11-324/11-624/11-724 Human Language for AI

Acoustic Phonetics II

David R. Mortensen

September 13, 2022

Language Technologies Institute
Carnegie Mellon University

Introduction: Consonants are Acoustically Diverse

Vowels and consonants both have distinctive acoustic structures. The acoustics of vowels are relatively straightforward, which is to be expected—vowels have basically one manner of articulation. Consonants have many manners of articulation, and each manner of articulation has its acoustic peculiarities.

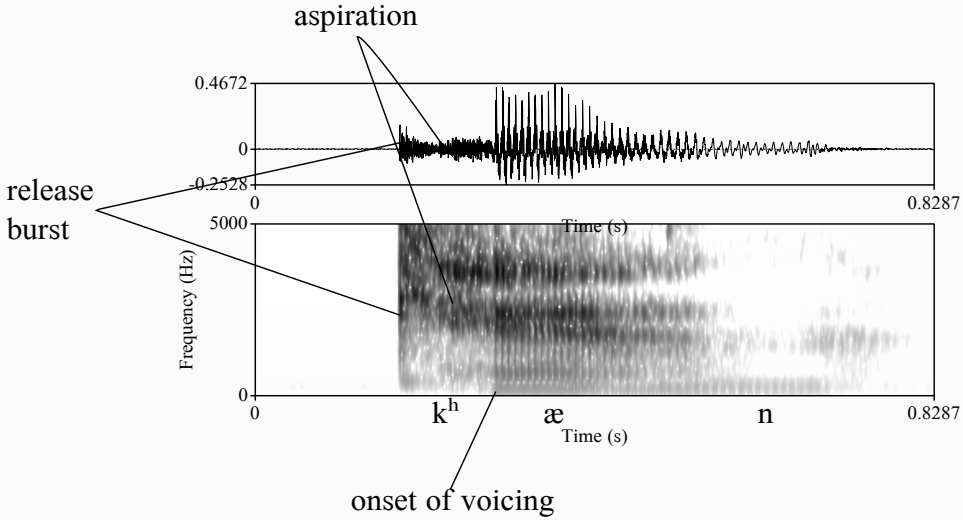
Learning Objectives

At the ends of today's lecture, students will have acquired the following knowledge and skills:

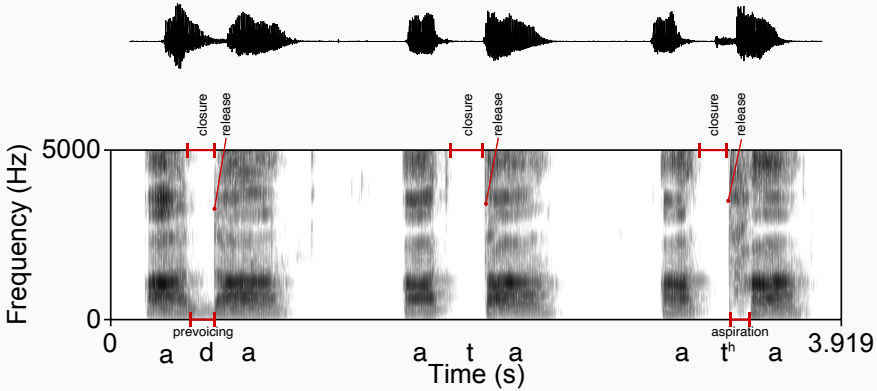
- Students will be able to distinguish vowels and consonants given a waveform and a spectrogram.
- Students will be able to identify the manner of articulation of different consonants given just a waveform and a spectrogram.
- Students will know how articulatory events map on to acoustic events
- They will know the following terms:
    - Closure
    - Release
    - Release burst
    - Voicing
    - Voice onset time
    - Turbulence
    - Aperiodic noise
    - Antiresonance and antiformant
    - Mel (unit of pitch)
    - MFCC
    - Acoustic model
- Students will know tone is realized acoustically (primarily as pitch)
- Students will know how phonetic knowledge can be applied to speech technologies research and development

Acoustics of Consonants

The Articulatory Landmarks of Plosives Correspond to Acoustic Landmarks

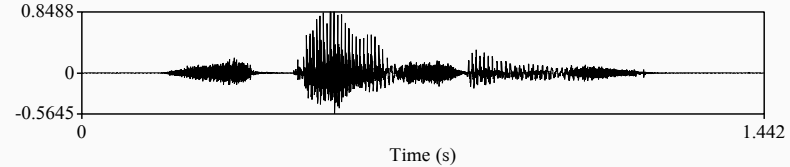


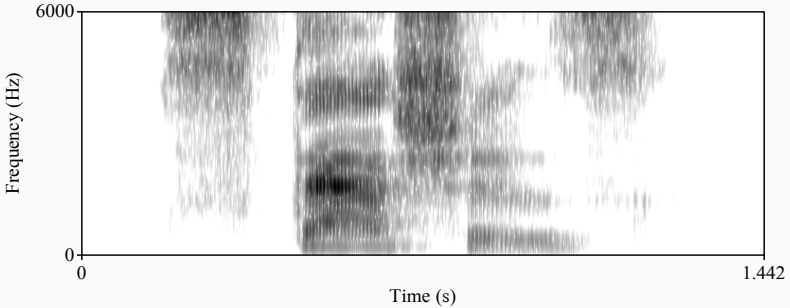
Plosives at One Place Can Differ in Voice Onset Time



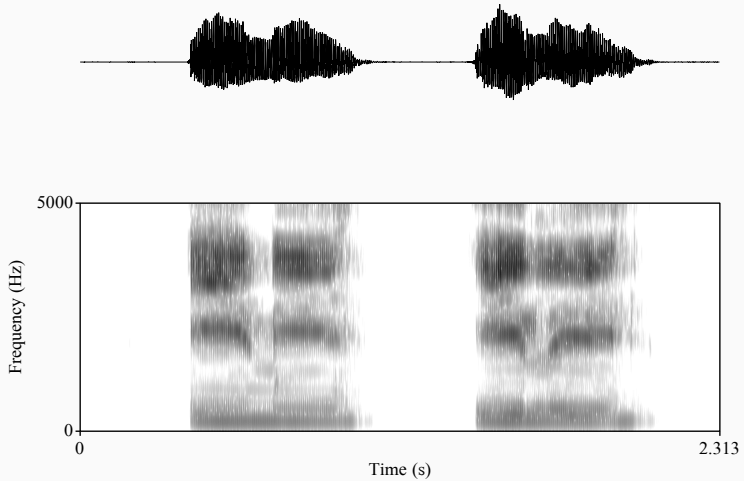
Waveform and spectrogram of [ada ata at^{h}a].

Fricatives at Different Places Can Differ in Center of Gravity

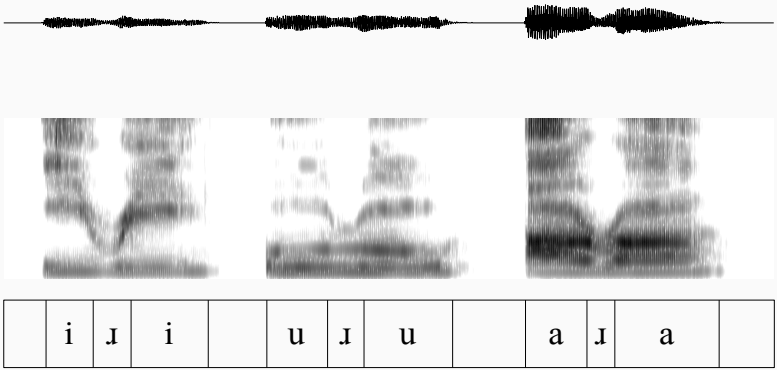




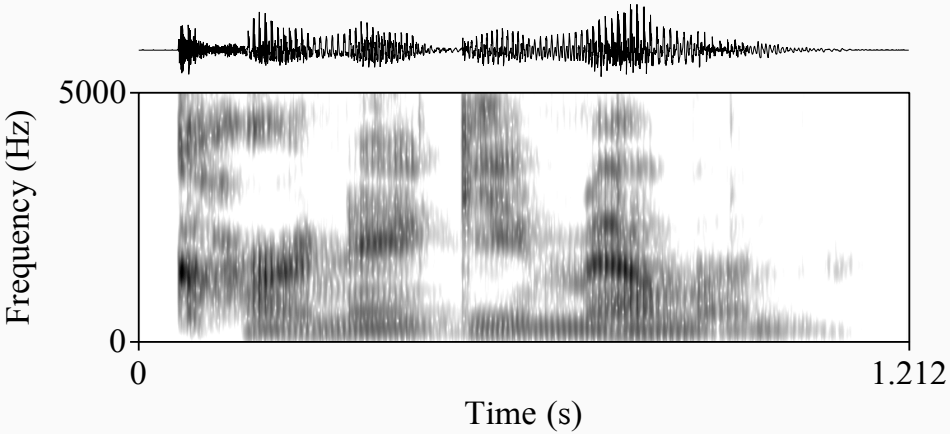
Acoustic Representations of [ini] and [ili]



Acoustics of /ɹ/ (and /ɻ/)

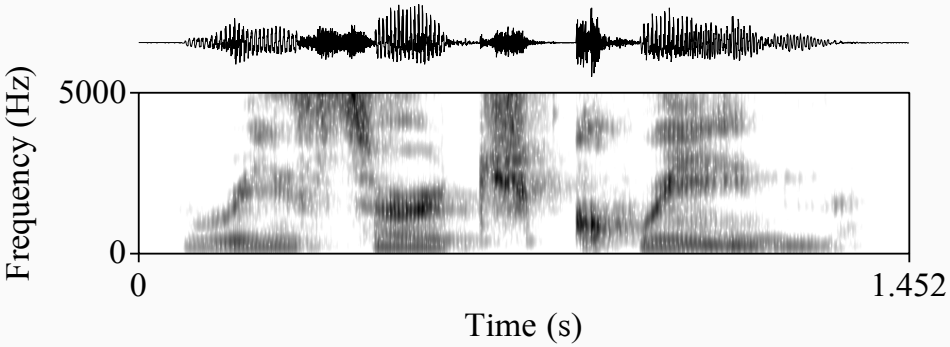


Mystery Spectrogram I



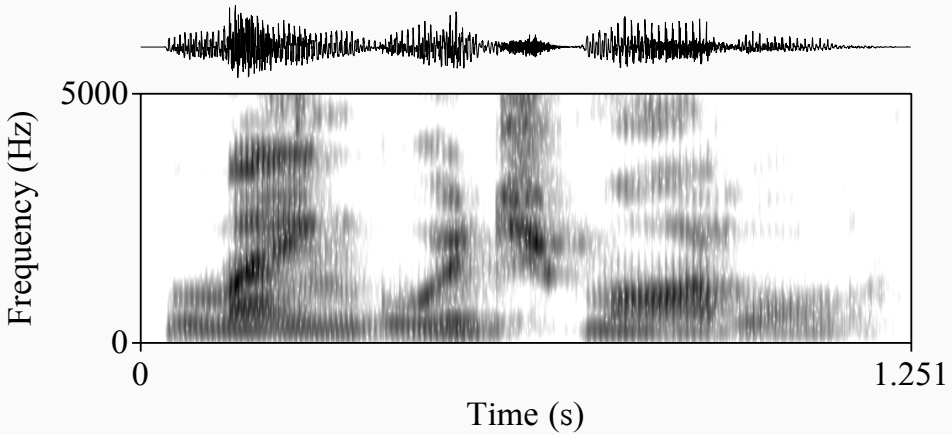
What does this spectrogram say?

Mystery Spectrogram II



What does this spectrogram say?

Mystery Spectrogram III



What does this spectrogram say?

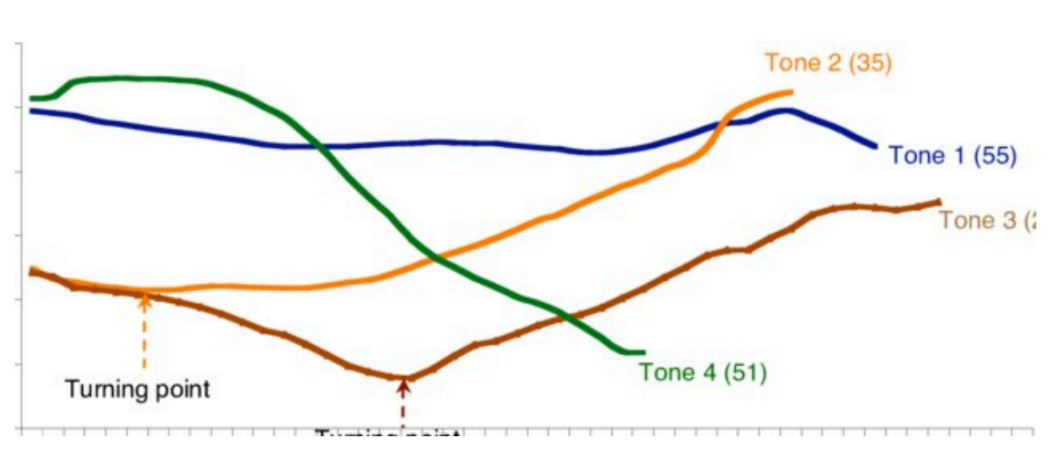
Tone and Intonation

Tone Is the Use of Pitch to Distinguish Words

- Many languages use pitch in the same way that they use consonant and vowel phones
- When used in this way, pitch differences are called TONE
- Well-known languages with tone include:
  - Chinese (all varieties)
  - Most languages of Southeast Asia (but not Malay, etc.)
  - Most languages of Sub-Saharan Africa (but not Swahili)
  - Swedish and Norwegian
  - Serbian and Croatian

To See Tone, Look at the Pitch Track

The four tones of Standard Mandarin (Putonghua) as pitch tracks:

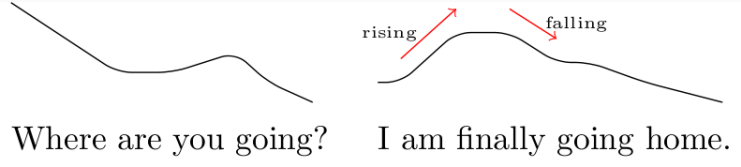


All Tones are Smooth

- Some tone languages are described as having “high” tones or “low” tones, but—when you look at a pitch track—all you will see is smooth contours
- This is true whether the tones are “level” tones or “contour” tones

Almost All Langugaes Have Intonation

- Intonation is when languages use pitch for linguistic purposes other than distinguishing words
- Almost all languages have it, even tone languages



Phonetics and Speech Technologies

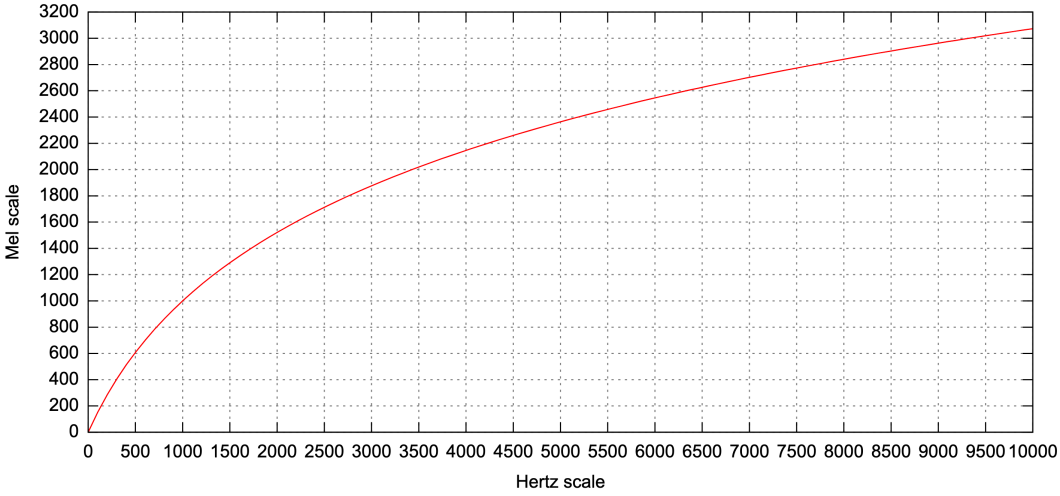
There Has Been Phonetic Forgetting in Speech Technologies

- Speech technologists used to know a **lot** about phonetics
- In many cases, attempts were made to model speech acoustics explicitly in both speech synthesis and speech recognition
  - Synthesizing speech by modeling the acoustics of the vocal tract
  - Recognizing speech by analyzing it into psycholinguistically relevant features like formant values.
- This did not work as well as data-oriented approaches to speech
- Gradually, as ML (especially neural) models have become better, less use has been made of explicit phonetic knowledge in speech technologies
- As a result, speech researchers have "forgotten" much of what they once knew about phonetics
- They still function

An Example of Forgetting: MFCCs versus Spectrograms

- At an earlier time, acoustic signals for speech recognition (ASR) were represented as **MFCCs** (Mel Frequency Cepstral Coefficients).
- Steps for making MFCCs
  - Take the Fourier transform of a windows signal
  - Map the resulting spectrum onto the MEL SCALE (see next slide)
    - Triangular overlapping windows or
    - Cosine overlapping windows
  - Take the log of the powers at each mel frequency
  - Treat the resulting list as if it were a signal and take the discrete cosine transform
  - The resulting amplitudes are the MFCCs
- MFCCs encode considerable phonetic biases
- Now, it is more common to just feed an end-to-end neural ASR system **spectrograms** and let it figure out all of the phonetic patterns itself

The Mel Scale is a Psychoacoustic Representation of Pitch



Phonetics is Still Relevant to Speech Technology

Does that mean that AI researchers and developers **shouldn't** bother to learn about phonetics? There are a number of reasons that speech technologies can still benefit from phonetic knowledge:

- **Analyzing systematic errors**. For example:
  - Confusion of similar sounds (in recognition)
  - Confusion of sounds in particular contexts (in recognition)
  - "Unnatural" combinations of sounds in synthesis

  Knowing how to augment training data or change how it is presented to a system can benefit from phonetic knowledge
- **Enriching or preprocessing training data with phonetic knowledge**.
  - "Grapheme-to-phoneme" (usually grapheme-to-phone) transduction
  - Injecting phonetic features
- **Data annotation protocols**

Questions?