transfer function of this kind of system is:

\[ H(z) = \frac{U_L(z)}{U_G(z)} = \frac{1 + r_G z^{-N/2}}{2} \prod_{i=1}^{N}(1 + r_i) \propto \frac{1}{1 - \sum_{i=1}^{N} a_i z^{-1}} \]  

(5.42)

(the system is just a factored form).

But this is just a simple all-pole (linear prediction) model. I.e.,

\[ y[n] = c + \sum_{i=1}^{N} \alpha_i y[n-i] + x[n] \]  

(5.43)

A case for linear-predictive coding of speech (get the \( \alpha \) coefficients and use/send them).

Question: how do we go directly from speech signal to the optimal parameters of all-pole model?

This is a parameter estimation problem, so lets review this for a bit.
Basics statistical parameter estimation

- Training data $D = \{(x_n, y_n)\}_{n=1}^N$ where $(x_n, y_n) \sim p(x, y)$ are drawn from a distribution.
- $x$ is $p$-dimensional column vector, $y$ is scalar.
- Goal: find $f : D_X \rightarrow D_Y$ with minimum error, where

$$\text{Error}_n = e_n = f(x_n) - y_n, \quad e = f(x) - y \quad (5.24)$$

and

$$E[e^2] = \int p(x, y)(f(x) - y)^2 dxdy \quad (5.25)$$

$$= \int p(x) \int (f(x) - y)^2 p(y|x) dydx \quad (5.26)$$

- Taking derivatives and setting to zero, we get best solution:

$$f(x) = \int y p(y|x) dy = E[y|x] \quad (5.27)$$

Note: $f$ might be linear or non-linear, we don’t know. But we can still find best solution under a linear model, where $f(x) = Ax$, and $A$ is a $1 \times p$ vector of “regression” coefficients, in which case $\hat{y} = Ax$.

- In general, assume $f()$ is parameterized by some parameters $A$ so

$$E = \frac{1}{N} \sum_{i=1}^N (f_A(x_n) - y_n)^2 \quad (5.28)$$

- Taking derivative of $E$ w.r.t. $A$ and set to zero to get:

$$\frac{\partial E}{\partial A} = 2 \frac{\sum_{n=1}^N (f_A(x_n) - y_n) \partial f_A(x_n)}{A} = 0 \quad (5.29)$$

- Under the linear $f(x) = Ax$ assumption, $\frac{\partial f_A(x_n)}{\partial A} = x^\top$
Basics statistical parameter estimation

- This gives us

\[
\frac{\partial E}{\partial A} = 2 \frac{A}{A} \sum_{n=1}^{N} (A^T x_n - y_n)x_n^{T} = 0 \quad (5.30)
\]

- We simplify this a bit by defining matrices associated with these quantities. First define the “design matrix” \( X \) and column vector \( Y \)

\[
X = \begin{pmatrix}
x_1^T \\
x_2^T \\
\vdots \\
x_N^T
\end{pmatrix}, \quad \text{and} \quad Y = \begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_N
\end{pmatrix} \quad (5.31)
\]

- With this, we get the “normal equations”

\[
X^T (XA - Y) = 0 \quad (5.32)
\]

Normal Equations

- The “normal equations”

\[
X^T (XA - Y) = X^T E = 0 \quad (5.33)
\]

i.e., \( Y \) lies in column space of matrix \( X \) (linear combinations of columns of \( X \)), when \( Y \) is being approximated by \( XA \).

- Called “normal equations” because \( X \) is orthogonal to the error \( E = (XA - Y) \):

Solution is of the form:

\[
A = (X^T X)^{-1}X^T Y \quad (5.34)
\]

where \((X^T X)^{-1}X^T\) is known as the Moore-Penrose pseudo-inverse of matrix \( X \).
We can apply this general setting to all-pole parameter estimation, but in this case, because of the special nature of the problem, there are some significant computational simplifications that can be made (over just computing Moore-Penrose pseudo-inverse naively).

We have the following model

\[ u[n] \rightarrow \text{Time-varying Vocal tract system function} \rightarrow s[n] \]

\( u \) is the glottal excitation, the speech signal is modeled as a complex waveform, and the time-varying vocal tract system function (as an all-pole model) is assumed to be essentially time-invariant over short enough time-windows.

and assume that over short enough time-window, this is essentially time-invariant (i.e., assume piece-wise TI).

LPC model, order \( p \)

- **Time domain**
  \[
  s[n] = \sum_{i=1}^{p} a_i s[n - i] + Gu[n]
  \]
  (5.35)

  where \( G \) is constant.

- **\( z \)-Domain**
  \[
  S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + G U(z)
  \]
  (5.36)

- **Transfer function \( z \) transform:**
  \[
  H(z) = \frac{S(z)}{G U(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}
  \]
  (5.37)
All-pole model

- All-pole models: $s[n]$ is predictable from weight sum of its past samples & the glottal pulse at that time.
- We assume approximation $\tilde{s}[n]$, unknown values $\{a_i\}$, with:

$$\tilde{s}[n] = \sum_{i=1}^{p} a_i s[n-i]$$

(5.38)

- We have the samples $s[n]$. What is not accounted for in this estimate is the glottal pulse, which we assume to be the error, i.e.,

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{i=1}^{p} a_i s[n-i]$$

(5.39)

- Hence, $S$ to error transfer function is inverse of transfer function $H$ before.

$$\frac{E(z)}{S(z)} = 1 - \sum_{i=1}^{p} a_i z^{-i}$$

(5.40)

Goal: find the $a_i$ coefficients that make $e[n]$ reflect the glottal pulse. In general, this is hard due to periodic nature of $e[n]$.

Instead, let's assume white Gaussian zero-mean unit variance noise source $u[n]$. Then estimation problem is then just error minimization (this is maximum likelihood estimate).
All-pole parameter estimation

- There are two methods:
  1. least-squares auto-correlation method, where we window the time signal
  2. least-squares auto-covariance method, where we window the error function itself.

- We do Auto-correlation method first.
- Define error cost function:

\[
E_n = \sum_m e_n^2(m) = \sum_m (s_n(m) - \tilde{s}_n(m))^2 \tag{5.41}
\]

\[
= \sum_m \left[ s_n(m) - \sum_{i=1}^{p} a_i s_n(m - i) \right]^2 \tag{5.42}
\]

where we window the speech signal \( s[m] \) with a window function \( w[m] \) (such as Hamming window) but first shift the speech signal to the left by \( n \) as in:

\[
s_n[m] = s[n + m]w[m] \tag{5.43}
\]

- Take derivatives to get vector \( a_1, a_2, \ldots, a_p \) for \( n^{th} \) time window. That is for all \( i \)

\[
\frac{\partial E_n}{\partial a_i} = 0 \tag{5.44}
\]

- For now, drop \( n \)-subscripts and show it done for one window, since each window is processed independently of the others.

- Other assumptions we make: ergodicity: window length is long enough to get good statistical estimate but short enough so that time properties can estimate ensemble properties. In other words:

\[
E_n = \frac{1}{K} \sum_m e_n^2(m) \approx E[e_n^2(m)] \tag{5.45}
\]

where \( K \) is the window length.
All-pole parameter estimation

- To estimate, we go through the process:
  \[
  \frac{\partial E[e_n^2(m)]}{\partial a_i} = 2E \left[ e_n(m) \frac{\partial e(m)}{\partial a_i} \right], \quad \text{and} \quad \frac{\partial e(m)}{\partial a_i} = -s_n(m - i) \quad (5.46)
  \]

- Define autocorrelation of a speech signal (random interpretation) as:
  \[ R_n(k, \ell) = E[s_n[k]s_n[\ell]] \quad (5.47) \]

- Then setting above derivative to zero, we get:
  \[
  \frac{\partial E[e_n^2(m)]}{\partial a_i} = 2E \left[ \sum_{k=1}^{p} a_k s_n(m - k)s_n(m - i) - s_n(m)s_n(m - i) \right] \\
  = \sum_{k=1}^{p} a_k R_n(m - k, m - i) - R_n(m, m - i) \quad (5.48) \\
  = 0 \quad (5.49)
  \]

- In other words, we get:
  \[
  \sum_{k=1}^{p} a_k R_n(m - k, m - i) = R_n(m, m - i) \quad (5.50)
  \]

- This is just the normal equations \((X^\top(XA - Y) = 0)\) again, but this time for the auto-correlation function.

- Features of the solution: Quadratic (so convex), so only one unique solution (unimodal)
- Also, there exist fast methods to get solution (better than matrix inverse) since the “matrices” are special.
All-pole parameter estimation

- In auto-correlation method, \( s(n) \) is windowed signal
  \[ s[m] = w[m]s_n[m] \]
  and we assume stationary ergodic, so that:
  \[ R(k, \ell) = R(k - \ell) = R(\ell - k) \quad (5.51) \]

- Hence, we get
  \[ \sum_{k=1}^{p} a_k R_n(i - k) = R_n(i), \quad \text{for } i = 1, \ldots, p \quad (5.52) \]

- Define \( \vec{a} = [a_1, a_2, \ldots, a_p]^\top \) and \( \vec{R} = [R(1), R(2), \ldots, R(p)]^\top \), and
  \[
  R = \begin{pmatrix}
  R(0) & R(1) & R(2) & R(3) & \cdots & R(p-1) \\
  R(1) & R(0) & R(1) & R(2) & \cdots & R(p-2) \\
  R(2) & R(1) & R(0) & R(1) & \cdots & R(p-3) \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  R(p-1) & R(p-2) & R(p-3) & R(p-4) & \cdots & R(0) 
  \end{pmatrix}
  \]

- \( R \) is symmetric Toeplitz \( p \times p \) matrix (same elements on diagonal),
  positive semi-definite (but might be singular).

- Then update equations become \( R\vec{a} = \vec{R} \) or \( \vec{a} = R^{-1}\vec{R} \).
- More on this in a bit
Window Effect

- \( s_n[m] \) is windowed signal, \( s_n[m] = w(m)s[m] \), window at \( n \) length \( N \).
- First note, \( w[m] = 0 \) outside of \( 0 \leq m \leq N - 1 \), so same for \( s[m] \).
- Hence \( e[m] = s_n[m] - \tilde{s}_n[m] = s_n[m] - \sum_{i=1}^{p} a_is_n[m - i] = 0 \) outside of \( 0 \leq m \leq N + p - 1 \), and overall error is

\[
E_n = \sum_{m=0}^{N+p-1} e_n^2(m) \tag{5.53}
\]

- To estimate \( R \), we do:

\[
R_n(-k,-i) = E[s_n[-k]s_n[-i]] \approx \sum_m s_n(m - k)s_n(m - i) \tag{5.54}
\]

- Since \( s[m] = 0 \) for \( m < 0 \) and \( m > N - 1 \), above is non-zero only for 

\( 0 \leq m - k \leq N - 1 \), or equivalently \( k \leq m \leq N + k - 1 \) (5.55)

and

\( 0 \leq m - i \leq N - 1 \), or equivalently \( i \leq m \leq N + i - 1 \) (5.56)

All-pole parameter estimation

- So we get, for \( 1 \leq \ell, k \leq p \), and setting \( \ell = m - i \):

\[
R_n(-k,-i) \approx \sum_{m=0}^{N+p-1} s_n(m - k)s_n(m - i) \tag{5.57}
\]

\[
= \sum_{\ell=-i}^{N+p-1-i} s_n(\ell + i - k)s_n(\ell) \tag{5.58}
\]

\[
= \sum_{\ell=0}^{N+1-(i-k)} s_n(\ell + i - k)s_n(\ell) \tag{5.59}
\]

which follows since zero for \( \ell < 0 \) and \( \ell > N - 1 - (i-k) \).

- result is a function only of \( i - k \), or \( R(i - k) \), so when we truncate the speech signal using the window method, the auto-correlation method for stationary signals is indeed appropriate.
All-pole parameter estimation

- Next, we outline Auto-covariance method
- This method for non-stationary signals where \( R(\ell, k) \neq R(i - k) \).
- We have windowed error function: \( E_n = w(n) \sum m e^2(n) \) rather than windowed speech function. We get:

\[
\sum_{k=1}^{p} a_k R(n - k, n - i) = R(n, n - i) \tag{5.60}
\]

- We define

\[
\phi(k, i) = R(n - k, n - i) \tag{5.61}
\]

and

\[
\begin{pmatrix}
\phi(1, 1) & \phi(2, 1) & \phi(3, 1) & \ldots & \phi(p, 1) \\
\phi(1, 2) & \phi(1, 3) & \ldots & \ldots & \phi(1, p) \\
\phi(1, 3) & \phi(1, 4) & \ldots & \ldots & \phi(1, p) \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\phi(1, p) & \phi(2, p) & \phi(3, p) & \ldots & \phi(p, p)
\end{pmatrix}
\begin{pmatrix}
a_1 \\
a_2 \\
a_3 \\
\vdots \\
a_p
\end{pmatrix}
= 
\begin{pmatrix}
\phi(0, 1) \\
\phi(0, 2) \\
\phi(0, 3) \\
\vdots \\
\phi(0, p)
\end{pmatrix}
\tag{5.62}
\]

Summary

- Complicated tube analysis gives us LPC.
- We can estimate LPC coefficients relatively easily.