6.003
Signal Processing

Week 8, Lecture B
Interlude: Speech

Adam Hartz
hz@mit.edu
Describe a **system** (physical, mathematical, or computational) by the way it transforms an **input signal** into an **output signal**. Can be represented multiple ways:

...by its **unit-sample/impulse response**.

\[ x[n] \xrightarrow{} h[n] \xrightarrow{} y[n] = (x \ast h)[n] \]

...by its **frequency response**.

\[ X(\Omega) \xrightarrow{} H(\Omega) \xrightarrow{} Y(\Omega) = X(\Omega)H(\Omega) \]

...as a **filter**.

\[ \sum_k X[k]e^{j\Omega_0 n} \xrightarrow{} H(\Omega) \xrightarrow{} \sum_k X[k]H(k\Omega_0)e^{j\Omega_0 n} \]
Review: STFT

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

![Diagram showing time and frequency windows]

- Window 0
- Window 1
- Window 2
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.
Model of Speech Production

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

Vowels sound different because mouth and lip positions are different.

bat
bit
but

bait
bite
boot

bet
bought
boat

beet
Speech Production

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

Same glottis signal + different formants $\rightarrow$ different vowels.
Characterizing Vowel Sounds

Characteristic peaks in frequency response are called formants.

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.
Detecting Vowel Sounds

We detect changes in the filter function to recognize vowels.

Example: scales
Detecting Vowel Sounds

We detect changes in the filter function to recognize vowels.

\[ |H^{ee}(\omega)| \]

\[ \omega \]

\[ \text{low} \]

\[ \omega \]

\[ \text{intermediate} \]

\[ \omega \]

\[ \text{high} \]

\[ 2\pi 400 \, 2\pi 2500 \, 2\pi 3200 \, \omega \]
Detecting Vowel Sounds

We detect changes in the filter function to recognize vowels.

\[ |H^{ee}(\omega)| \]

- **low**
- **intermediate**
- **high**
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
Source/Filter Model

\[ s[n] \]

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 to characterize this filter?
The speech signal $s[\cdot]$ 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:

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

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

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

And rearrange:

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

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:

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

which can be expressed more concisely as a matrix equation:
Linear Predictive Coding

Summary of LPC procedure:
1. select a region of time using a window function \( w[n] \)
2. calculate the autocorrelation function \( R[i] \)
3. 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 \( H(\Omega) \)?
Check Yourself!

How would we find the frequency response $H(\Omega)$?

**Method 1:**
Use the difference equation to calculate the unit-sample response $h[n]$.

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

Set $x[n] = \delta[n]$ and solve for $h[n] = y[n]$ for $n \geq 0$.

Calculate the Fourier transform $H(\Omega)$ from $h[n]$.
Check Yourself!

How would we find the frequency response $H(\Omega)$?

**Method 2:**
Take the Fourier transform of the difference equation:

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

$$Y(\Omega) = X(\Omega) + \sum_{k=1}^{P} a_k e^{-j\Omega k} Y(\Omega)$$

Since $Y(\Omega) = X(\Omega)H(\Omega)$, we can compute $H(\Omega)$ as

$$H(\Omega) = \frac{Y(\Omega)}{X(\Omega)} = \frac{1}{1 - \sum_{k=1}^{P} a_k e^{-j\Omega k}}$$
Examples

Examples:
- vowel recognition
- speech generation (pitch shift)