Review: Filtering

We can describe a system (physical, mathematical, or computational) by the way it transforms an input signal into an output signal. We can represent a system in multiple ways:

...by its unit sample response:

\[ x[n] \rightarrow h[n] \rightarrow y[n] = (x \ast h)[n] \]

...or by its frequency response:

\[ X(\Omega) \rightarrow H(\Omega) \rightarrow Y(\Omega) = X(\Omega)H(\Omega) \]

Review: STFT

STFT is a compromise between time- and frequency-domain representations, representing the frequency content of the signal at various points in time.
Review: Spectrograms

The STFT enhances our ability to reason about the frequency content of signals at various points in time. It is often visualized using a spectrogram, which is defined to be the magnitude squared of the STFT.

Today: Speech

Our model: speech is generated by the passage of air from the lungs, through the vocal cords, mouth, and nasal cavity.

Model of Speech Production

Controlled by complicated muscles, vocal cords are set in vibration by the passage of air from the lungs.

During voiced speech, glottis generates puffs of air about 4 ms in duration. Frequency of puffs ranges from 120–240 Hz.
Model of Speech Production

Vibrations of the vocal cords are “filtered” by the mouth and nasal cavities to generate speech.

Vowels sound different because mouth and lip positions are different.

Example: x-ray movie showing speech in production
Ken Stevens, 1962

Source/Filter Model

Acoustic Sources: pulse train for voiced utterances, Gaussian noise for unvoiced utterances

Gain: $G$ controls loudness

Vocal Tract: filter represents shape of mouth, tongue, and lips
Model of Speech Production

Vowels sound different because mouth and lip positions are different.

![Waveforms of different vowels](image1)

Characterizing Vowel Sounds

Same glottis signal + different formants → different vowels.

![Formant analysis](image2)

Formants

Resonant frequencies of the vocal tract.

<table>
<thead>
<tr>
<th>Formant</th>
<th>heed</th>
<th>head</th>
<th>had</th>
<th>hod</th>
<th>haw'd</th>
<th>who'd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>F1</td>
<td>270</td>
<td>530</td>
<td>660</td>
<td>730</td>
<td>570</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>2290</td>
<td>1840</td>
<td>1720</td>
<td>1090</td>
<td>840</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>3010</td>
<td>2480</td>
<td>2410</td>
<td>2440</td>
<td>2410</td>
</tr>
<tr>
<td>Women</td>
<td>F1</td>
<td>310</td>
<td>610</td>
<td>860</td>
<td>850</td>
<td>590</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>2790</td>
<td>2330</td>
<td>2050</td>
<td>1220</td>
<td>920</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>3310</td>
<td>2990</td>
<td>2850</td>
<td>2810</td>
<td>2710</td>
</tr>
<tr>
<td>Children</td>
<td>F1</td>
<td>370</td>
<td>690</td>
<td>1010</td>
<td>1030</td>
<td>680</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>3200</td>
<td>2610</td>
<td>2320</td>
<td>1370</td>
<td>1060</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>3730</td>
<td>3570</td>
<td>3320</td>
<td>3170</td>
<td>3180</td>
</tr>
</tbody>
</table>

[http://www.sfu.ca/sonic-studio/handbook/Formant.html](http://www.sfu.ca/sonic-studio/handbook/Formant.html)
**Characterizing Vowel Sounds**

Characteristic peaks in frequency response are called formants.

![Formant Diagram]

Formants can be used to characterize vowels. For example, in an “ee” sound, F1 and F2 are around 270Hz and 2290Hz; in an “oo” sound, they are around 300Hz and 870Hz.

We detect changes in the filter function to recognize vowels. Example: Scales

---

**Source/Filter Model**

![Source/Filter Model Diagram]

**Acoustic Sources**: pulse train for voiced utterances, Gaussian noise for unvoiced utterances

**Gain**: G controls loudness

**Vocal Tract**: filter represents shape of mouth, tongue, and lips

How can we characterize this filter?

---

**Linear Predictive Coding**

The speech signal $s[n]$ is shaped by the vocal tract. We will develop a “predictive model”

$$s[n] = \sum_{k=1}^{P} a_k s[n-k]$$

where output at time $n$ is a linear combination of $P$ previous inputs.

Our goal: estimate the $a_k$ values.
Linear Predictive Coding

Let $\hat{s}[n]$ be our model’s prediction for $s[n]$:

$$\hat{s}[n] = \sum_{k=1}^{P} a_k s[n-k]$$

We want to find the $a_k$ values that minimize the squared difference between the two:

$$E = \sum_{n} (s[n] - \hat{s}[n])^2 = \sum_{n} (s[n] - \sum_{k=1}^{P} a_k s[n-k])^2$$

Set derivative w.r.t $a_i$ equal to zero for $1 \leq i \leq P$:

$$\frac{\partial E}{\partial a_i} = 0 = \sum_{n} 2(s[n] - \sum_{k=1}^{P} a_k s[n-k])(-s[n-i])$$

And rearrange:

$$\sum_{n} s[n]s[n-i] = \sum_{k=1}^{P} a_k \sum_{n} s[n-k]s[n-i]$$

We can rewrite this expression in terms of the autocorrelation function $R[i] = R[-i] = \sum_{n} s[n]s[n-i]$.

Our final result is:

$$R[i] = \sum_{k=1}^{P} a_k R[i-k] \text{ for } 1 \leq i \leq P$$

which can be expressed more concisely as a matrix equation:


Summary of LPC procedure:

- select a region of time using a window function $w[n]$
- calculate the autocorrelation function $R[i]$
- solve the set of linear equations to find $a_k$.

Now, our filter is represented by:

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

How would we find the frequency response $H(\Omega)$?
Examples of Applying This Model

- Vowel recognition
- Speech synthesis