SPRING 2023

CS 378: INTRO TO SPEECH AND AUDIO PROCESSING

Speech Sounds 1

DAVID HARWATH
Assistant Professor, UTCS

The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Welcome!

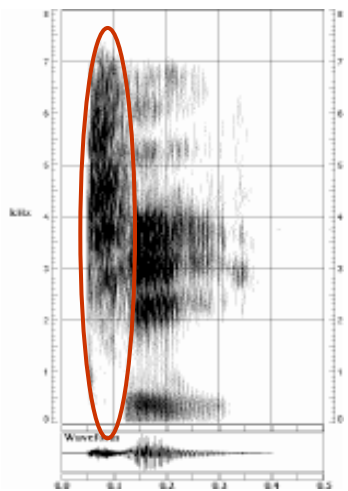
- Problem set 1 out on Canvas; try working on problem 1

- Today: learning to read spectrograms
  - Reading spectrograms is lots of fun
  - Many good ideas for ML algorithms are born from an *intuition about the underlying phenomenon being modeled*
  - Spectrogram reading will give you an intuitive understanding of the speech signal that you cannot get any other way

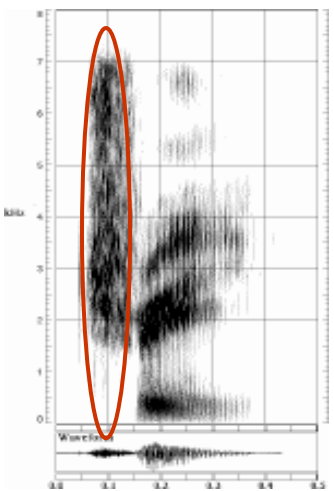
A few quick definitions

- ***Word:*** The unit of language that is the primary carrier of meaning. Composed of a sequence of phonemes.
- ***Phoneme***: The *abstract* "building block" of a word. If you change a phoneme, you change the word.
  - OOP Analogy: abstract class – can't be instantiated
- ***Phone***: A *concrete acoustic realization* of a phoneme. Multiple phones can map to the same phoneme (allophones)
  - OOP Analogy: subclass of an abstract class – can be instantiated

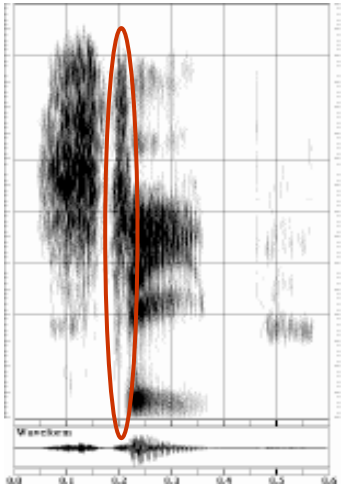
Some allophones of the phoneme /t/



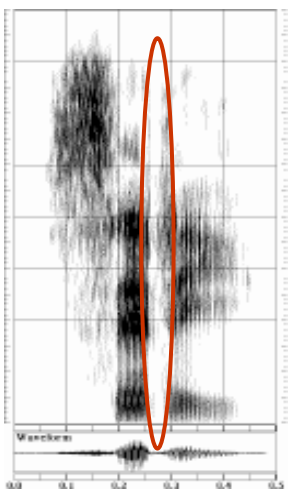
TEA



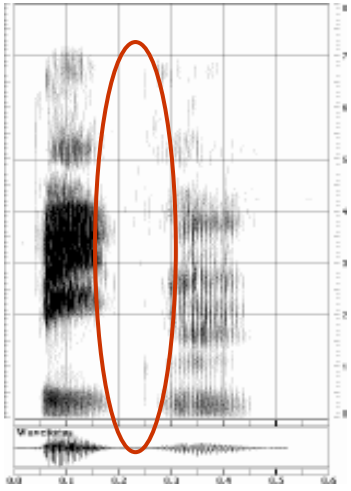
TREE



STEEP



CITY



BEATEN

Phonemes in American English*

*A slightly non-standard notation based on the International Phonetic Alphabet (IPA) that we will use in this class.

Vowels

/iʸ/	beat
/ɪ/	bit
/eʸ/	bait
/ɛ/	bet
/æ/	bat
/ɑ/	bob
/ɔ/	bought
/ʌ/	but
/oʷ/	boat
/ʊ/	book
/uʷ/	boot
/ɝ/	Burt
/ɑʸ/	bite
/ɔʸ/	Boyd
/aʷ/	bout
/ə/	about

Fricatives

/s/	see
/š/	she
/f/	fee
/θ/	thief
/z/	z
/ž/	Gigi
/v/	v
/ð/	thee

Stops

/p/	pea
/t/	tea
/k/	key
/b/	bee
/d/	Dee
/g/	geese

Semivowels

/w/	wet
/r/	red
/l/	let
/y/	yet

Nasals

/m/	meet
/n/	neat
/ŋ/	sing

Affricates

/č/	church
/ǰ/	judge

Aspirant

/h/	heat

The International Phonetic Alphabet (IPA)

- English has approximately 40 phonemes, depending on who you ask.
- There are hundreds (thousands?) of different speech sounds used across all the world's 7,000 languages
- The International Phonetic Alphabet (IPA) is an *orthographic system* used by linguists to organize and transcribe phonemes across languages
- Today we will only look at American English phonology
  - But later in the semester we will discuss multilingual ASR

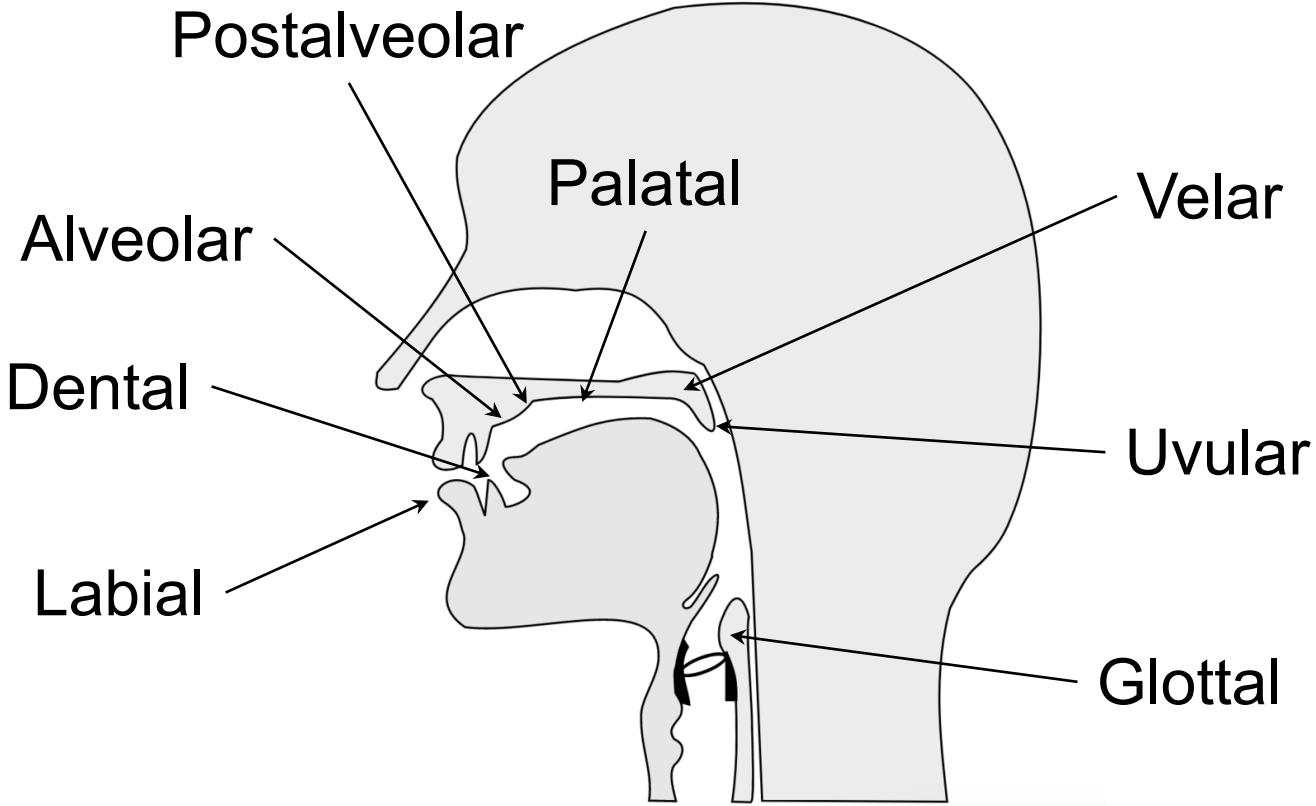
Manner and Place of Articulation

- Manner of Articulation: **How** you use your speech articulators to produce a phone
  - Vowels, fricatives, nasals, stops…
- Place of Articulation: **Where** the production of a phone is happening at a *specific location in your mouth*
  - Labial, dental, alveolar, palatal, velar, uvular…

English Manner Classes

- Vowels: No constriction of the vocal tract. Excitation from vocal fold vibration
  - Example: “aaaaaah”
- Fricatives: Constriction of the vocal tract resulting in turbulent noise
  - Example: “ssssss”
- Stops: Complete closure of vocal tract, followed by a burst/release
  - Example: “p”
- Nasals: Lowering of the velum, allowing airflow into nasal cavity
  - Example: ”mmmmm”
- Semivowels: A somewhat constricted vowel
  - Example: “llllll”
- Affricates: A stop concatenated with a fricative
  - Example: “ch”
- Aspirant: No constriction, just turbulent airflow
  - Example: “hhhhhh”

Places of Articulation



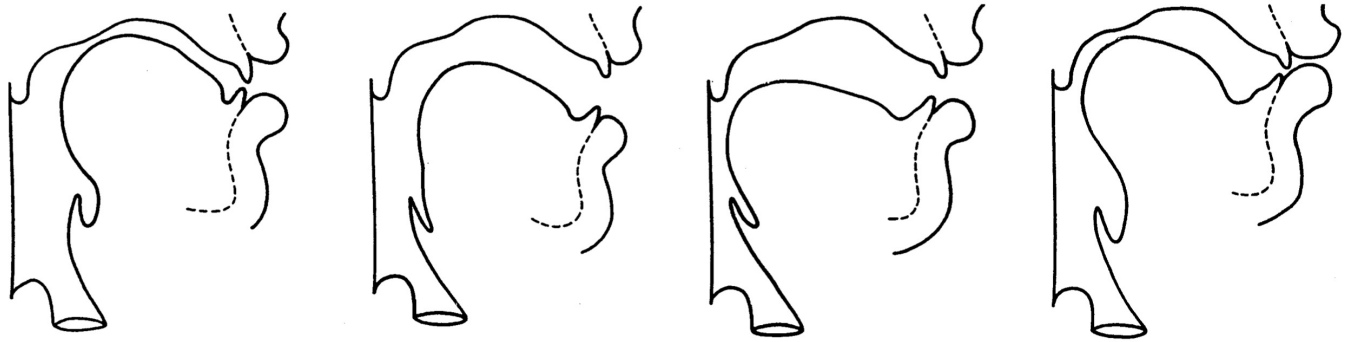
Context Dependency

- Phonemes define articulatory “targets” that your articulators attempt to reach when you speak
- You cannot move your articulators instantaneously.
  - You will typically start moving your articulators towards the next phonemic target while still producing the current phone
- Neighboring phonemes “compete” for your articulators, giving rise to co-articulation (and phonetic realization)

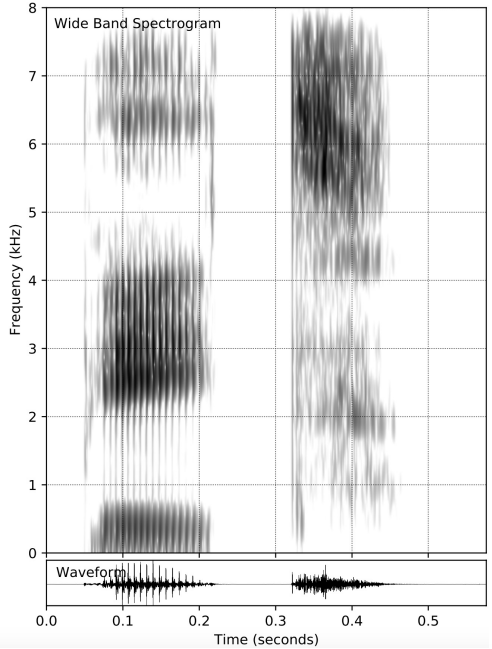
Vowels

- Production: No constriction in vocal tract + excitation by vocal fold vibration (voicing)
- Acoustic characteristics (formant frequencies) depend on shape of vocal tract (position of tongue, jaw, lips)

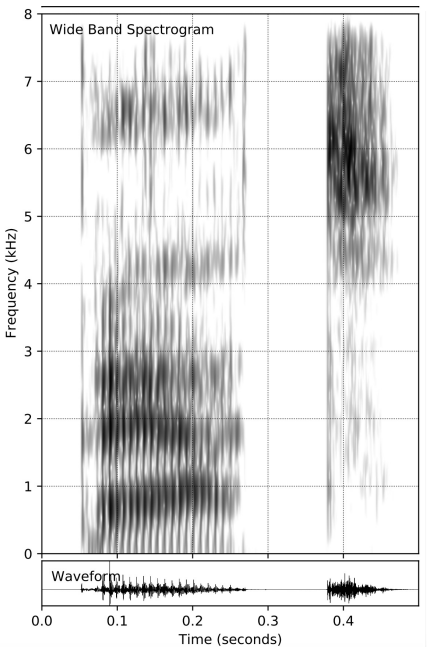
[i] [æ] [ɑ] [u]



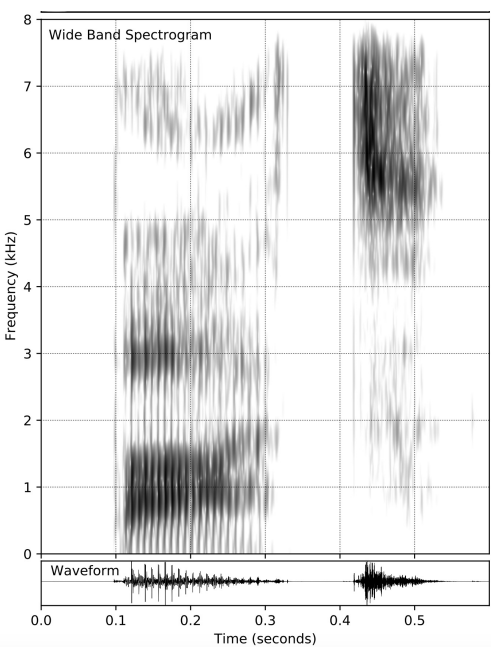
English Cardinal Vowels



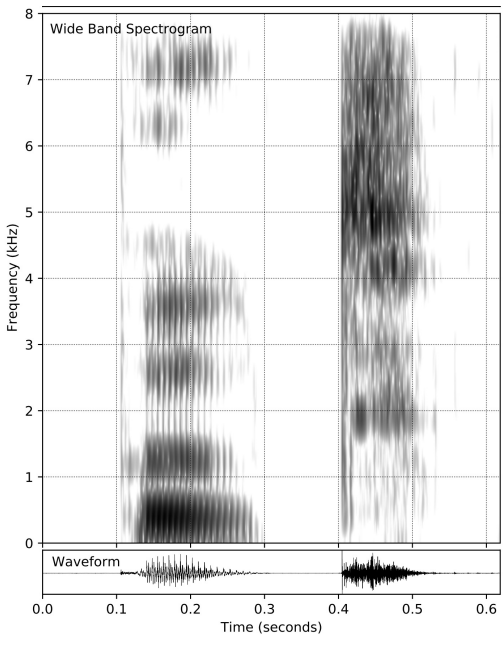
beet
/biʸt/



bat
/bæt/



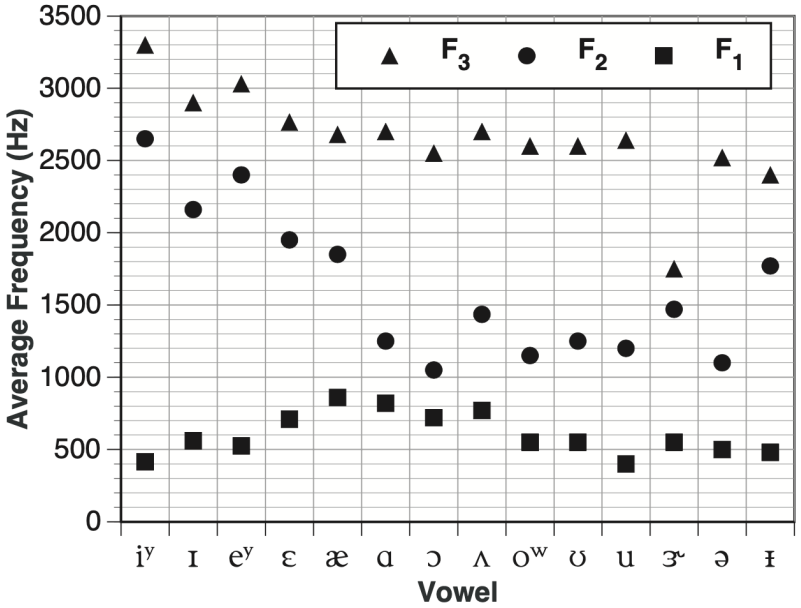
bott
/bɑt/



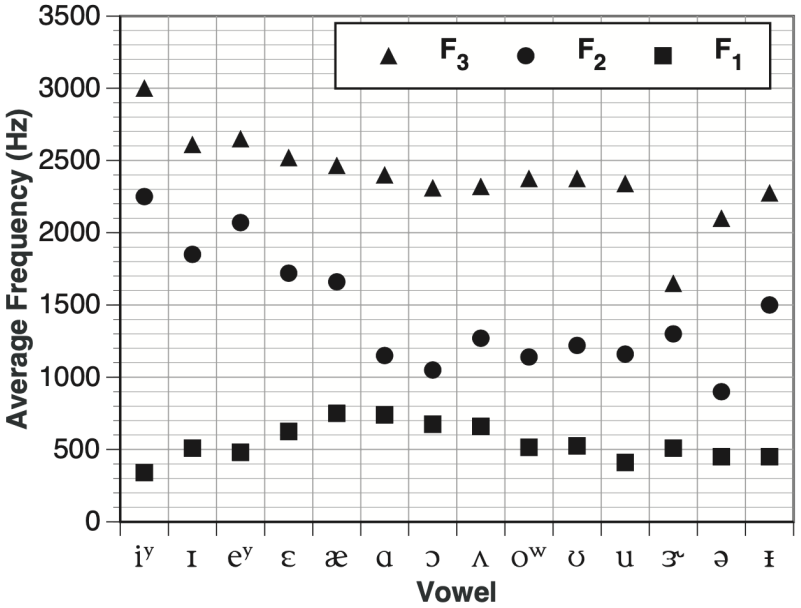
boot
/but/

Average Vowel Formant Locations

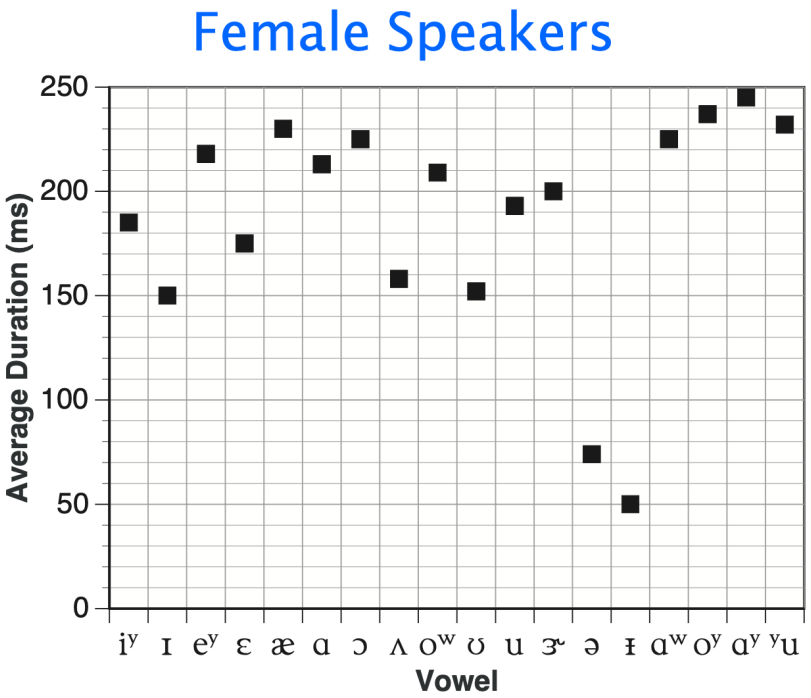
Female Speakers

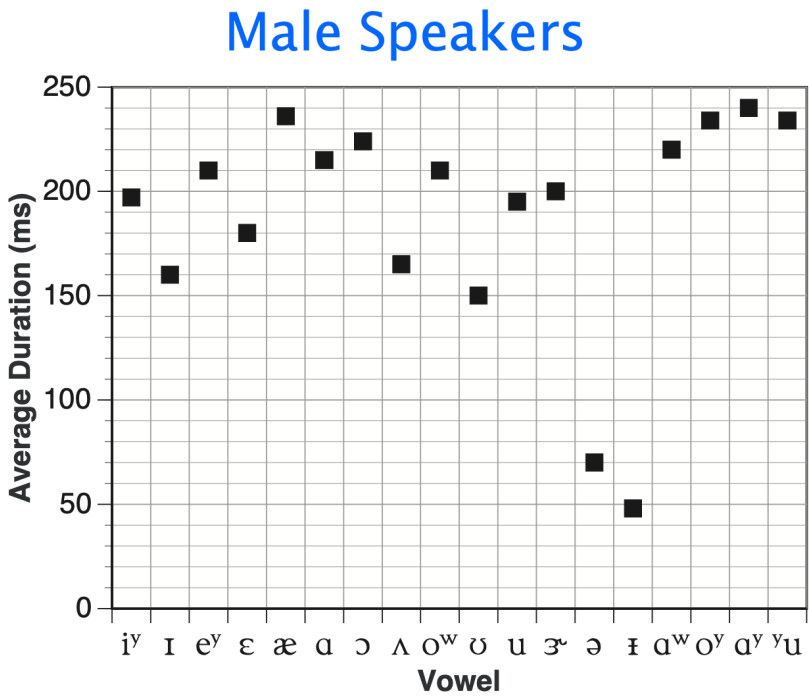


Male Speakers

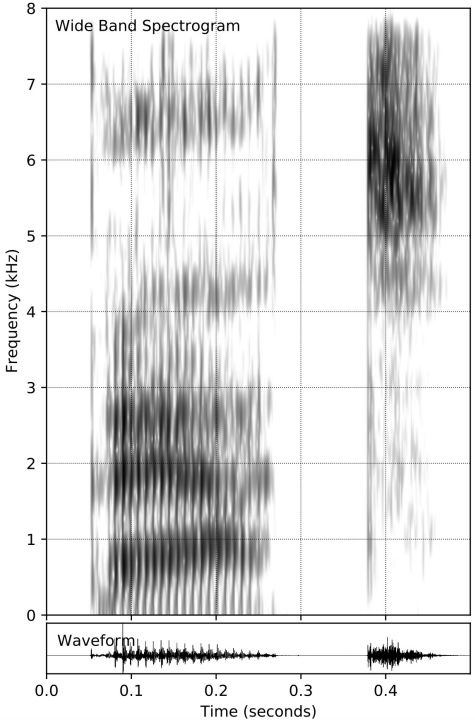


Average Vowel Durations

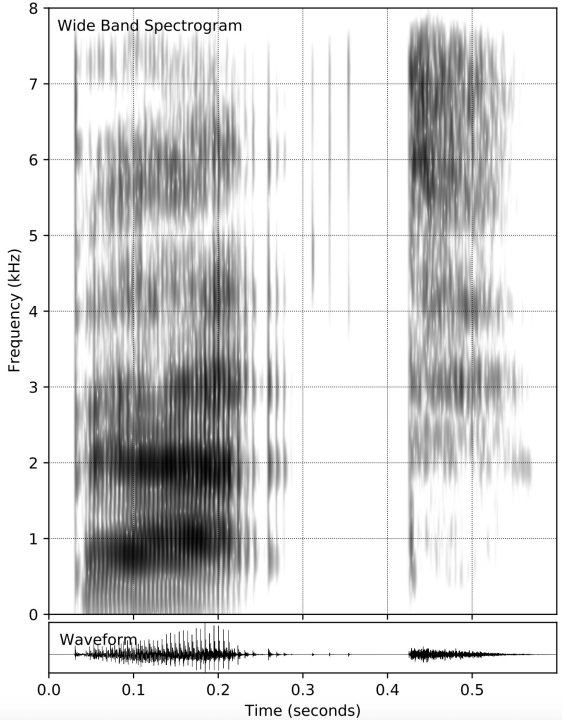




Average F0



Male speakers: ~120 Hz



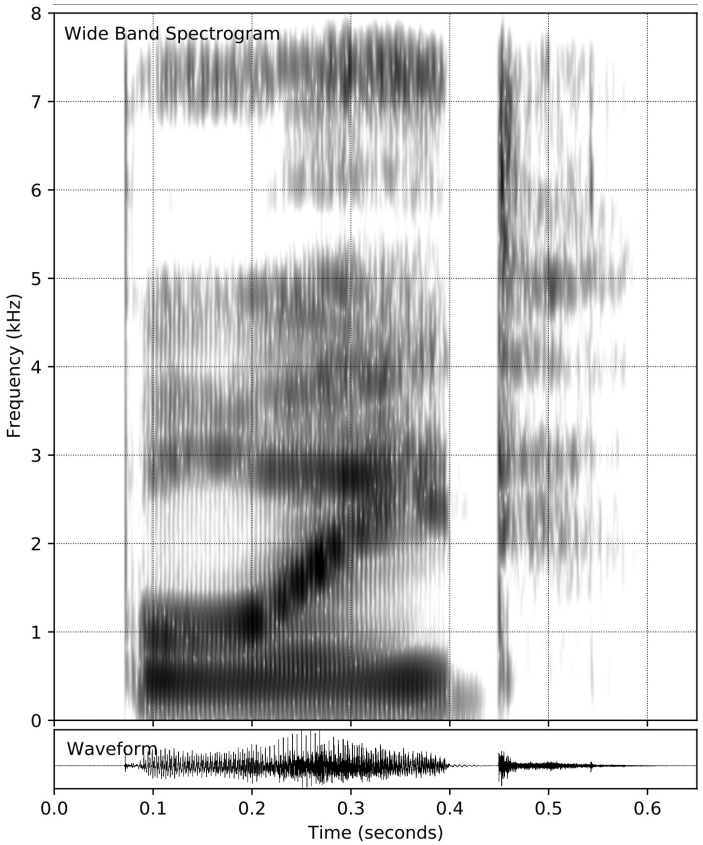
Female speakers: ~210 Hz

Dipthongs

- A “moving” vowel (or: two vowels merged together)
- Typically have a long duration (200-250ms)

Dipthongs

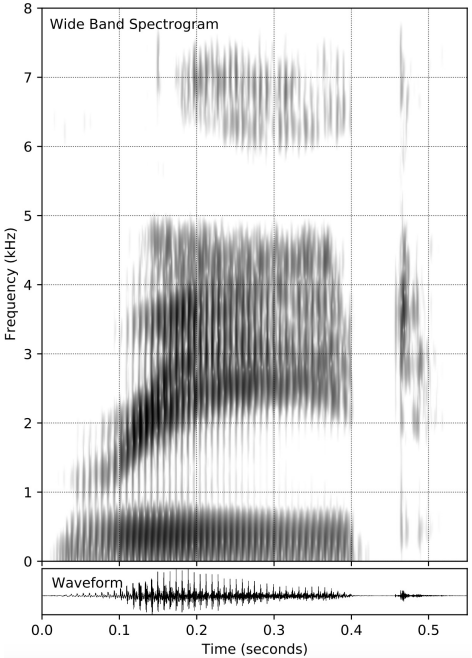
$$a^{w} = a \rightarrow w$$
$$o^{y} = o^{w} \rightarrow i^{y}$$
$$a^{y} = a \rightarrow i^{y}$$
$$^{y}u = y \rightarrow u$$
$$e^{y} = \varepsilon \rightarrow i^{y}$$



Boyd

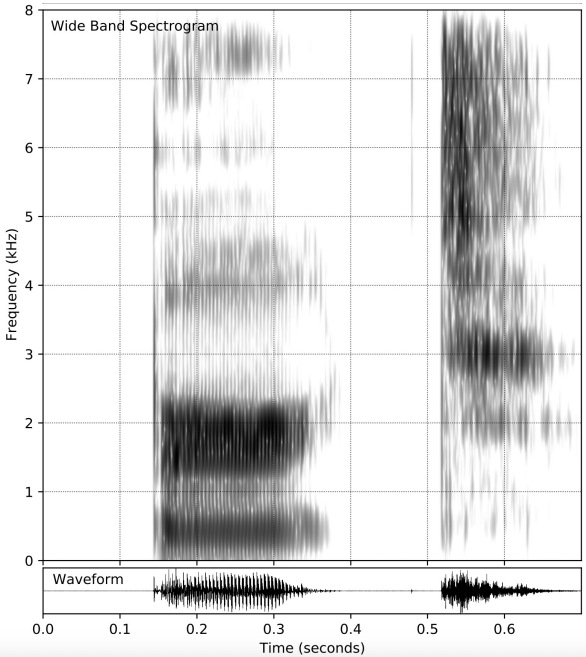
/bo^{y}d/

Retroflexion



reed

/riʸd/

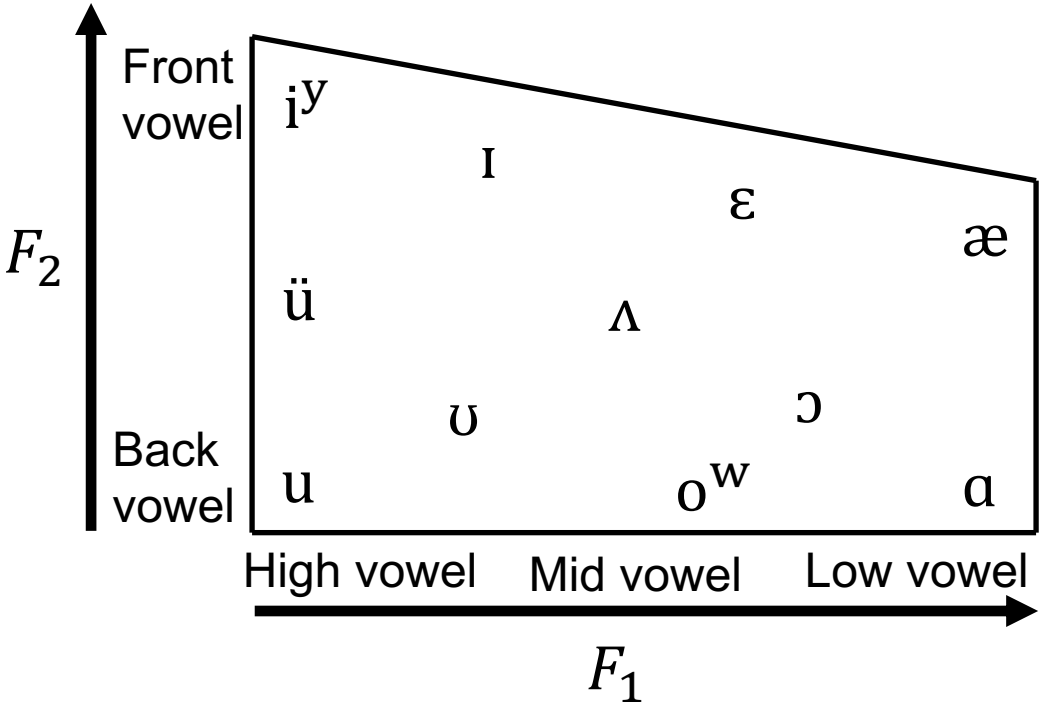


Burt

/bɜˠt/

The Vowel Chart

Retroflex (very low F_3): ɾ (short) ɝ (syllabic)



Front schwa: ɨ (High F_2, <50ms duration)

Back schwa: ə (Low F_2, <50ms duration)

Dipthongs (200-250ms)

$$ɑ^w = ɑ \rightarrow w$$
$$o^y = o^w \rightarrow i^y$$
$$a^y = ɑ \rightarrow i^y$$
$$^y u = y \rightarrow u$$
$$e^y = ɛ \rightarrow i^y$$

Short Vowels (~75-175ms):

ɪ, ɛ, ʌ, ʊ

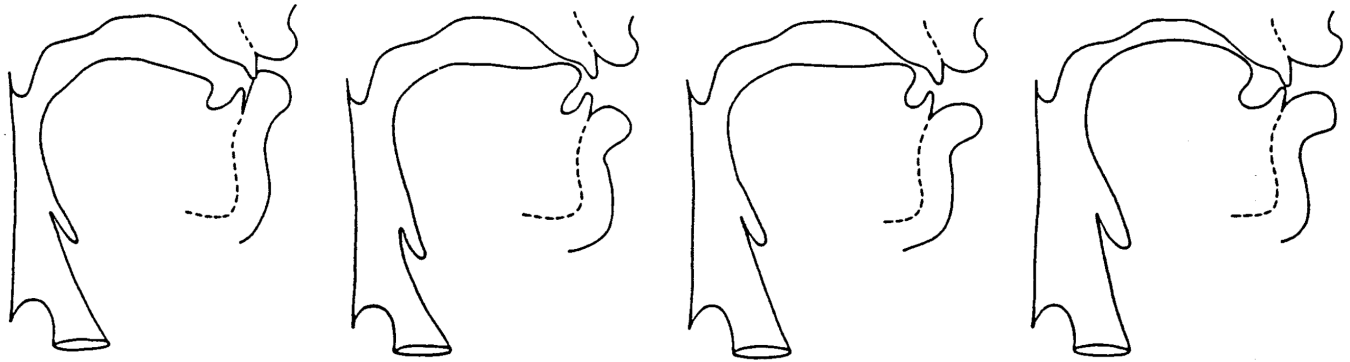
Long Vowels (~150-250ms):

i^y, æ, ɑ, ɔ, o^w, u, ɝ

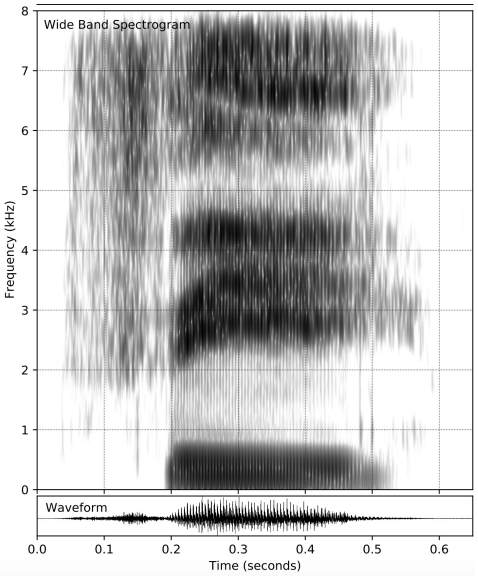
Fricatives

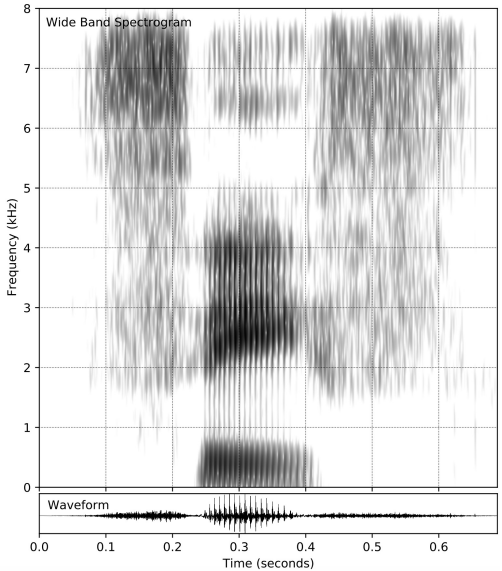
- A constriction in the vocal tract that is excited either by turbulent airflow (unvoiced) or vocal fold vibration (voiced)
- Spectral characteristics depend on location of constriction

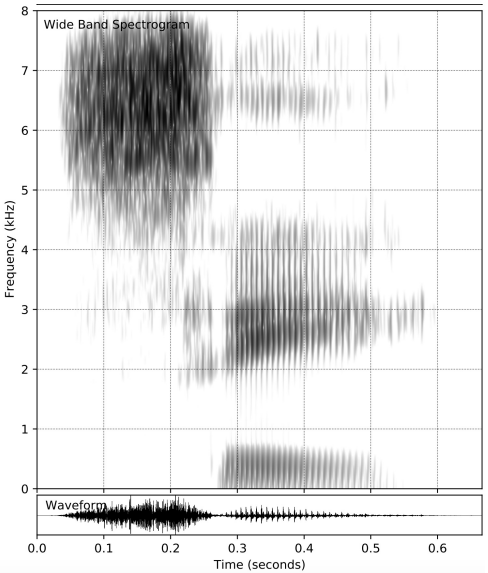
[f] [θ] [s] [š]

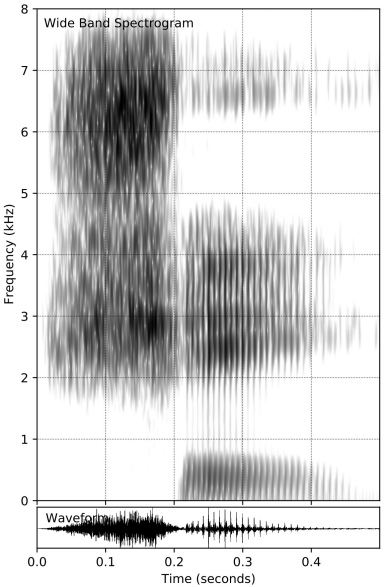


Unvoiced Fricatives









fee /fi^{y}/

thief /θi^{y}f/

see /si^{y}/

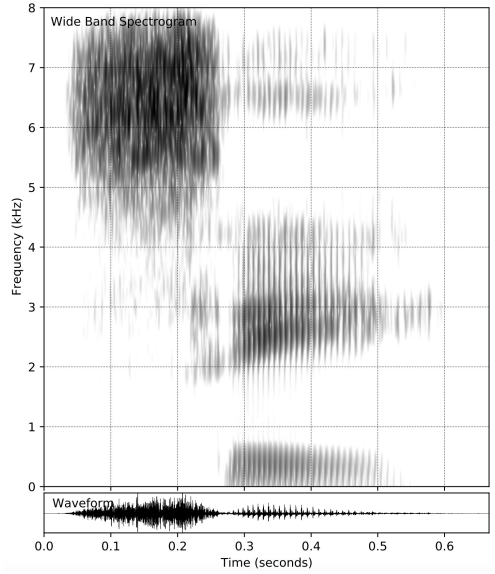
she /ši^{y}/

American English Fricatives

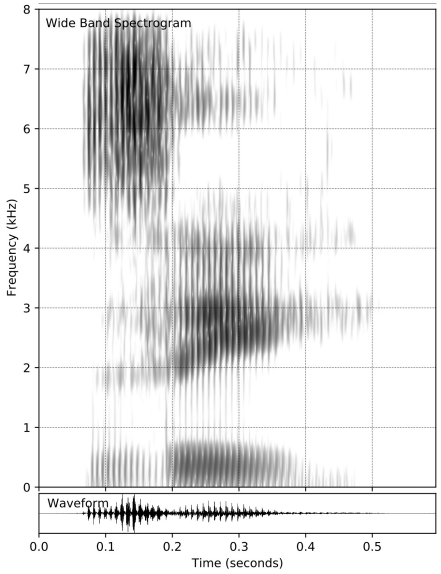
- 8 different fricatives in American English
- Defined by 2 features: voicing and place of articulation (PoA)

	Place of Articulation	Unvoiced	Voiced
Non-strident (weak)	Labial (Labiodental)	/f/ fee	/v/ v
	Dental	/θ/ thief	/ð/ thee
Strident (strong)	Alveolar	/s/ see	/z/ z
	Palatal	/š/ she	/ž/ Gigi

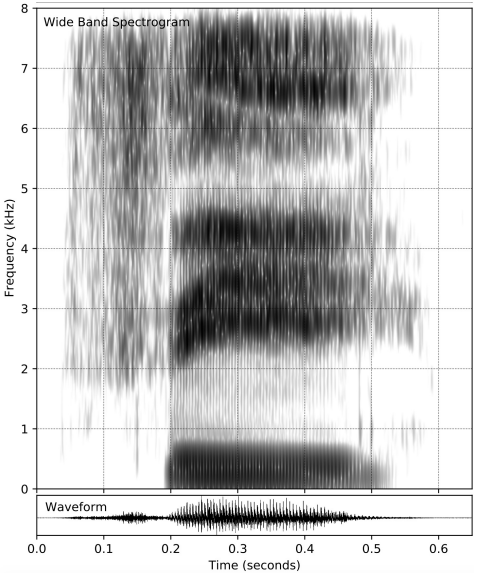
Voiced vs. Unvoiced Fricatives



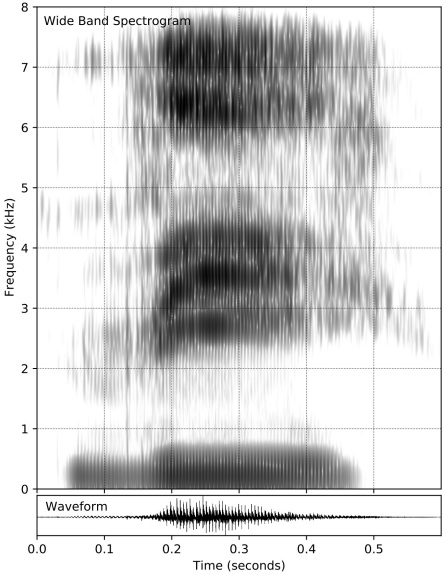
see
/si^{y}/



z
/zi^{y}/

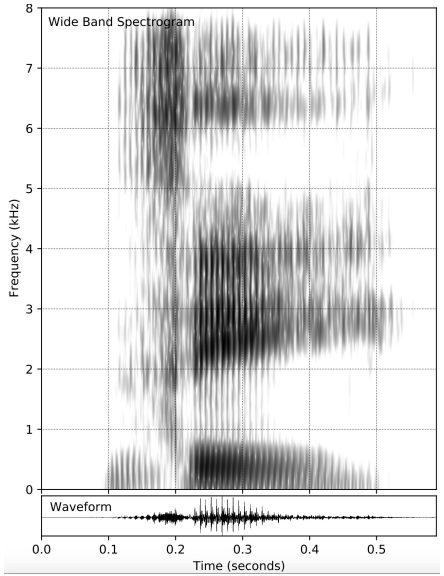


fee
/fi^{y}/

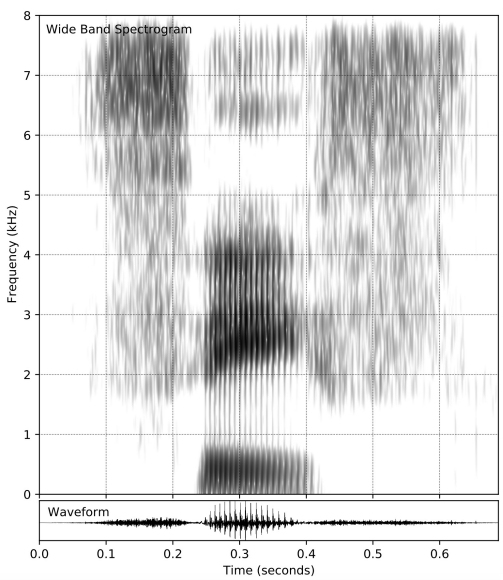


v
/vi^{y}/

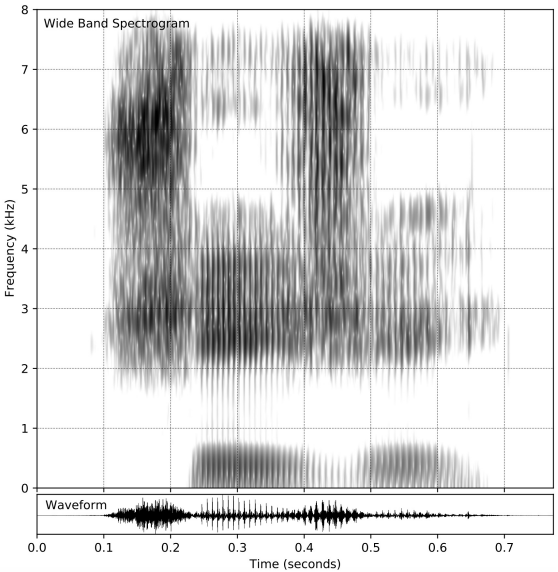
Voiced vs. Unvoiced Fricatives



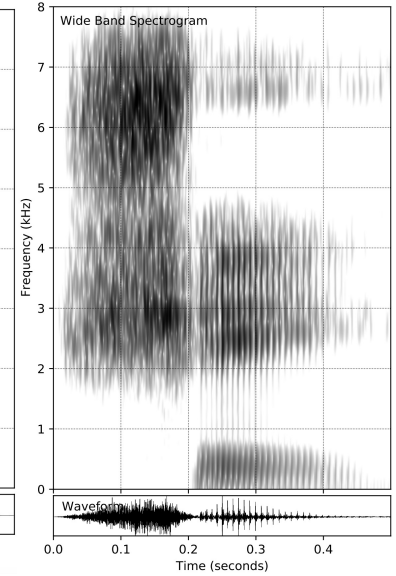
thee
/ðiʸ/



thief
/θiʸf/

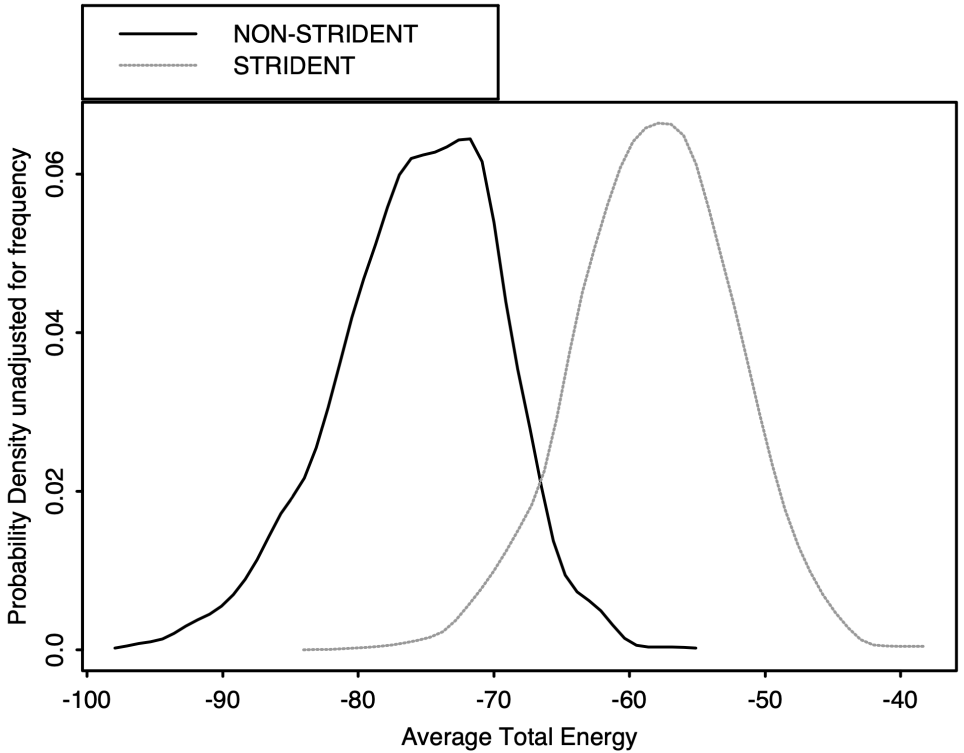


gigi
/žiʸžiʸ/



she
/šiʸ/

Fricative energy depends on PoA



Fricative duration depends on voicing