Acoustics of Speech

Julia Hirschberg

CS 4706
Goal 1: Distinguishing One Phoneme from Another, Automatically

- ASR: Did the caller say ‘I want to fly to Newark’ or ‘I want to fly to New York’?
- Forensic Linguistics: Did the accused say ‘Kill him’ or ‘Bill him’?
- What evidence is there in the speech signal?  
  - How accurately and reliably can we extract it?
Goal 2: Determining *How* things are said is sometimes critical to understanding

- Intonation
  - Forensic Linguistics: ‘Kill him!’ or ‘Kill him?’
  - TTS: ‘Are you leaving tomorrow./?’
  - What information do we need to extract from/generate in the speech signal?
  - What tools do we have to do this?
Today and Next Class

• How do we define cues to segments and intonation?
  – Fundamental frequency (pitch)
  – Amplitude/energy (loudness)
  – Spectral features
  – Timing (pauses, rate)
  – Voice Quality

• How do we extract them?
  – *Praat*
  – Wavesurfer
  – Xwaves…
Sound Production

- Pressure fluctuations in the air caused by a musical instrument, a car horn, a voice
  - Sound waves propagate thru e.g. air (marbles, stone-in-lake)
  - Cause eardrum (tympanum) to vibrate
  - Auditory system translates into neural impulses
  - Brain interprets as sound
  - Plot sounds as change in air pressure over time
- From a speech-centric point of view, sound not produced by the human voice is noise
  - Ratio of speech-generated sound to other simultaneous sound: Signal-to-Noise ratio
How ‘Loud’ are Common Sounds – How Much Pressure Generated?

<table>
<thead>
<tr>
<th>Event</th>
<th>Pressure (Pa)</th>
<th>Db</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Whisper</td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td>Quiet office</td>
<td>2K</td>
<td>40</td>
</tr>
<tr>
<td>Conversation</td>
<td>20K</td>
<td>60</td>
</tr>
<tr>
<td>Bus</td>
<td>200K</td>
<td>80</td>
</tr>
<tr>
<td>Subway</td>
<td>2M</td>
<td>100</td>
</tr>
<tr>
<td>Thunder</td>
<td>20M</td>
<td>120</td>
</tr>
<tr>
<td><em>DAMAGE</em></td>
<td>200M</td>
<td>140</td>
</tr>
</tbody>
</table>
Voiced Sounds are Typically Periodic

- Simple Periodic Waves (sine waves) defined by
  - **Frequency**: how often does pattern repeat per time unit
    - **Cycle**: one repetition
    - **Period**: duration of cycle
    - **Frequency**=# cycles per time unit, e.g. sec.
      - Frequency in Hz = cycles per second or 1/period
      - E.g. 400Hz pitch = 1/.0025 (1 cycle has a period of .0025; 400 cycles complete in 1 sec)
  - **Zero crossing**: where the waveform crosses the x-axis
– **Amplitude**: peak deviation of pressure from normal atmospheric pressure

– **Phase**: timing of waveform relative to a reference point
Figure 1.3  A 100 Hz sine wave with the duration of one cycle (the period) and the peak amplitude labeled.

Figure 1.4  Two sine waves with identical frequency and amplitude, but 90° out of phase.
Complex Periodic Waves

• Cyclic but composed of \textit{multiple} sine waves
• \textbf{Fundamental frequency (F0):} rate at which largest pattern repeats (also GCD of component frequencies) + \textit{harmonics}
• Any complex waveform can be analyzed into its component sine waves with their frequencies, amplitudes, and phases (\textit{Fourier’s theorem})
2 Sine Waves $\rightarrow$ 1 Complex periodic wave

Figure 1.5  A complex periodic wave composed of a 100 Hz sine wave and a 1,000 Hz sine wave. One cycle of the fundamental frequency ($f_0$) is labeled.
4 Sine Waves → 1 Complex periodic wave
Power Spectra and Spectrograms

• Frequency components of a complex waveform represented in the power spectrum
  – Plots frequency and amplitude of each component sine wave
• Adding temporal dimension → spectrogram
• Obtained via Fast Fourier Transform (FFT), Linear Predicative Coding (LPC),…
  – Useful for analysis, coding and synthesis
Examples and Terms

- **Vowels.wav, speechbeach1.wav, speechbeach2.wav**
- **Spectral slice**: plots amplitude at each frequency
- **Spectrograms**: plots changes in amplitude and frequency over time
- **Harmonics**: components of a complex waveform that are multiples of the fundamental frequency (F0)
- **Formants**: frequency bands that are most amplified by the vocal tract
Aperiodic Waveforms

- Waveforms with random or non-repeating patterns
  - Random aperiodic waveforms: white noise
    - Flat spectrum: equal amplitude for all frequency components
  - Transients: sudden bursts of pressure (clicks, pops, lip smacks, door slams)
    - Flat spectrum with single impulse
  - Voiceless consonants
Speech Waveforms in Particular

• Lungs plus vocal fold vibration filtered by the resonances of the vocal tract produce complex periodic waveforms
  – *Pitch range, mean, max*: cycles per sec of lowest frequency component of signal = fundamental frequency (F0)
  – *Loudness*:
    • RMS amplitude: \( \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2} \)
    • Intensity: in Db, where \( P_0 \) is auditory threshold pressure
      \[
      10 \log_{10} \frac{1}{NP_0} \sum_{i=1}^{N} x_i^2
      \]
How do we capture speech for analysis?

• Recording conditions
  – A quiet office, a sound booth, an anechoic chamber
• Microphones convert sounds into electrical current: oscillations of air pressure become oscillations of voltage in an electric circuit
  – Analog devices (e.g. tape recorders) store these as a continuous signal
  – Digital devices (e.g. computers, DAT) first convert continuous signals into discrete signals (digitizing)
Sampling

- **Sampling rate**: how often do we need to sample?
  - At least 2 samples per cycle to capture periodicity of a waveform component at a given frequency
  - 100 Hz waveform needs 200 samples per sec
  - **Nyquist frequency**: highest-frequency component captured with a given sampling rate (half the sampling rate) – e.g. 8K sampling rate (telephone speech) captures frequencies up to 4K
Sampling/storage tradeoff

• Human hearing: ~20K top frequency
  – Do we really need to store 40K samples per second of speech?
• Telephone speech: 300-4K Hz (8K sampling)
  – But some speech sounds (e.g. fricatives, stops) have energy above 4K…
  – Peter/teeter/Dieter
• 44k (CD quality audio) vs.16-22K (usually good enough to study pitch, amplitude, duration, …)
• Golden Ears…
Sampling Errors

• **Aliasing**:  
  – Signal’s frequency higher than the Nyquist frequency  
  – Solutions:  
    • Increase the sampling rate  
    • Filter out frequencies above half the sampling rate (anti-aliasing filter)
Quantization

- **Measuring the amplitude** at sampling points: what resolution to choose?
  - Integer representation
  - 8, 12 or 16 bits per sample
- Noise due to quantization steps avoided by higher resolution -- but requires more storage
  - How many different amplitude levels do we need to distinguish?
  - Choice depends on data and application (44K 16bit stereo requires ~10Mb storage)
– But *clipping* occurs when input volume (i.e. amplitude of signal) is greater than range that can be represented

– Watch for this when you are recording for TTS!

– Solutions

• Increase the resolution

• Decrease the amplitude

• Example: clipped.wav
Filtering

- **Acoustic filters** block out certain frequencies of sounds
  - *Low-pass filter* blocks high frequency components of a waveform
  - *High-pass filter* blocks low frequencies
  - *Band-pass filter* blocks both around a band
  - *Reject band* (what to block) vs. *pass band* (what to let through)

- But if frequencies of two sounds overlap…. *source separation issues*
Estimating pitch

- **Pitch tracking**: Estimate F0 over time as a function of *vocal fold vibration* (*vowels.wav*)
- **How?** Autocorrelation approach
  - A periodic waveform is correlated with itself since one period looks much like another
  - Find the period by finding the ‘lag’ (offset) between two windows on the signal for which the correlation of the windows is highest
  - Lag duration (T) is 1 period of waveform
  - Inverse is F0 (1/T)
• Microprosody effects of consonants (e.g. /v/)
• Creaky voice $\rightarrow$ no pitch track
• Errors to watch for in reading pitch tracks:
  – **Halving**: shortest lag calculated is too long $\rightarrow$ estimated cycle too long, too *few* cycles per sec (*underestimate* pitch)
  – **Doubling**: shortest lag too short and second half of cycle similar to first $\rightarrow$ cycle too short, too *many* cycles per sec (*overestimate* pitch)
then I don't know if I can explain it to you

1 1 1 0 1 1 1 1 1 1
Next Class

• Download Praat from the course syllabus page
• Read the Praat tutorial
• Record 2 files: your name in one file and these English vowels in another file (/iy/, /ih/, /ei/, /ae/, /ow/, /aa/) and save them to disk
• Bring a laptop with the files and headphones to class (if you have – otherwise we’ll share)