Acoustic theory of speech production
Acoustics of vowels

Oct. 6, 2008
Standing waves

- **What is a standing wave:**
  - A stationary vibration pattern that results from the combination of reflection and interference such that the reflected waves interfere constructively with the incident waves.
  - It has nodes – points where the medium doesn’t move, and antinodes – points where the motion is a maximum.
  - The medium appears to vibrate in segments or regions and the fact that these vibrations are made up of traveling waves is not apparent – hence the term “standing wave”.

[Animation from: Daniel Russell website]
Standing waves

- Standing waves on a string that is fixed at both ends:

[illustrations from: C.R. Nave HyperPhysics (GSU) website]
Standing waves

- Standing waves on a string that is fixed at both ends:

\[ \lambda_n = \frac{2L}{n} \quad f_n = \frac{v}{\lambda_n} = n \frac{v}{2L} \quad n = 1, 2, 3... \]

<table>
<thead>
<tr>
<th>( L = \lambda/2 )</th>
<th>( \lambda = 2L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f = v/(2L) )</td>
<td>first harmonic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( L = \lambda )</th>
<th>( \lambda = L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f = v/(L) )</td>
<td>second harmonic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( L = 3\lambda/2 )</th>
<th>( \lambda = 2L/3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f = 3v/(2L) )</td>
<td>third harmonic</td>
</tr>
</tbody>
</table>
Standing waves

- Standing waves in an air tube that is closed at one end and open at the other:
  - A node for air displacement is always an antinode for air pressure, and vice versa.

[Diagram showing standing wave phenomena]
Standing waves

- Standing waves in an air tube that is closed at one end and open at the other:

\[ \lambda_n = \frac{4L}{n} \quad f_n = \frac{v}{\lambda_n} = n \frac{v}{4L} \quad n = 1, 3, 5... \]

<table>
<thead>
<tr>
<th>(L = \lambda/4)</th>
<th>(\lambda = 4L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L = 3\lambda/4)</td>
<td>(\lambda = 4L/3)</td>
</tr>
<tr>
<td>(L = 5\lambda/4)</td>
<td>(\lambda = 4L/5)</td>
</tr>
</tbody>
</table>
Standing waves

- Demo:

[Applet from: Fredrick Olness website]
Standing waves

- Standing waves as a property of a system: filter, resonator.

![Standing waves diagram](Picture from: Haskins Labs website)
The source filter model of speech production

- The source filter model:
  - The production of speech consists of two kinds of operations: (1) the generation of sound sources, at the glottis or at some point along the length of the vocal tract, and (2) the filtering of these sources by the vocal tract.
  - There are two principle kinds of sources: a) turbulence noise (present in [s], [f], etc.), and b) vocal-fold vibration (present in vowels, nasals etc.).
The source filter model of speech production

- The glottal source (impulses):
  - The vibration of the vocal folds are periodic, which determines the fundamental frequency.
  - The spectrum of the glottal source decreases in amplitude with increasing frequency at a rate of around −12dB per octave — that is for each doubling in frequency, the amplitude of the spectrum decreases by around 12dB (later, radiation from the lips will raise the tilt by 6 dB/octave).
  - Open Quotient (OQ): the ratio of the time in which the vocal folds are open and the whole pitch period duration.

[Movie from: UCL phonetics website; figures from: Steve Cassidy website]

LING 120 Introduction to Phonetics I, Fall 2008
Open quotients and phonation types

[From: Johnson 2003]
The source filter model of speech production

- The vocal tract filter:

  ![Diagram showing vocal tract positions and formants for different vowel sounds](From: UCL phonetics website)

- Formants are the resonance frequencies of the vocal tract.
Deriving schwa

- When the tongue is in the position of the neutral vowel, the schwa, the vocal tract (without the nasal cavity) can be approximated as a tube that is closed at one end (glottis) and open at the other (mouth).

\[
\lambda_n = \frac{4L}{n} \quad \text{for} \quad n = 1, 3, 5, \ldots
\]

If \( L = 17.5 \text{ cm}; \ v = 35,000 \text{ cm/s} \):
- \( F_1 = \frac{v}{\lambda_1} = \frac{v}{(4L)} = 500\text{Hz} \)
- \( F_2 = \frac{v}{\lambda_3} = \frac{v}{(4/3L)} = 1500\text{Hz} \)
- \( F_3 = \frac{v}{\lambda_5} = \frac{v}{(4/5L)} = 2500\text{Hz} \)

[Figure from: Ladefoged 1996]
Vowel formants

- Vowels are associated with a steady-state articulatory configuration and a steady-state acoustic pattern. Vowels often have been characterized with a very simple set of acoustic descriptors, namely, the frequencies of the first three formants.
- American English: *heed, hid, head, had, hod, hawed, hood, who’d*

[From: Ladefoged, 2005]
Measuring formants

- Spectrogram + formant tracking over a time span

- LPC spectrum at one time point
Spectrograms

- Wideband and narrowband spectrograms
  - Traditionally, speech spectrograms are either *wideband* or *narrowband*, so called from the size of the electronic bandpass filter (330Hz or 45Hz respectively) that swept the frequency range on the original speech spectrographs.
  - The practical significance is that wideband spectrograms show formant structure while narrowband spectrograms reveal the harmonic structure.
Spectrograms

- Trade off between time and frequency resolution
  - The general rule is that the wider the passband the coarser the analysis in frequency but the finer the analysis in time, whereas the narrower the passband the finer the analysis in frequency but the coarser the analysis in time.

**DFT (to compute spectrograms):**

\[
X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}n^k}, \quad k = 0,1,..., N-1
\]

**FFT:** a computationally efficient algorithm for implementing DFT. It must use power of two points (zero padding).

In DFT and FFT, time resolution and frequency resolution is a trade-off, due to the connection between window size \((N)\) and frequency resolution \((2\pi / N)\).
Spectral analysis in Praat

- **View Range**: the range of frequencies to display.
- **Window length**: the duration of the analysis window. The window length determines the *bandwidth* of the spectral analysis. To get a “wide-band” spectrogram (bandwidth 260 Hz), keep the standard window length of 5 ms; to get a “narrow-band” spectrogram (bandwidth 43 Hz), set it to 30 ms (0.03 seconds).
- **Dynamic range (dB)**: All values that are more than Dynamic range dB below the maximum will be drawn in white. Values in-between have appropriate shades of grey.
Formant tracking and LPC

• Formant tracking from spectrograms (using eyes, image processing, etc.)
• The most popular formant tracking algorithms are, however, based on LPC.
• Linear Predictive Coding (LPC): Each sample of the speech signal is modeled as the weighted sum of past samples. The weights (or coefficients) are used to derive a linear prediction filter. The LP filter represents the vocal tract, the frequency response of the filter (peaks in the spectrum of the LP filter) provides estimates of the formants of the speech signal.

\[
\hat{y}[n] = \sum_{k=1}^{p} a_k y[n - k].
\]

\[
y[n] = b_1 y[n-1] + b_2 y[n-2] + \cdots
\]
Linear Predictive Coding (LPC)

• Predict $y[n]$ from a linear combination of its past values.

$$
\hat{y}[n] = \sum_{k=1}^{p} a_k y[n - k].
$$

• Find the predictor coefficients by minimizing the mean squared error $E$:

$$
\epsilon[n] \triangleq y[n] - \hat{y}[n] = y[n] - \sum_{k=1}^{p} a_k y[n - k].
$$

$$
E = \frac{1}{N} \sum_{n=0}^{N-1} \epsilon[n]^2 = \frac{1}{N} \sum_{n=0}^{N-1} \left( y[n] - \sum_{k=1}^{p} a_k y[n - k] \right)^2.
$$

• The linear predictor is an all-pole IIR filter. The all-pole IIR filter represents the vocal tract, therefore, the frequency response of the filter shows the resonances that shape the formants of the speech wave.

• In order for LPC to work, the vocal tract (the filter) must not have zeros. In mathematical terms, side branches introduces zeros. So LPC doesn’t work well for nasals.
LPC spectrum and FFT spectrum

- LPC spectrum: Linear prediction estimates the vocal tract filter and shows the resonances that shape the formants of the speech wave. There is no distinction between wideband and narrowband analysis in LPC. The LPC slice shows the envelope of the vocal tract resonances and looks superficially like a wideband FFT.
- FFT spectrum: Wideband FFT slices show the formant structure at the selected location, narrowband FFT slices show more detailed spectral structure.
LPC analysis in Praat

- **Prediction order**: the number of filter coefficients; this is a critical setting; the absolute minimum setting is twice the number of formants in the frequency range of the signal (but a better setting is twice the number of formants plus at least 2);

- **Analysis width**: the amount of data from the signal that is needed for the computation; a good setting includes at least two glottal periods;

- **Time step**: the time interval between computed analysis frames;

- **Pre-emphasis from**: the effect is to high pass filter the signal by +6dB/octave, enhancing energy at higher frequencies.
Acoustic vowel space

[From: *Acoustics of American English Speech*, 1993]

- First formant: vowel height; Second formant: backness
Acoustic vowel space

[From: Peterson and Barney 1952]
Acoustic vowel space

[From: www.helsinki.fi/hum/hyfl/projektit/vokaalikartat_eng.html]

LING 120 Introduction to Phonetics I, Fall 2008
Diphthongs

[From: *Acoustics of American English Speech*, 1993]

LING 120 Introduction to Phonetics I, Fall 2008