SPRING 2023

CS 378: INTRO TO SPEECH AND AUDIO PROCESSING

Speech Sounds 2

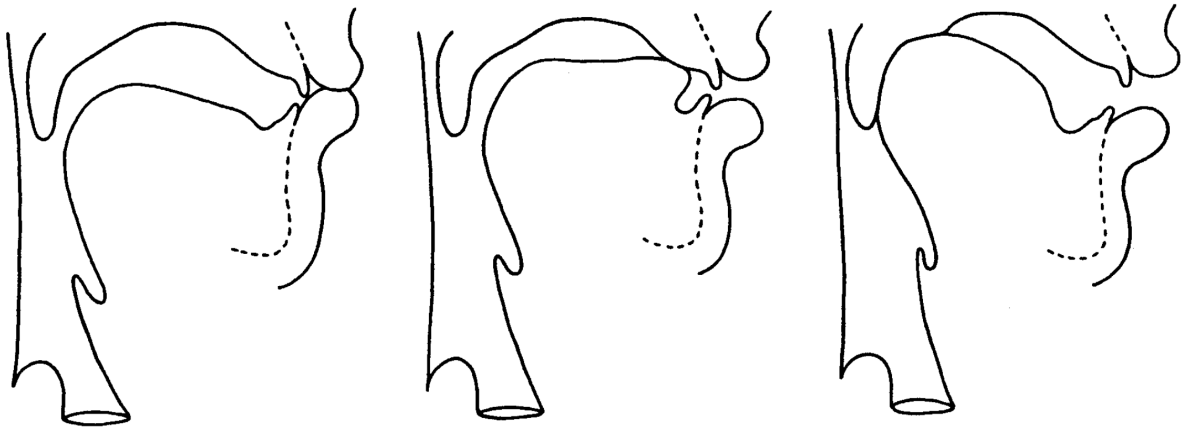
DAVID HARWATH
Assistant Professor, UTCS

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

Nasals

- The lips or tongue are used to *close off* the front cavity, and the velum is lowered to allow airflow out through the nose
- Excitation is from vocal fold vibration

[m] [n] [ŋ]

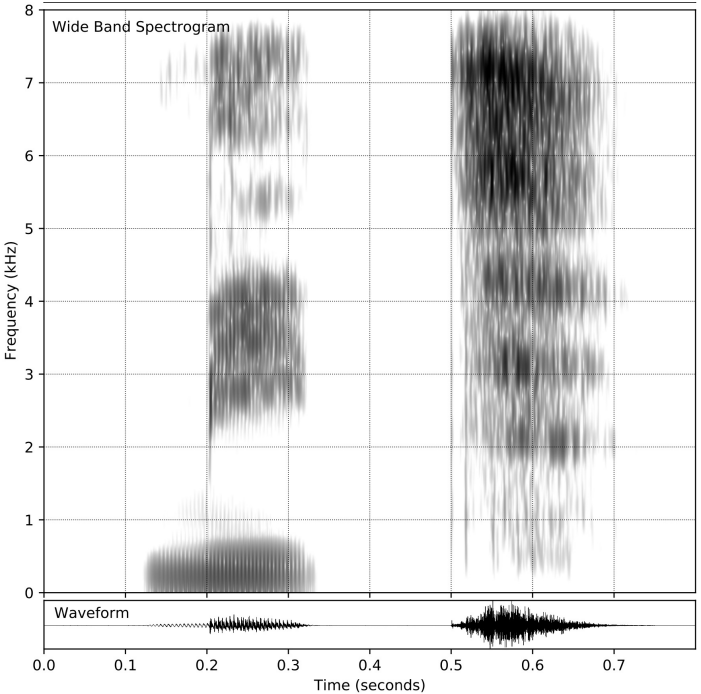


American English Nasals

- All nasals look very similar spectrally (low frequency murmur)
- English vowels are always attached to a vowel
  - The main cues to look for are formant transitions
    - /m/ generally pulls formants down
    - /ŋ/ often introduces velar pinch
    - /n/ doesn't impact neighboring formants much

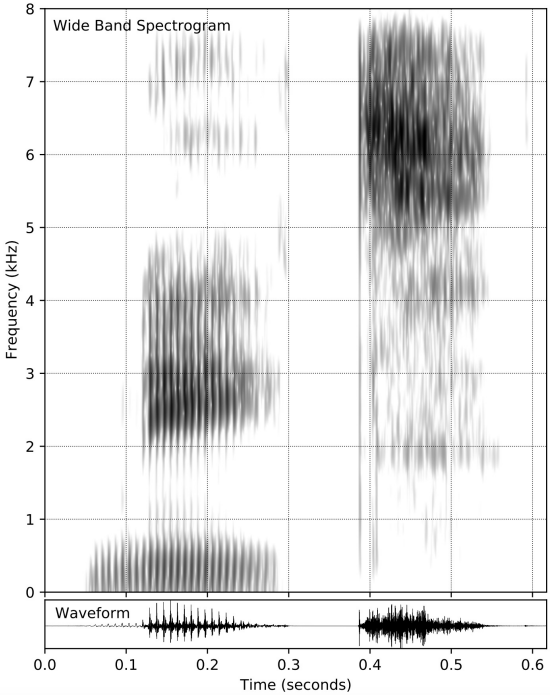
Place of Articulation	Phoneme
Labial (Bilabial)	/m/ meet
Alveolar	/n/ neat
Velar	/ŋ/ sing

Nasal Spectrograms



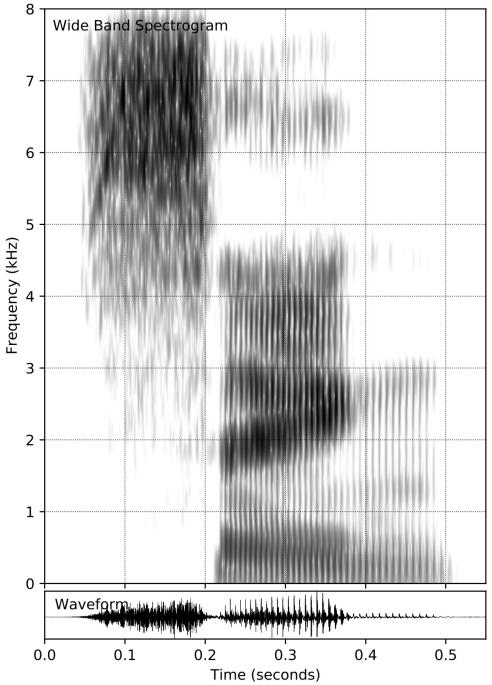
meet

/miʸt/



neat

/niʸt/



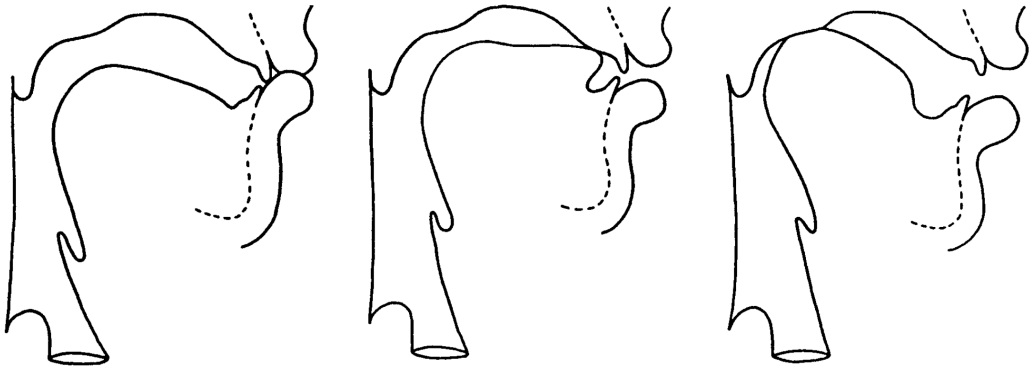
sing

/siʸŋg/

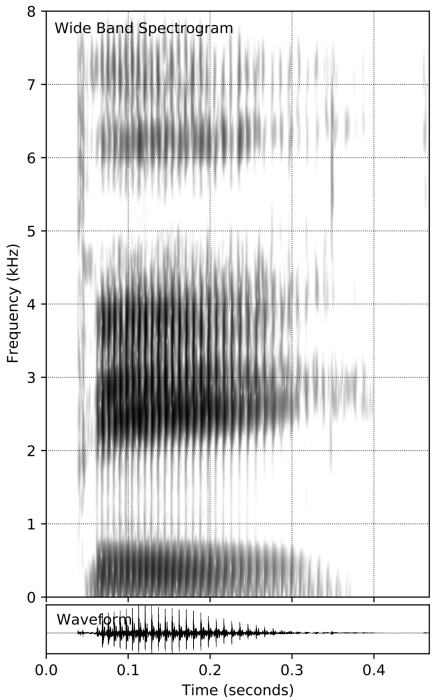
Stops

- *Complete* closure of the vocal tract (silence), followed by pressure release (burst), followed by frication/aspiration
- Spectral characteristics depend on PoA and voicing

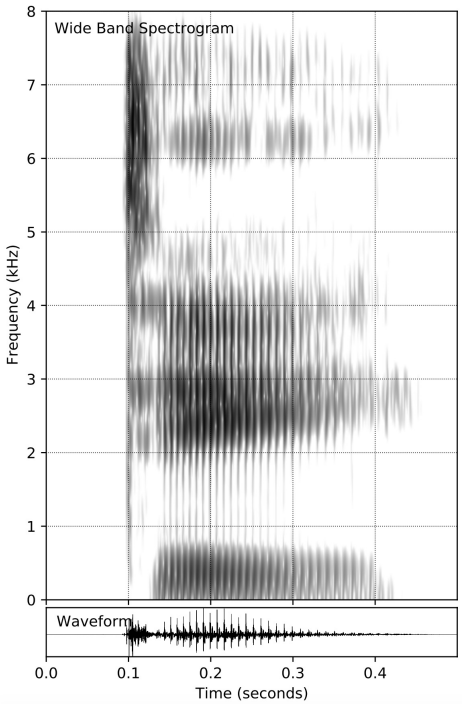
[b] [d] [g]



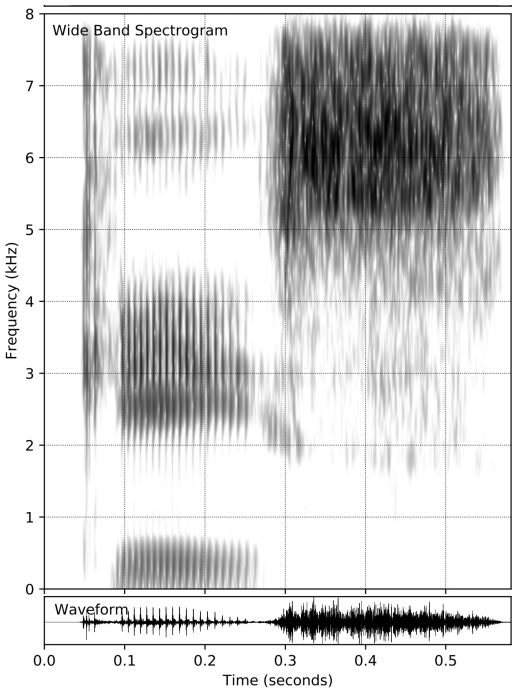
Voiced Stops



bee
/bi^{y}/

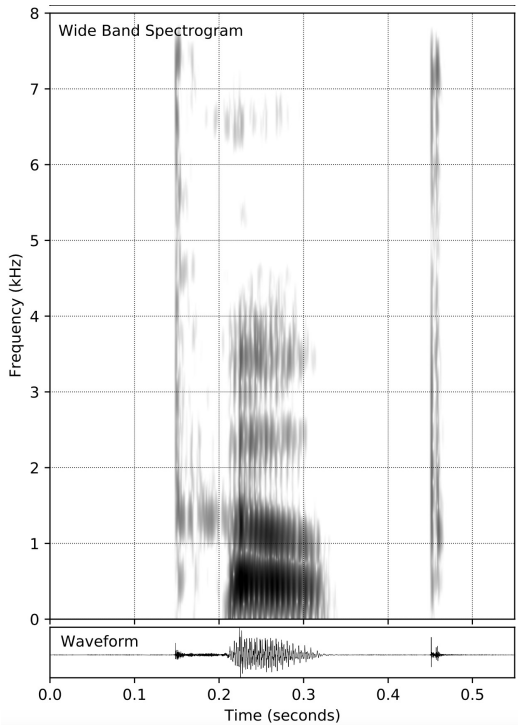


Dee
/di^{y}/

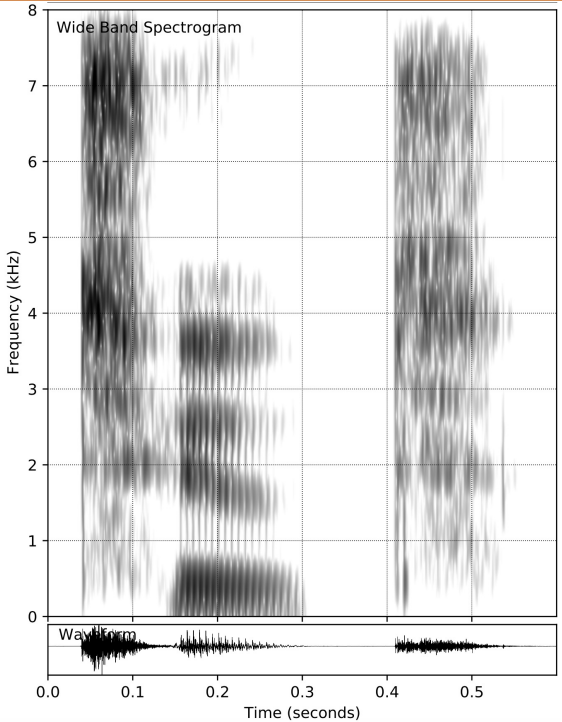


geese
/gi^{y}s/

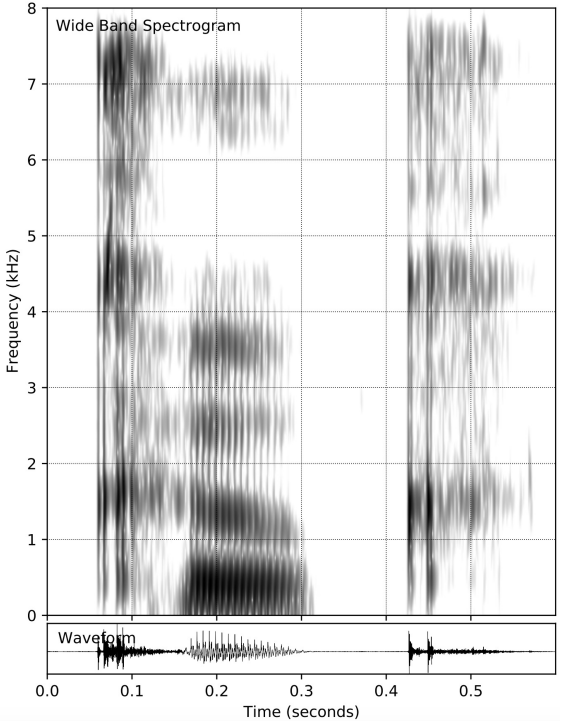
Unvoiced Stops



poop
/pup/



toot
/tut/



kook
/kuk/

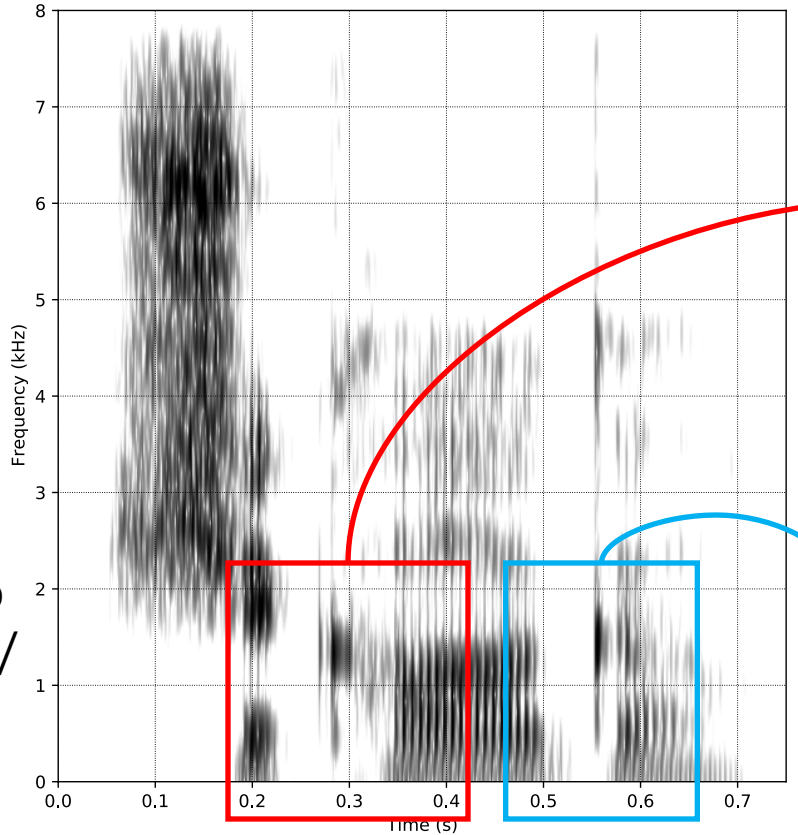
American English Stops

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

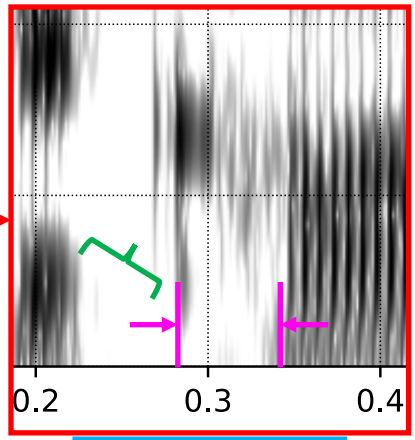
Place of Articulation	Unvoiced	Voiced
Labial (Bilabial)	/p/ fee	/b/ v
Dental	/t/ thief	/d/ thee
Velar	/k/ see	/g/ z

Acoustic cues: voicebar, voice onset time, burst spectrum, aspiration noise, neighboring formant transitions

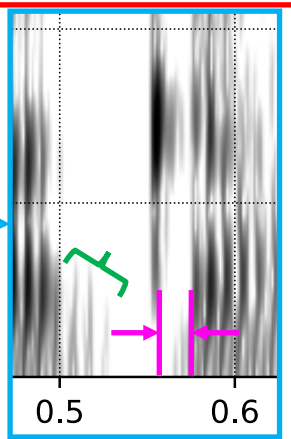
Voice Onset Time (VOT) and Voicebars



Chicago
/šɬkɔgoʷ/

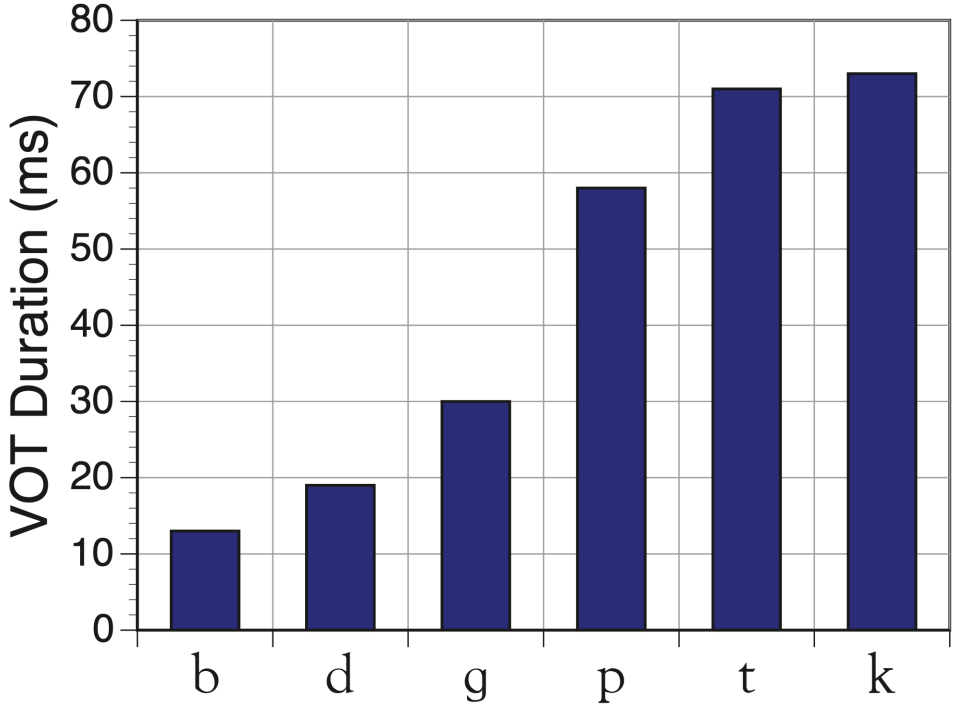


/k/
Unvoiced
No voicebar
Long VOT



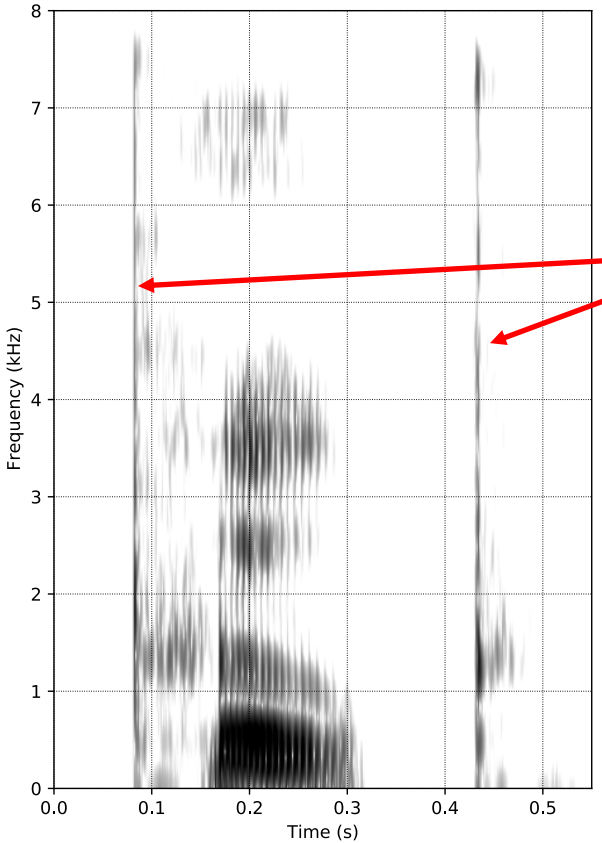
/g/
Voiced
Has voicebar
Short VOT

Stop Voice Onset Times (Average)



Spectra: Labial Stops

poop
/pup/

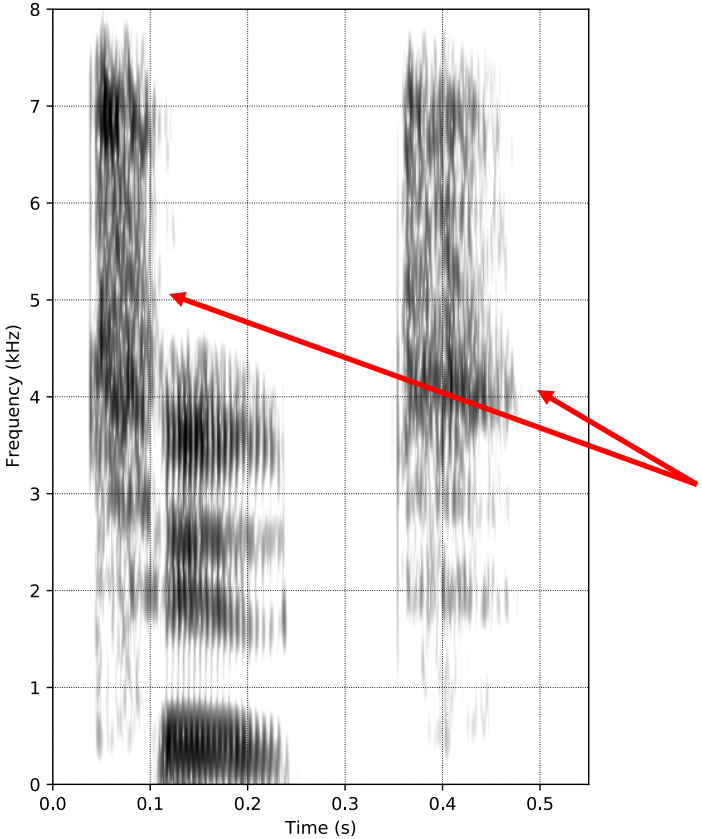


Labial stops (/p/ and /b/) often have a uniform, wideband burst that looks like a vertical pencil line

Aspiration is often weak but wideband

Spectra: Dental Stops

toot
/tut/

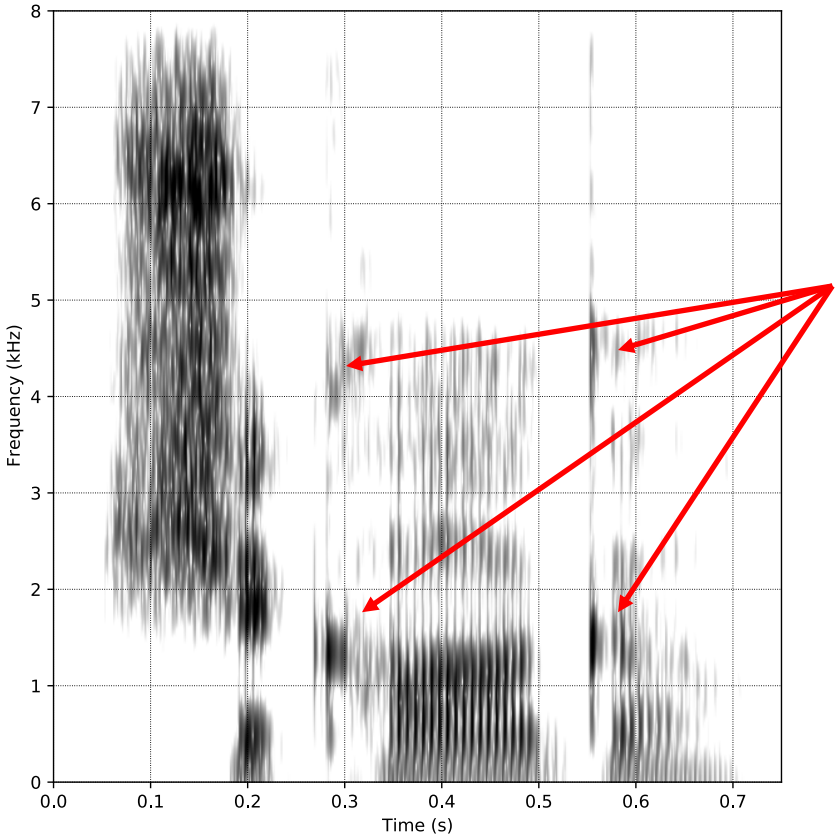


Dental stops (/t/ and /d/) often have a messier burst than labial stops

But, there is often a stronger frication after the burst that can look /s/-like

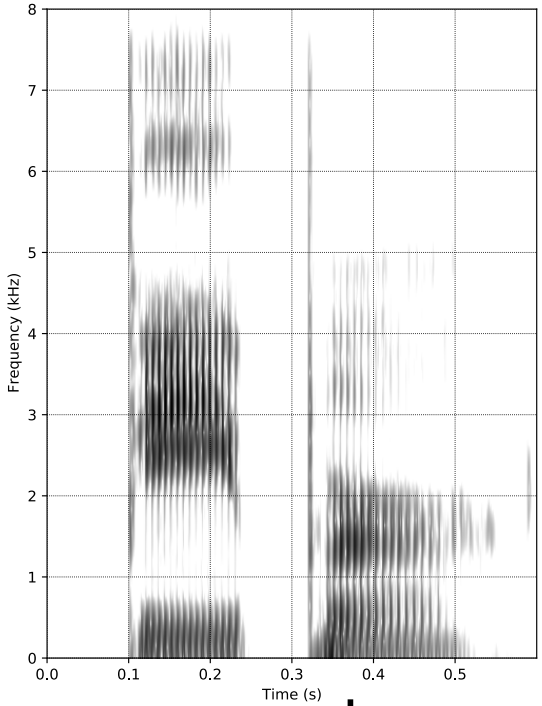
Spectra: Velar stops

Chicago
/šɬkɔgo^{w}/

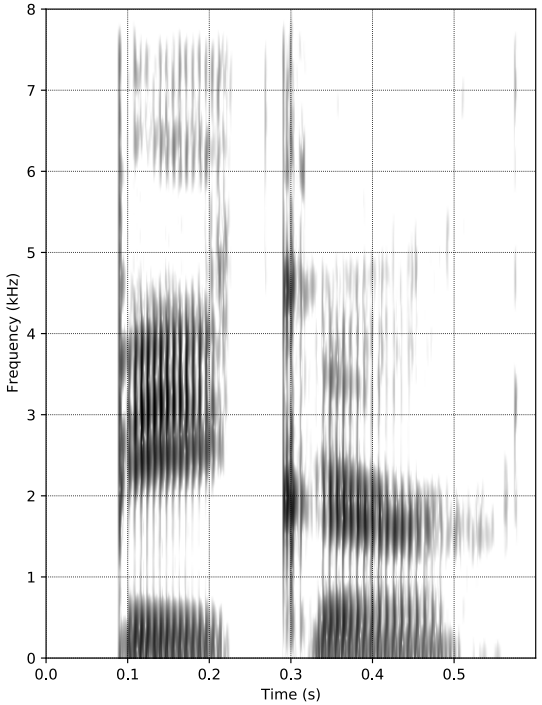


Velar stop bursts often have characteristic concentrations of energy at 1.5 kHz and 4.5 kHz

Co-articulation with vowels



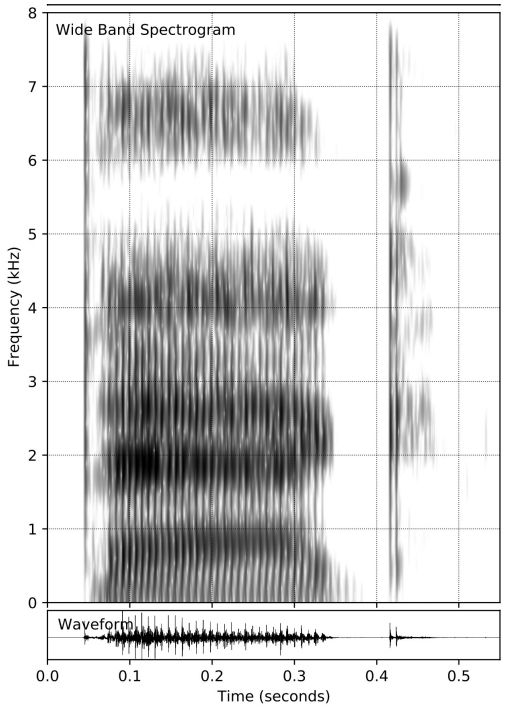
beeper
/biʸpɜˠ/



beaker
/biʸkɜˠ/

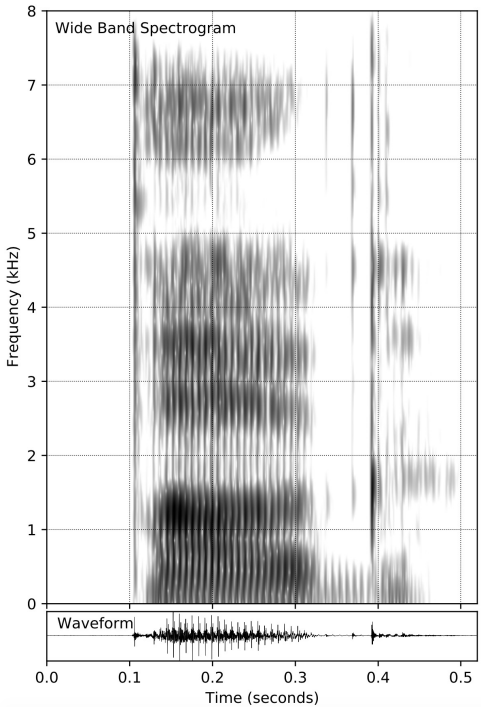
Labial stops can pull down surrounding formants

Co-articulation with vowels



bag

/bæg/

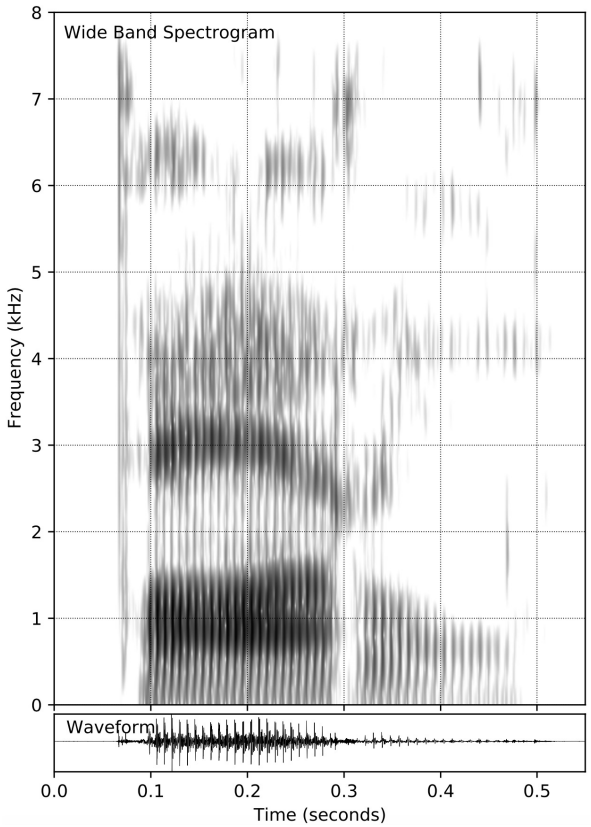


bug

/bʌg/

Velar stops “pinch” f2 and f3 of surrounding front vowels (but generally not back vowels)

Flaps



bottle
/bɑrl/

A flap is a heavily reduced stop, usually a /t/ or /d/ phonemically.

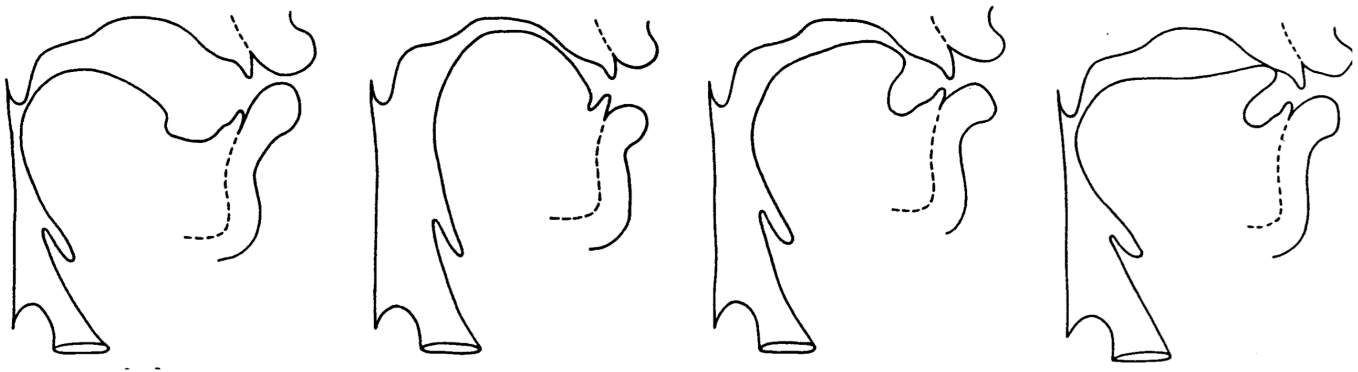
Rather than fully articulating the consonant, you quickly tap your tongue against the roof of your mouth

Think of the /t/ in words like “bottle” or “butter”

Semivowels

- A vowel with a constriction that does not produce turbulence noise
- Can be thought of as an “extreme” vowel

[w] [y] [r] [l]

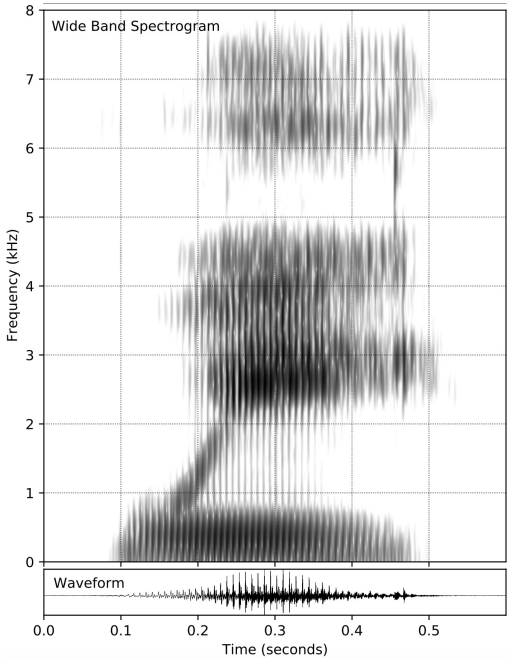


American English Semivowels

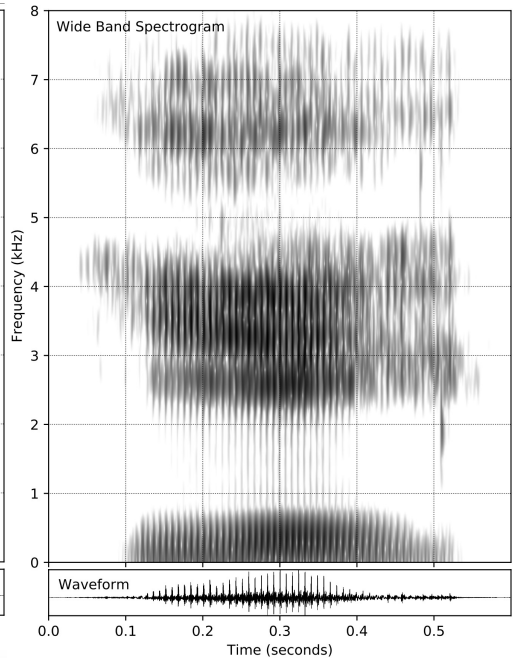
- 4 semivowels in American English
- Glides use tongue body to form constriction: /w/ and /y/
- Liquids use tongue tip to form constriction: /l/ and /r/
- Formants are similar to their corresponding vowel, but in more extreme positions.

Semivowel	Closest Vowel
/w/	/u/
/y/	/i^{y}/
/r/	/ɜ˞/
/l/	/o^{w}/

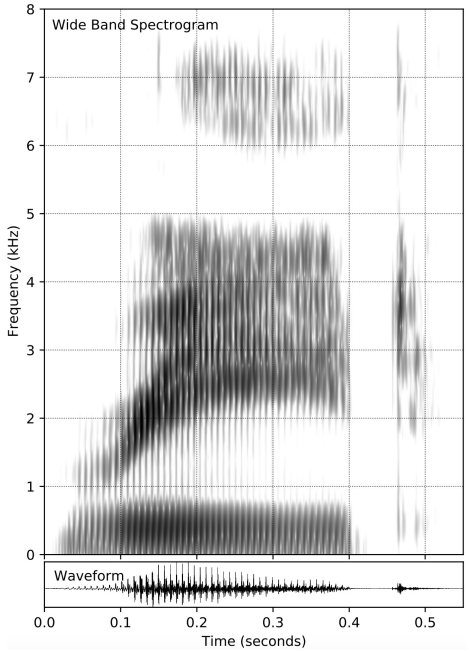
Semivowel Spectrograms



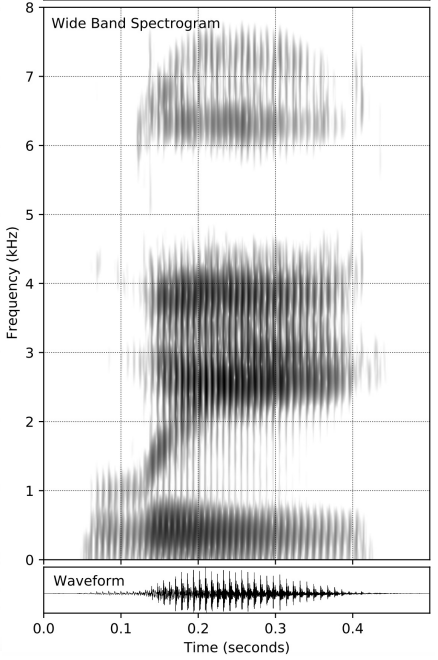
we
/wi^{y}/



ye
/yi^{y}/

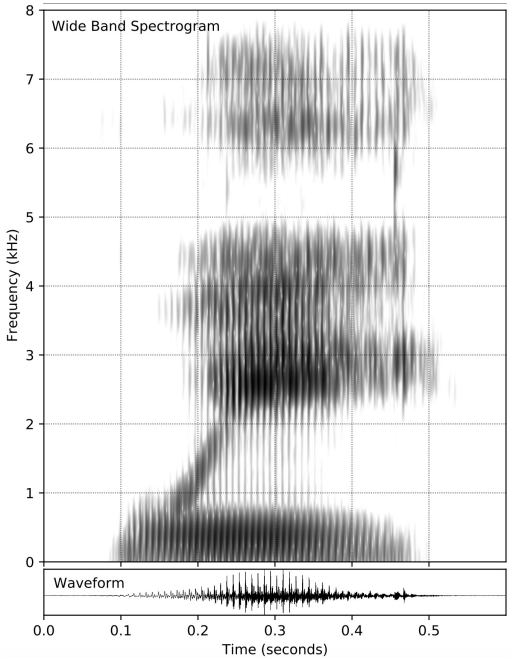


reed
/ri^{y}d/



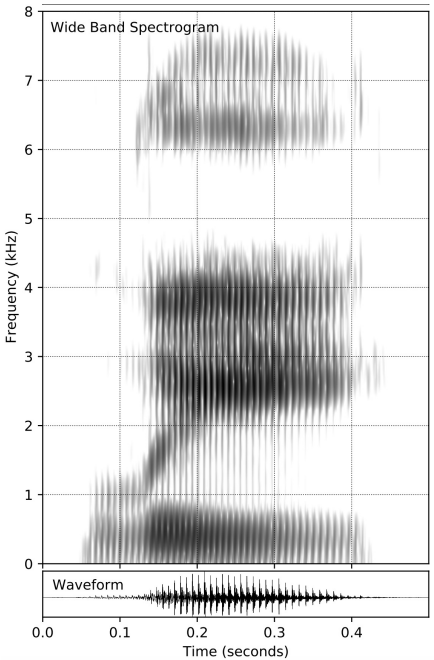
lee
/li^{y}/

Spectral Cues for Semivowels



we

/wi^{y}/



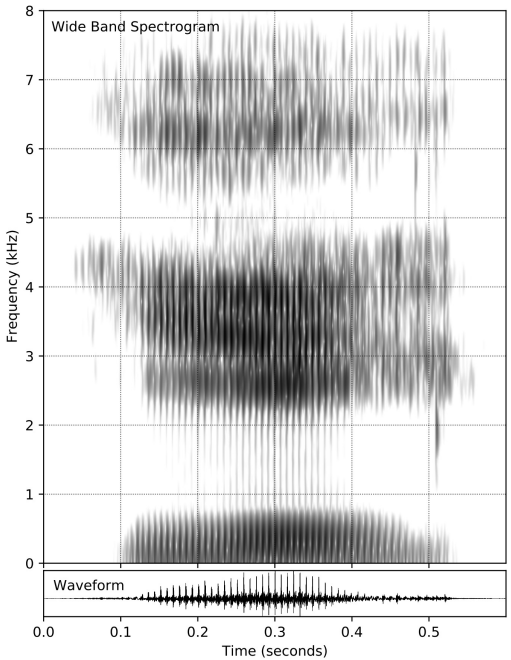
lee

/li^{y}/

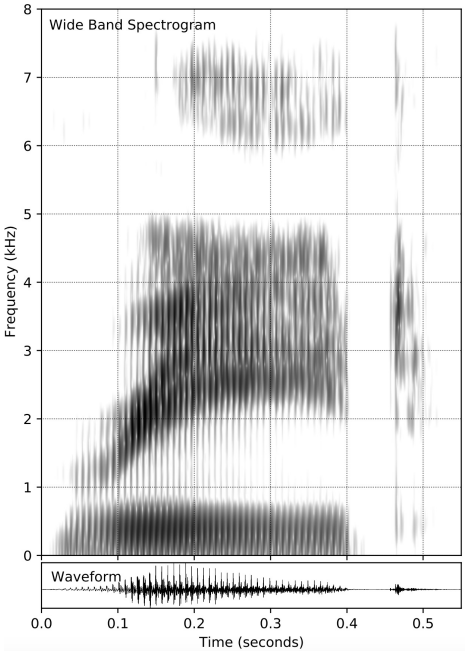
/w/ and /l/ are tricky to distinguish.

Both have very low F1 and F2, but /w/ usually has no energy above F2 whereas /l/ often does

Spectral Cues for Semivowels



ye
/yi^{y}/



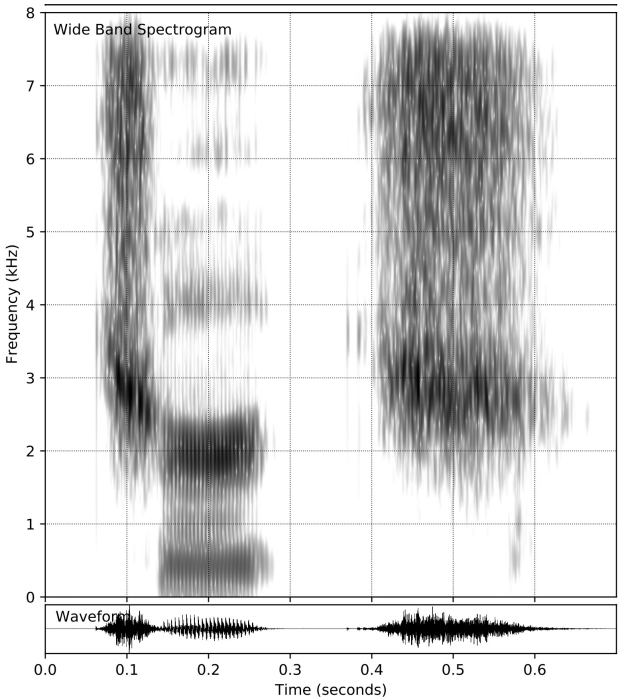
reed
/ri^{y}d/

/y/ looks like an extreme /i^{y}/

/r/ has a very low F3 (below 2 kHz). It looks like an extreme /ɝ/ and is usually shorter in duration

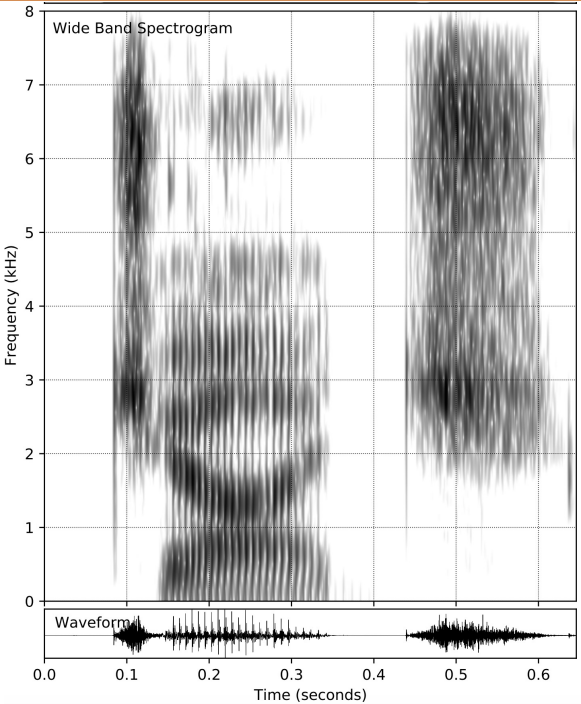
Affricates

- 2 affricates in American English
- Think of them as a stop + fricative combined
- Voiced: /ǰ/ = /d/+ /ž/
- Unvoiced: /č/= /t/+ /š/



church

/čɜ˞č/

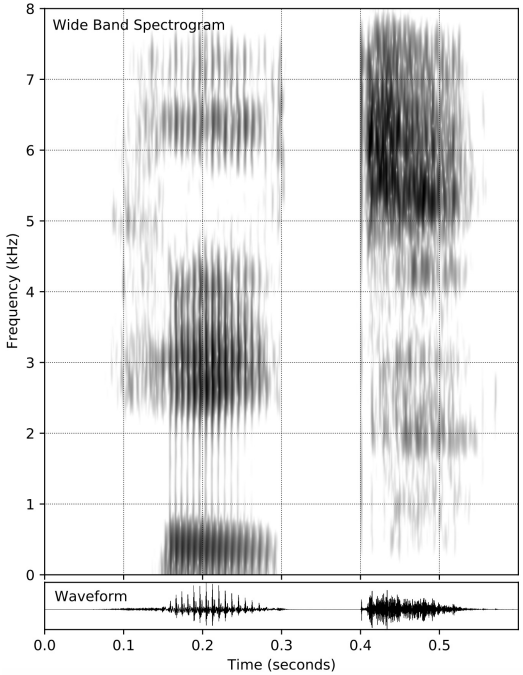


judge

/ǰʌǰ/

Aspirants

- Only 1 aspirant in English: /h/
- No constriction, no voicing; excitation from turbulent airflow from lungs
- Can look similar to a weak fricative like/*f*/, but energy is often concentrated at formants (except not at F1)



heat

/hi^{y}t/

The Consonant Chart

Manner	Labial	Dental	Alveolar	Palatal	Velar
Stop	p b		t d		k g
Fricative	f v	θ ð	s z	š ž	
Nasal	m		n		ŋ

For boxes with pairs: voiced unvoiced

w is extreme u
y is extreme i^y
l is extreme o
r is extreme ɜ˞

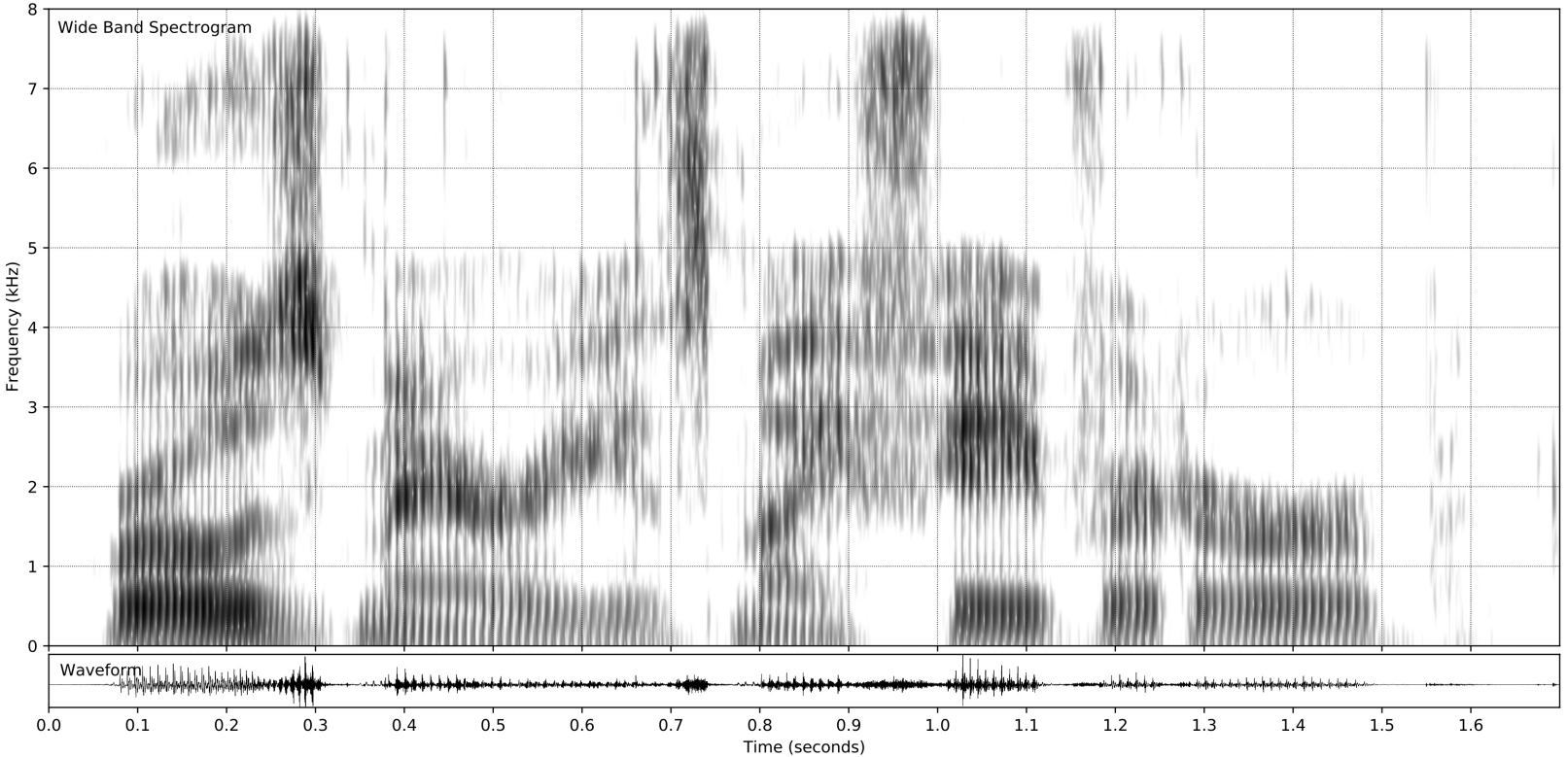
ɾ (flap: reduced stop)
ʔ (glottal stop)

h (aspirant)

ǰ = d + ž
č= t + š

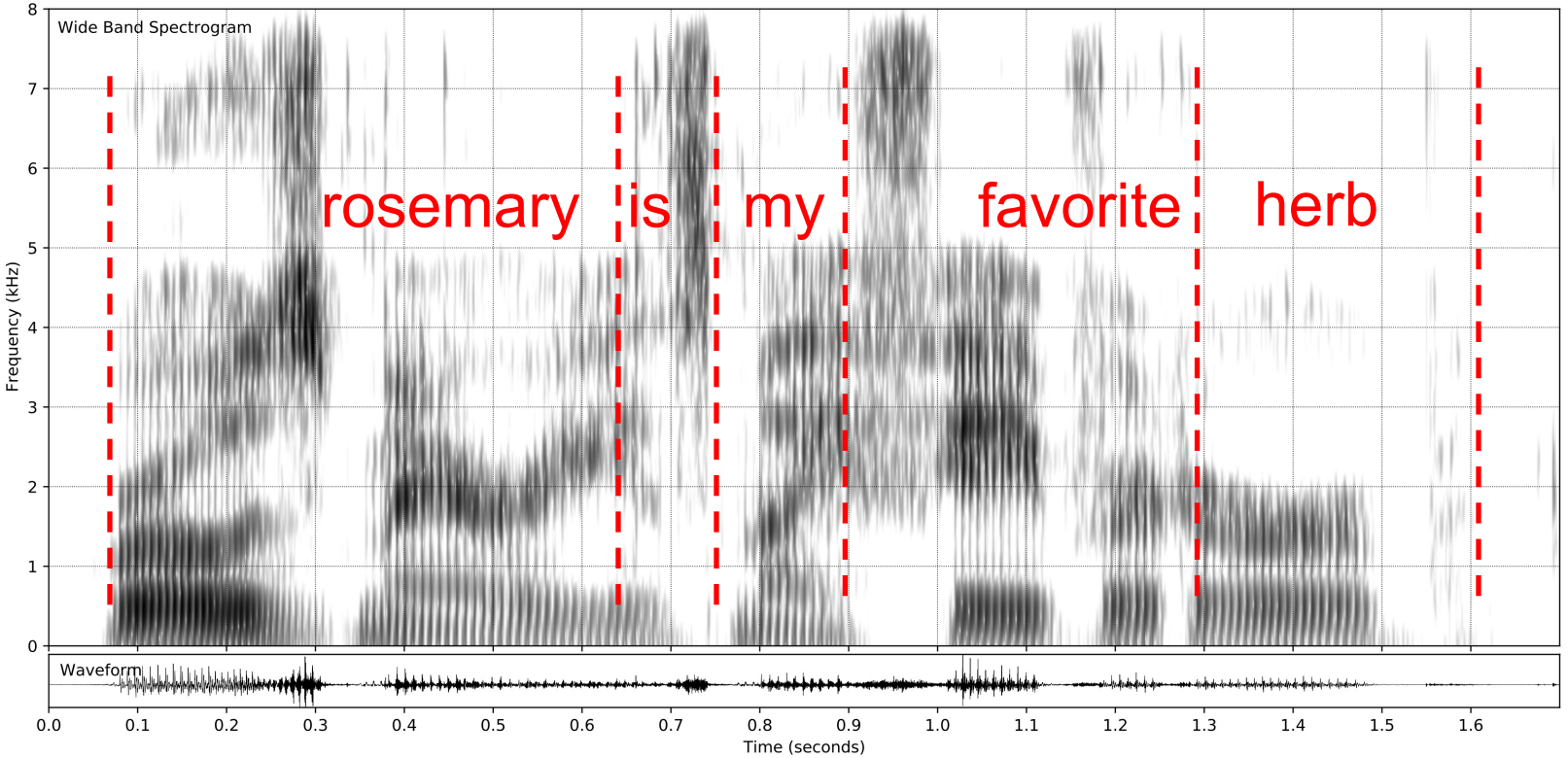
Speech doesn’t have “whitespace”

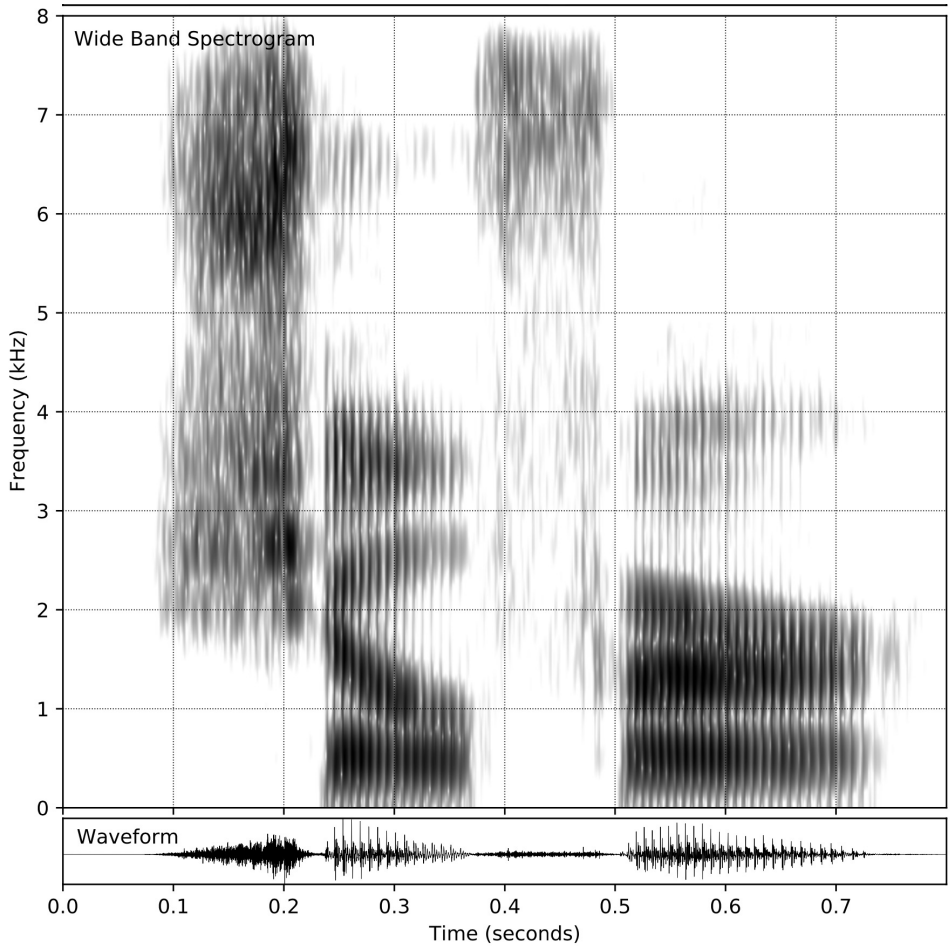
Where do the words begin and end in this utterance?



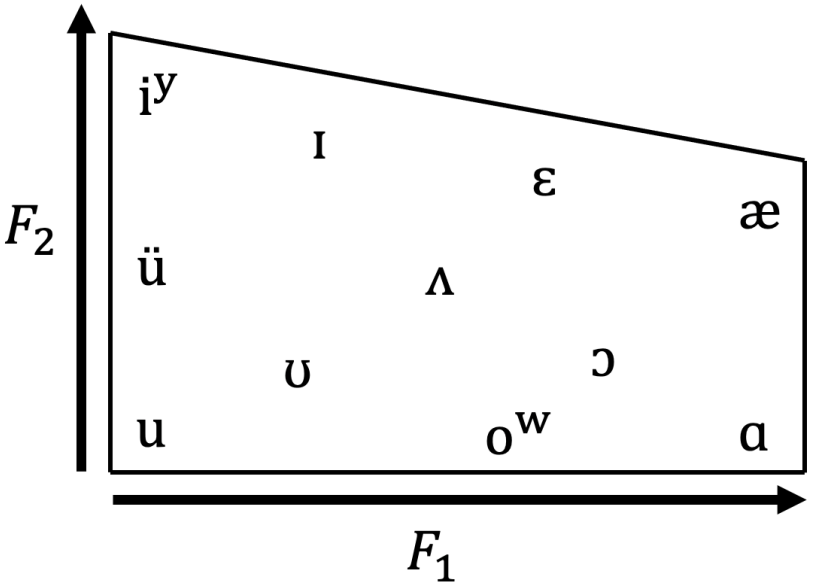
Speech doesn’t have “whitespace”

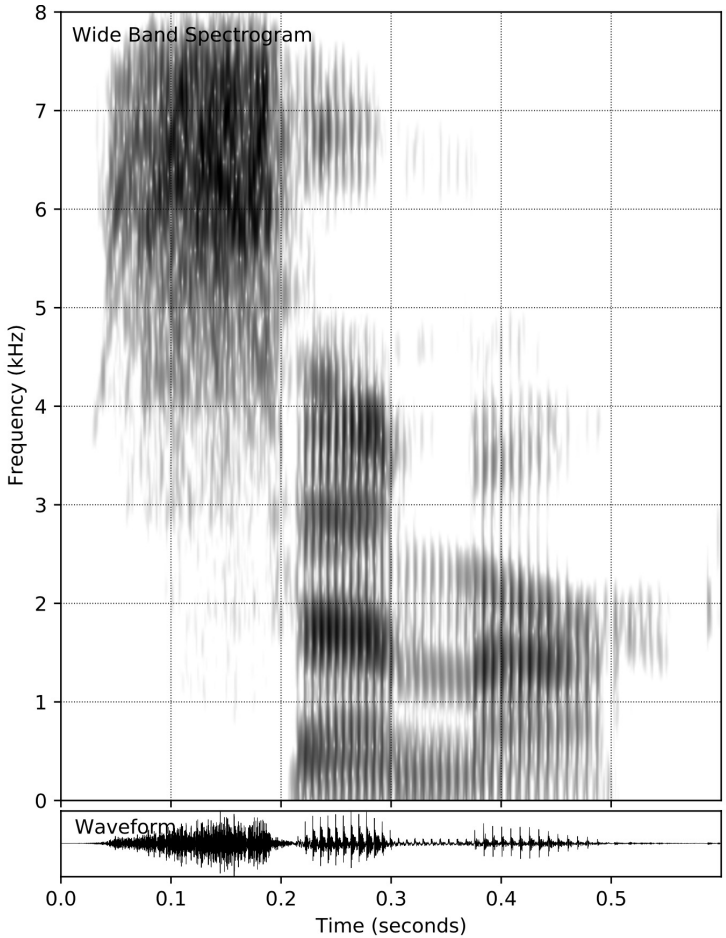
Where do the words begin and end in this utterance?



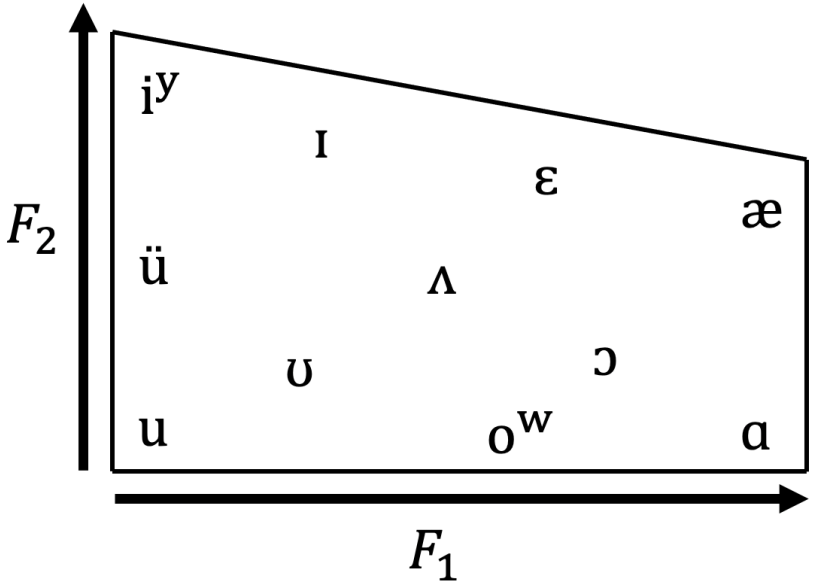


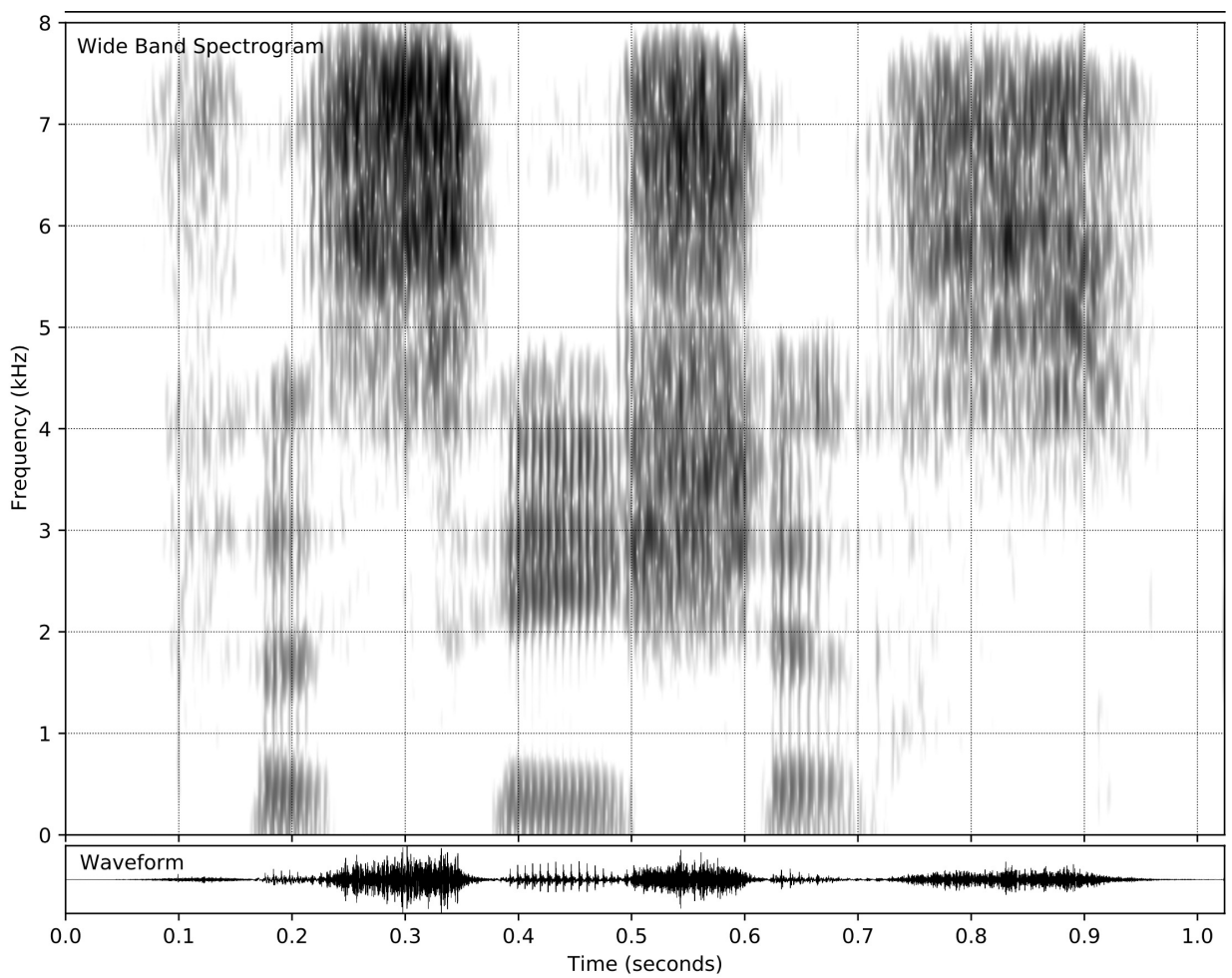
(Mystery word 1)



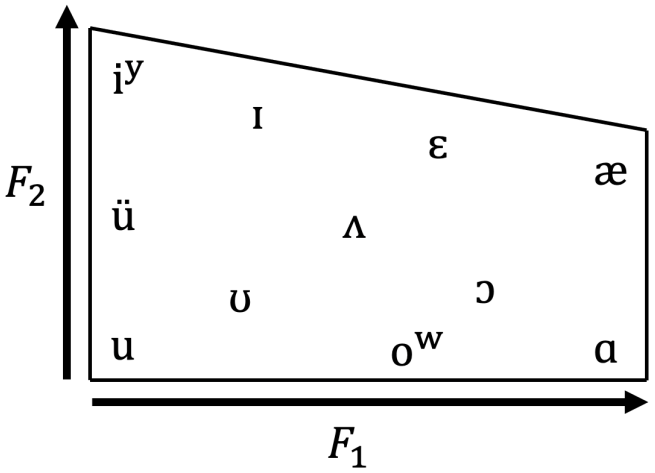


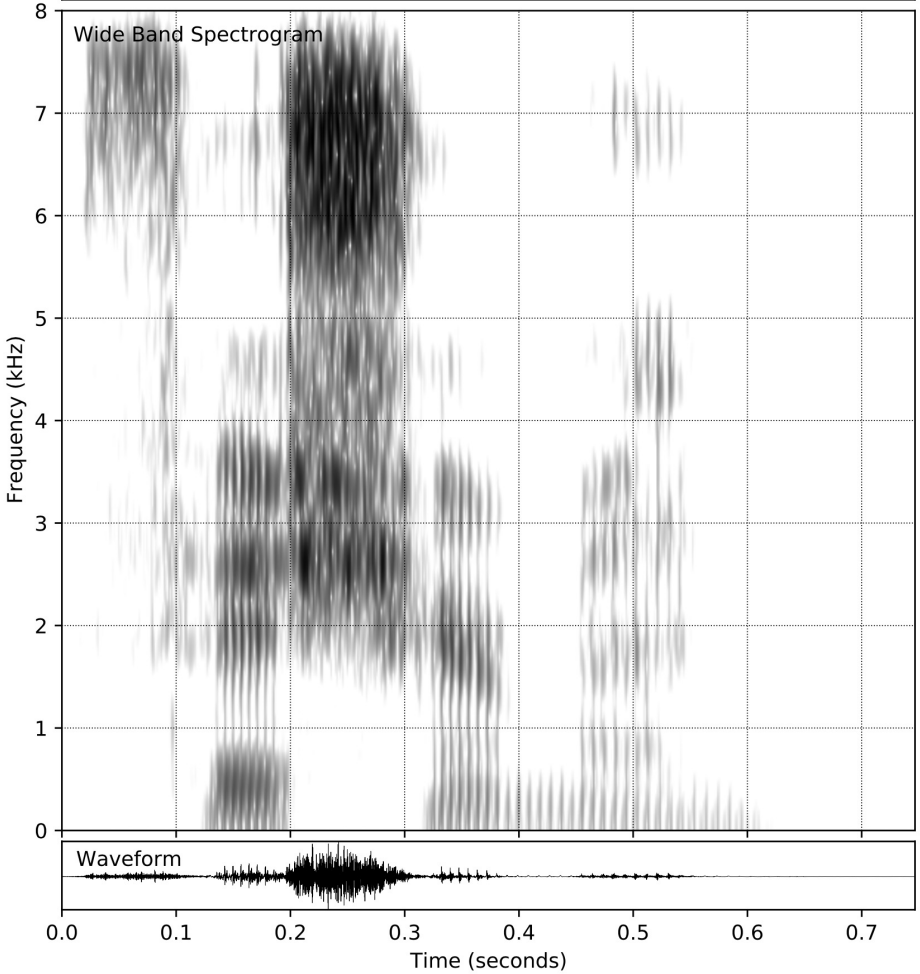
(Mystery word 2)



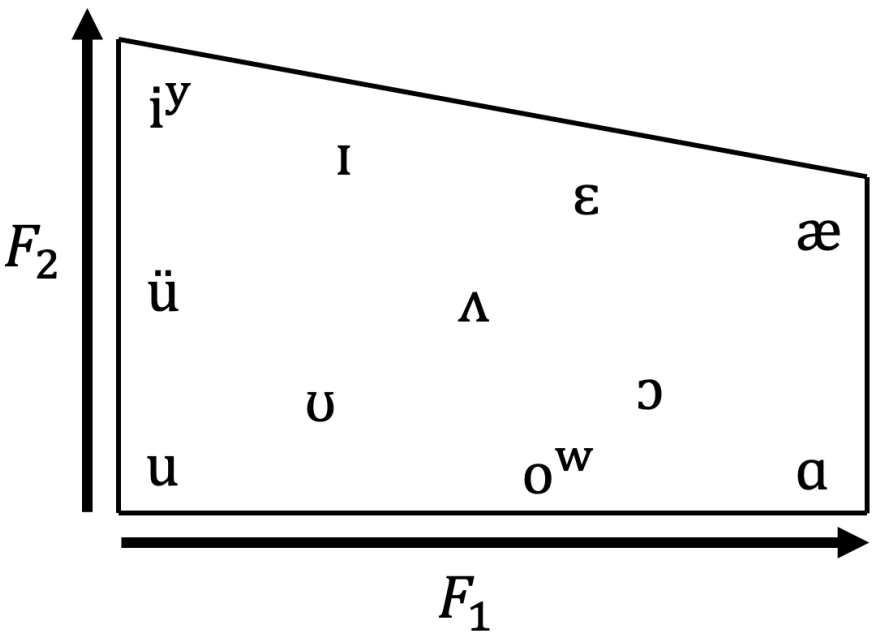


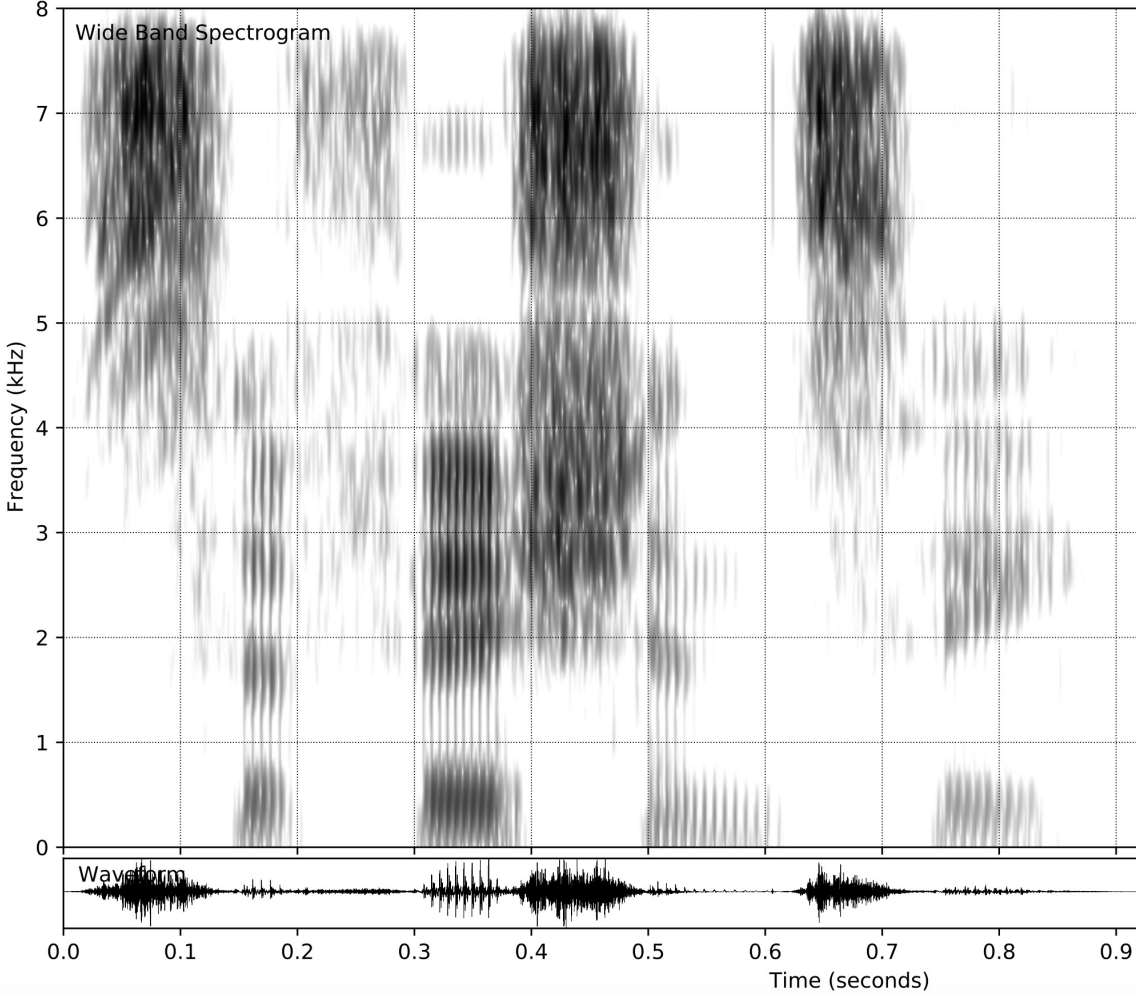
(Mystery word 3)



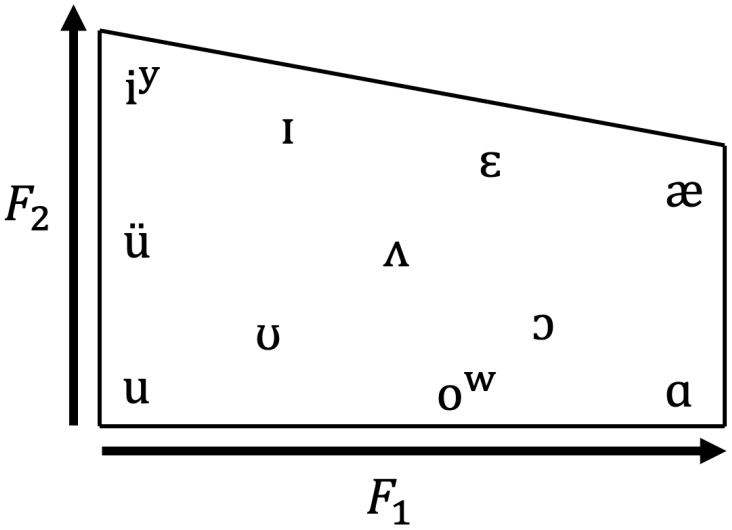


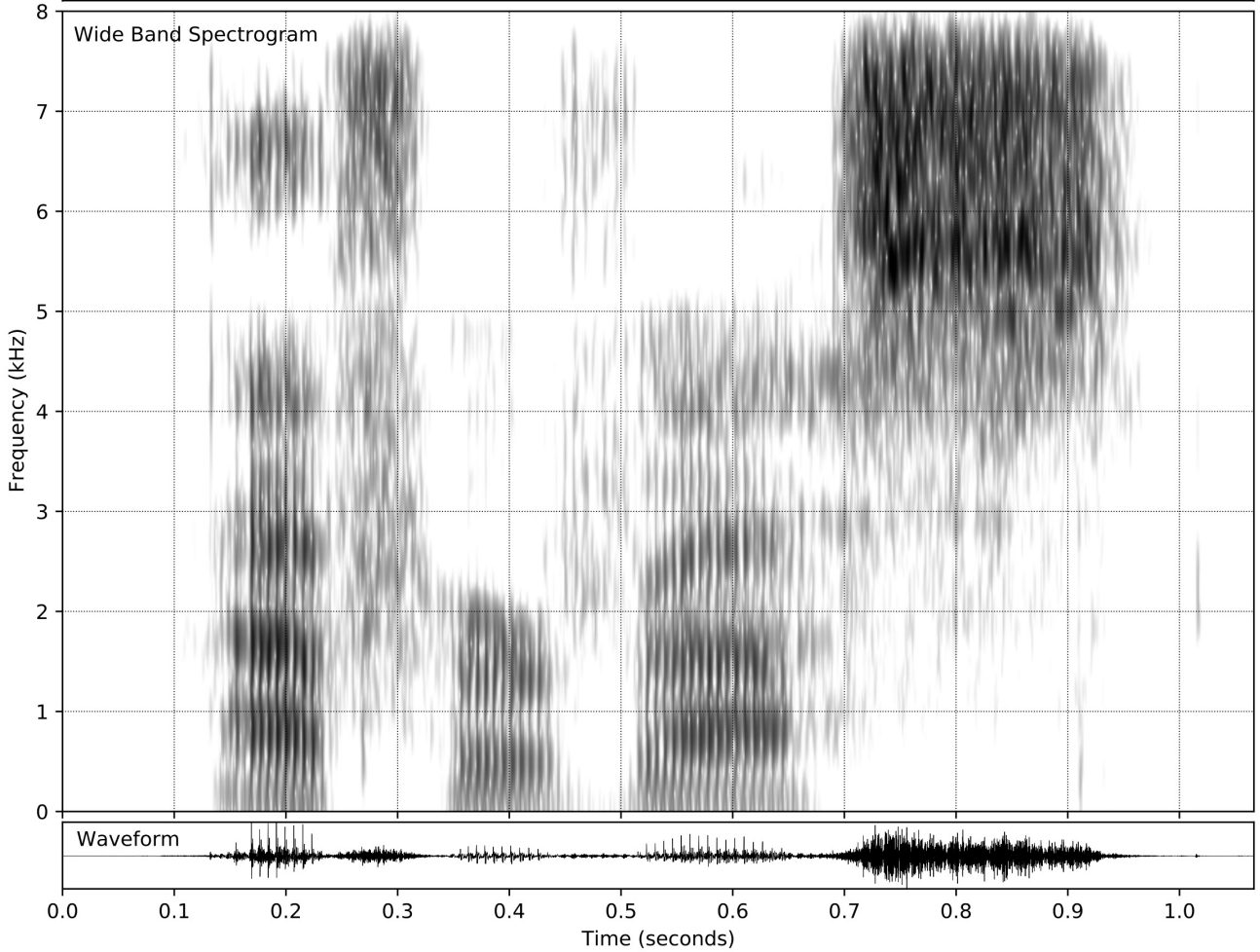
(Mystery word 4)



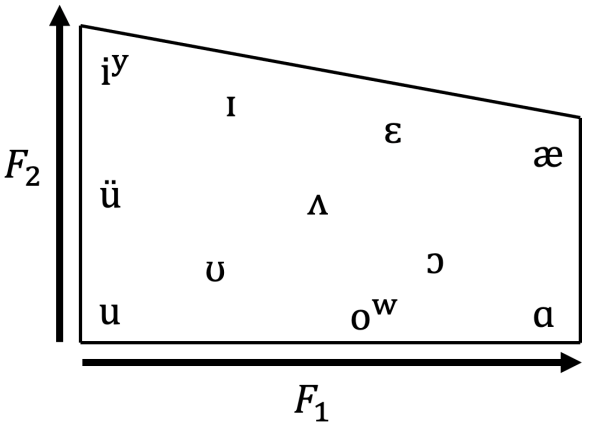


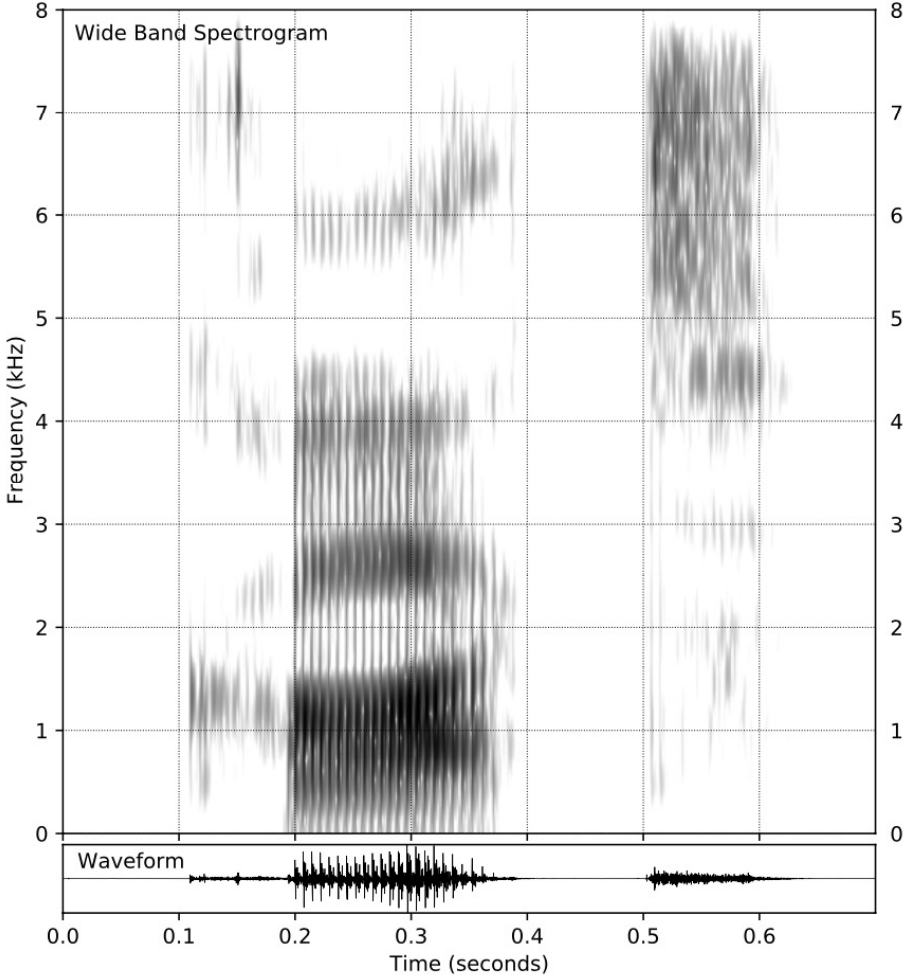
(Mystery word 5)



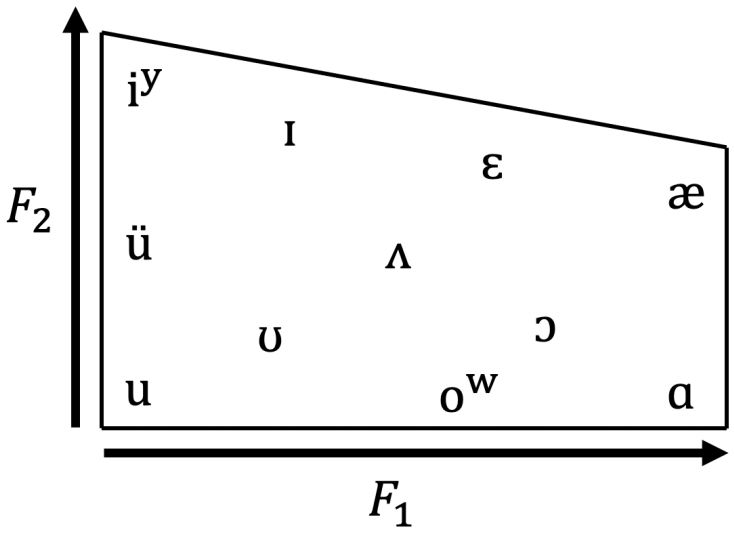


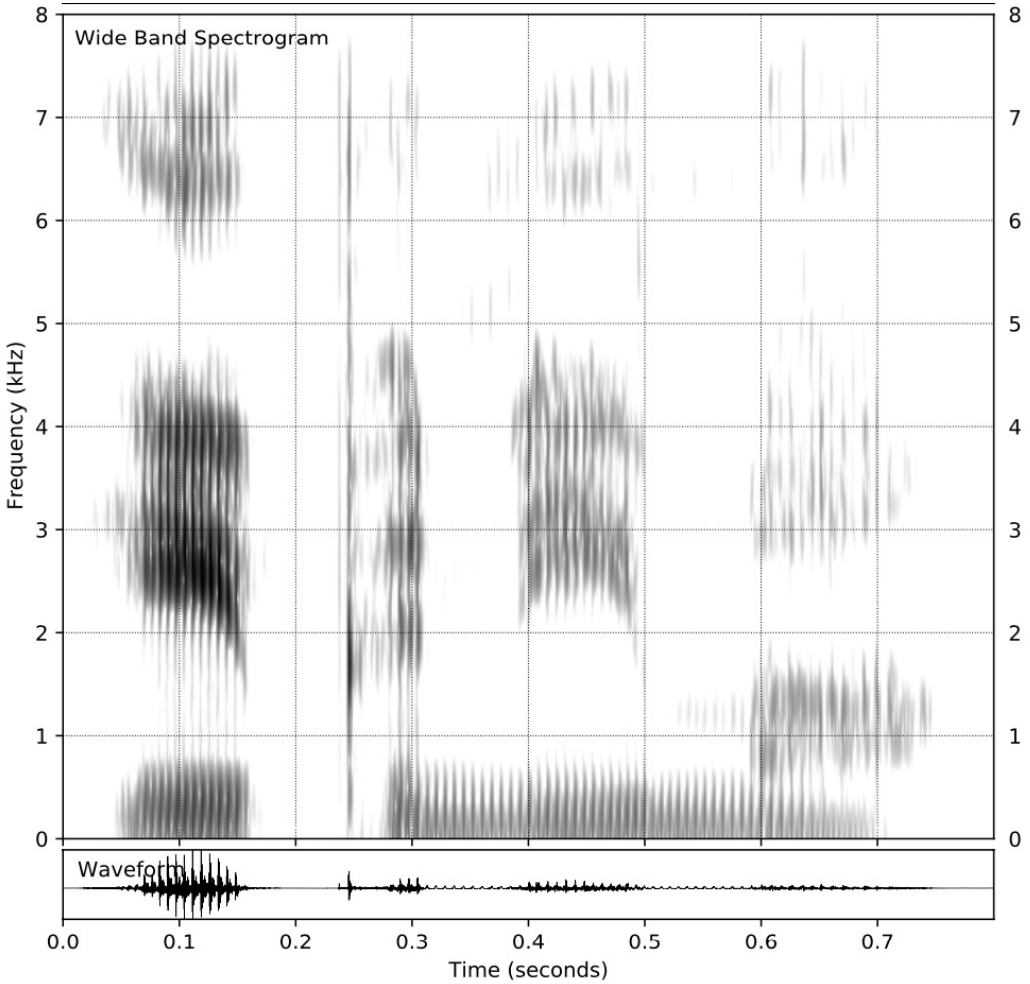
(Mystery word 6)



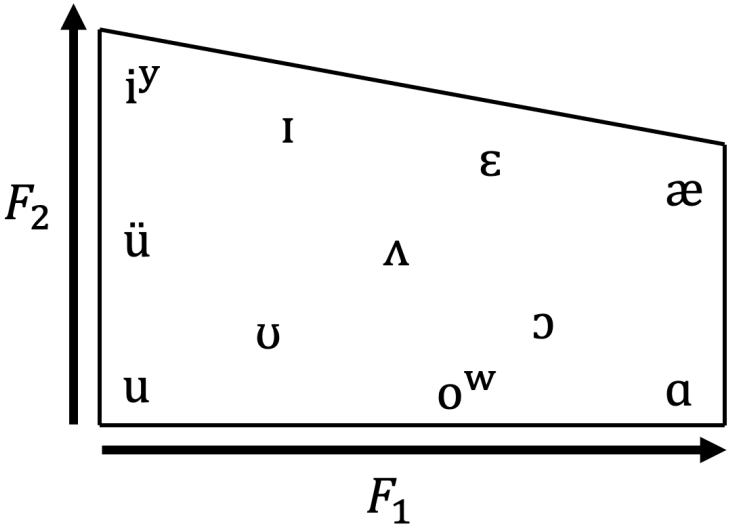


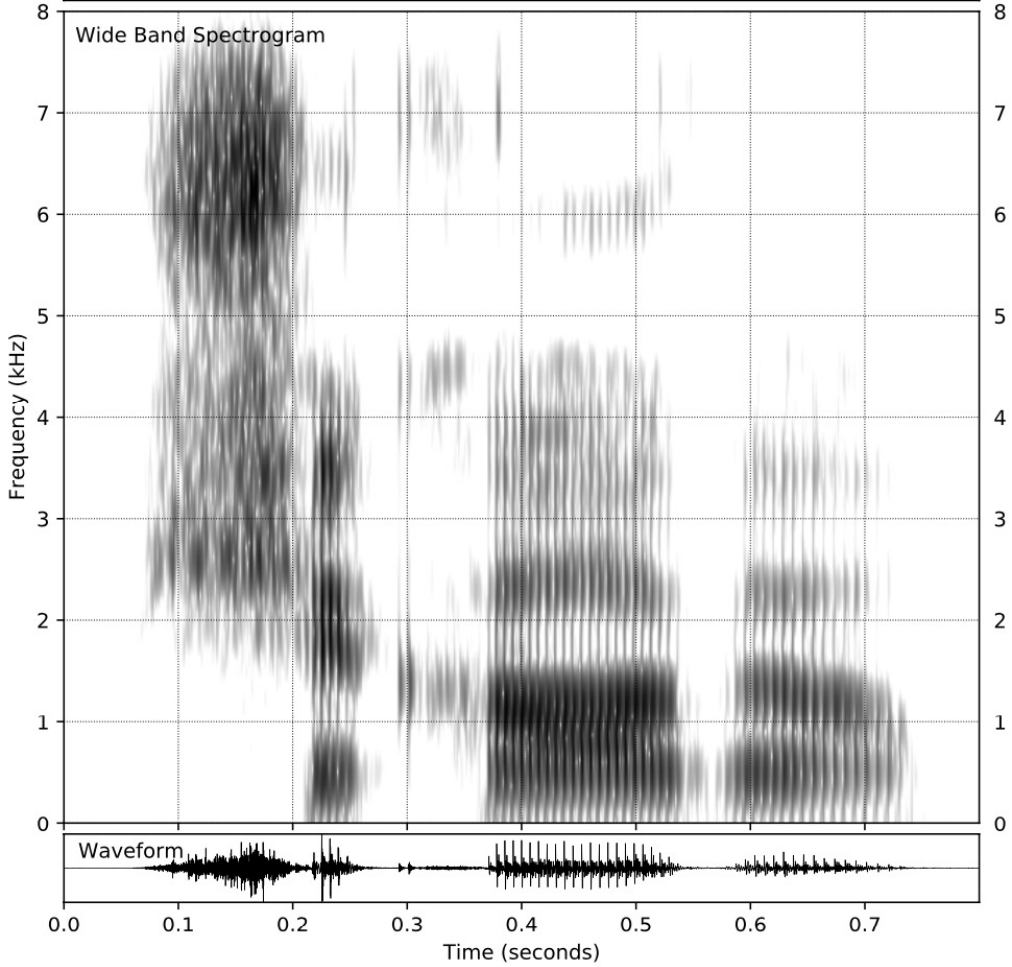
(Mystery word 7)





(Mystery word 8)





(Mystery word 9)