Speech

CSC401/2511 – Natural Language Computing – Spring 2026
Lecture 10
University of Toronto

This lecture

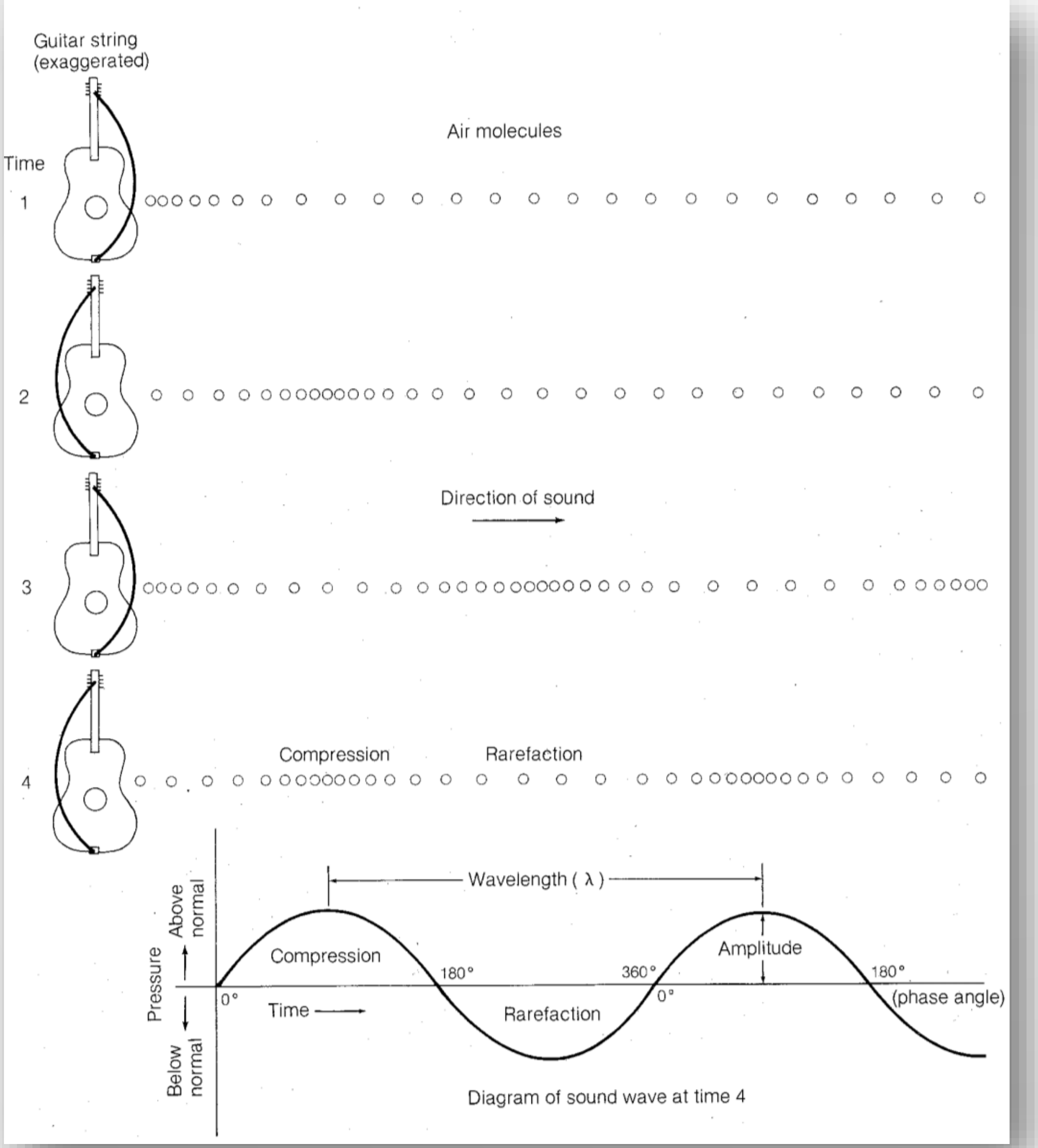
- Speech signals
- Articulatory phonetics

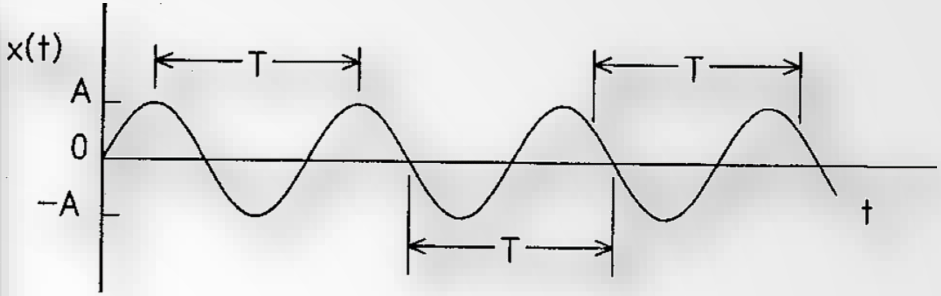
- Some images from Gray's Anatomy, Jim Glass' course 6.345 (MIT), the Jurafsky & Martin textbook, Encyclopedia Britannica, the Rolling Stones, the Pink Floyds.

What is sound?

- **Sound** is a time-variant pressure wave created by a vibration.
  - Air particles **hit** each other, setting others in motion.
    - High pressure ≡ **compressions** in the air (C).
    - Low pressure ≡ **rarefactions** within the air (R).

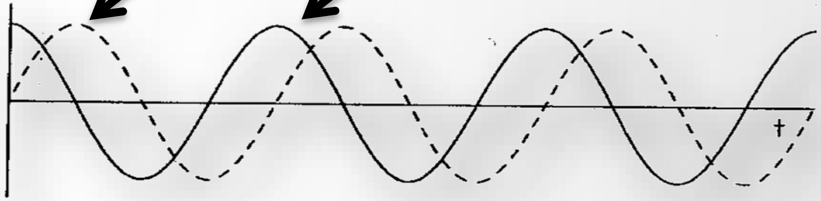
What is sound?





Frequency $F = 1/T$

sine cosine



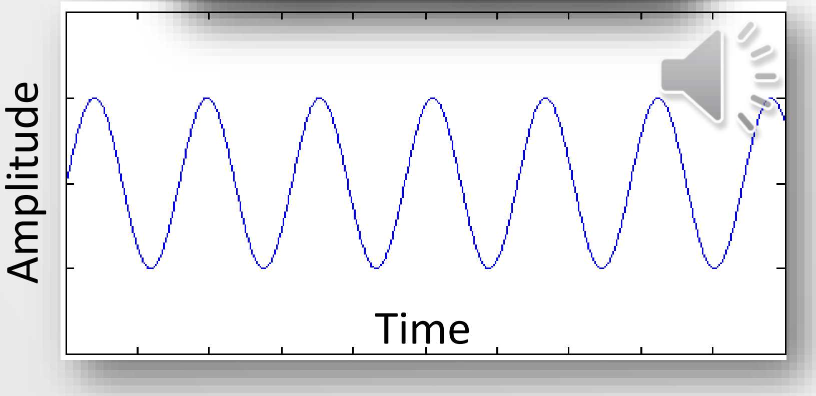
phase ϕ is displacement of a signal in time. E.g., with $\phi = \pi/2$,

$$\sin(x + \phi) = \cos(x)$$

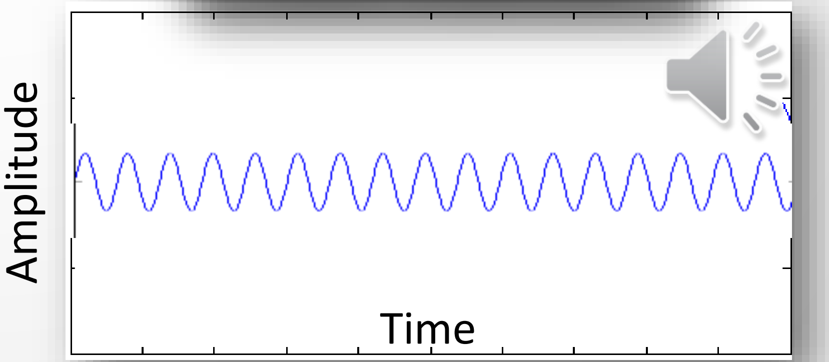
What is sound?

- A single **tone** is a sinusoidal function of pressure and time.
  - **Amplitude**: *n.* The degree of the displacement in the air. This is similar to 'loudness'.
  - **Frequency**: *n*. The number of cycles within a unit of time. e.g., **1 Hertz (Hz) = 1 oscillation/second**

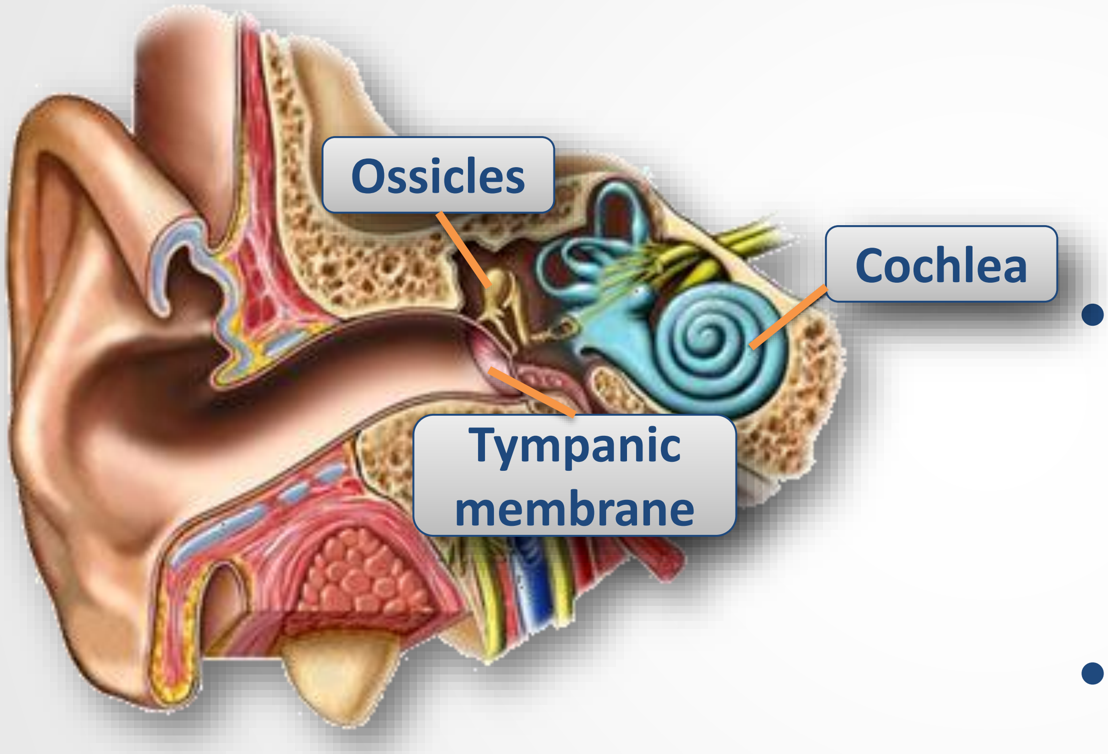
Lower frequency, higher amplitude



Higher frequency, lower amplitude



The inner ear

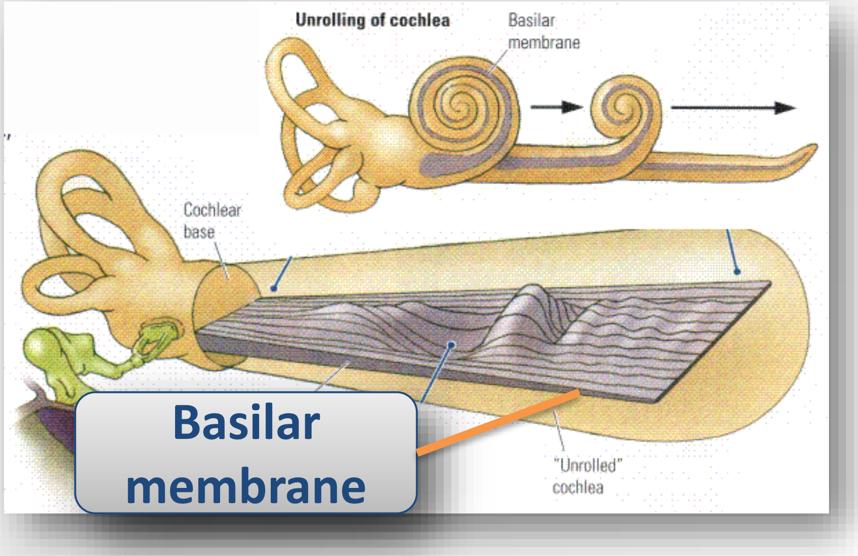


- Time-variant waves enter the ear, vibrating the **tympanic membrane**.

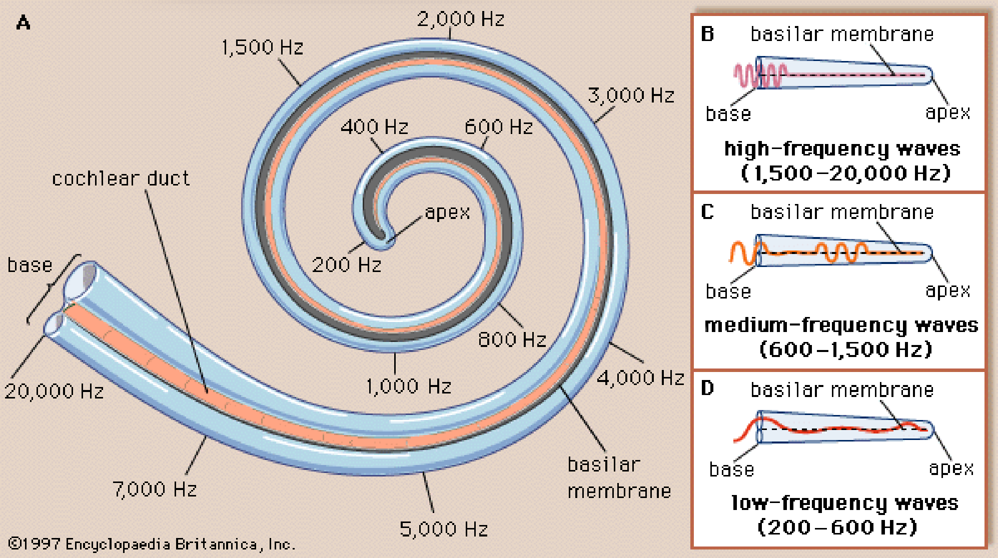
- This membrane causes tiny bones (the **ossicles**) to vibrate.

- These bones in turn vibrate a structure within a shell-shaped bony structure called the **cochlea**.

The cochlea and basilar membrane



Basilar membrane



©1997 Encyclopaedia Britannica, Inc.

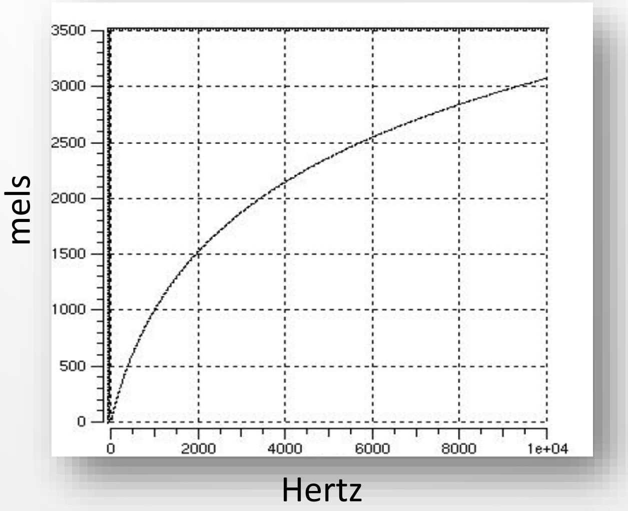
- The **basilar membrane** is covered with tiny hair-like nerves – some near the **base**, some near the **apex**.
- **High** frequencies are picked up near the base, **low** frequencies near the apex.
- These nerves fire when activated, and communicate to the brain.

The Mel-scale

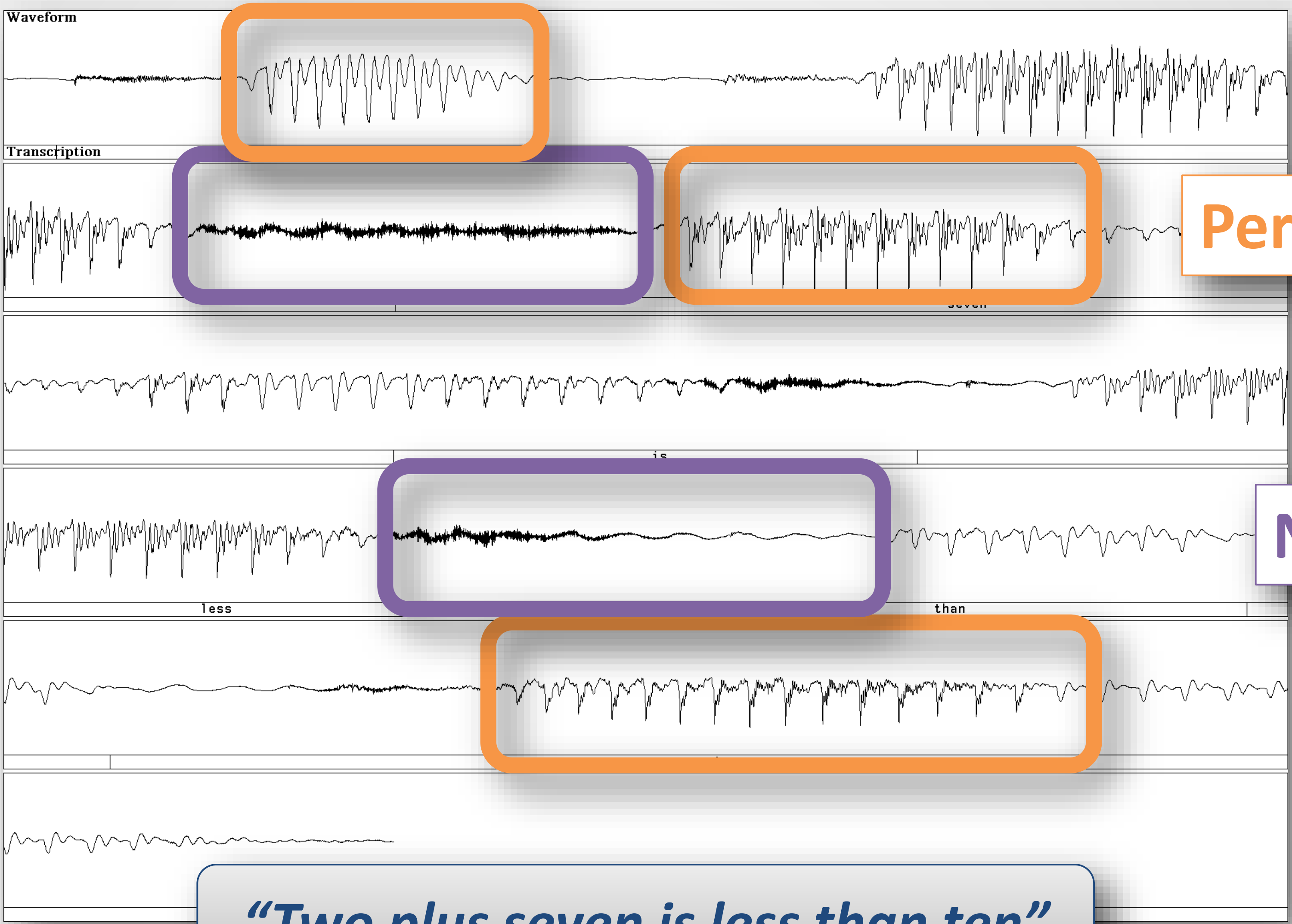
- Human hearing is **not** equally sensitive to **all** frequencies.
  - We are **less** sensitive to frequencies > 1 kHz.

- A **mel** is a unit of pitch. Pairs of sounds which are **perceptually** equidistant in pitch are separated by an equal number of **mels**.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$



Speech waveforms



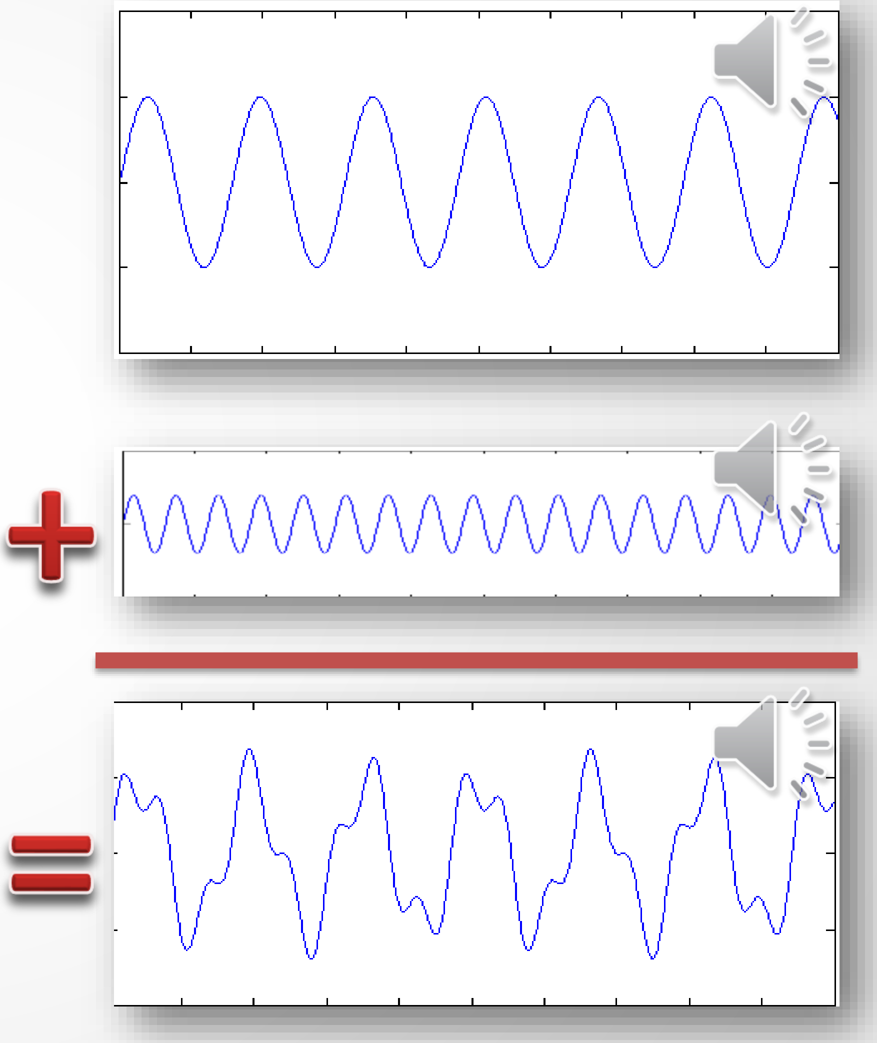
Periodic

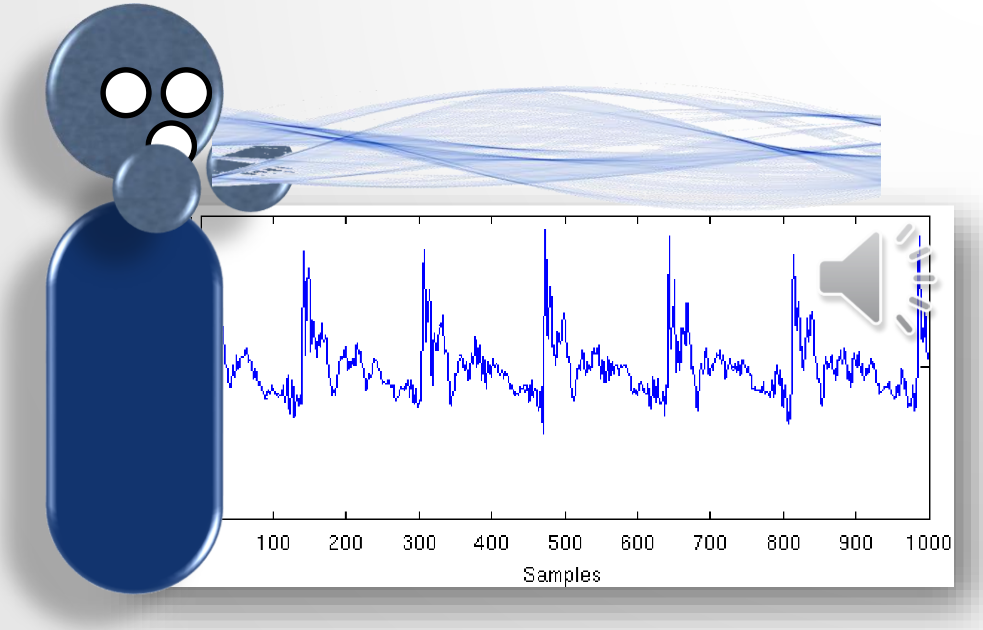
Noisy

"Two plus seven is less than ten"

Superposition of sinusoids

- **Superposition**: *n.* the adding of sinusoids together.

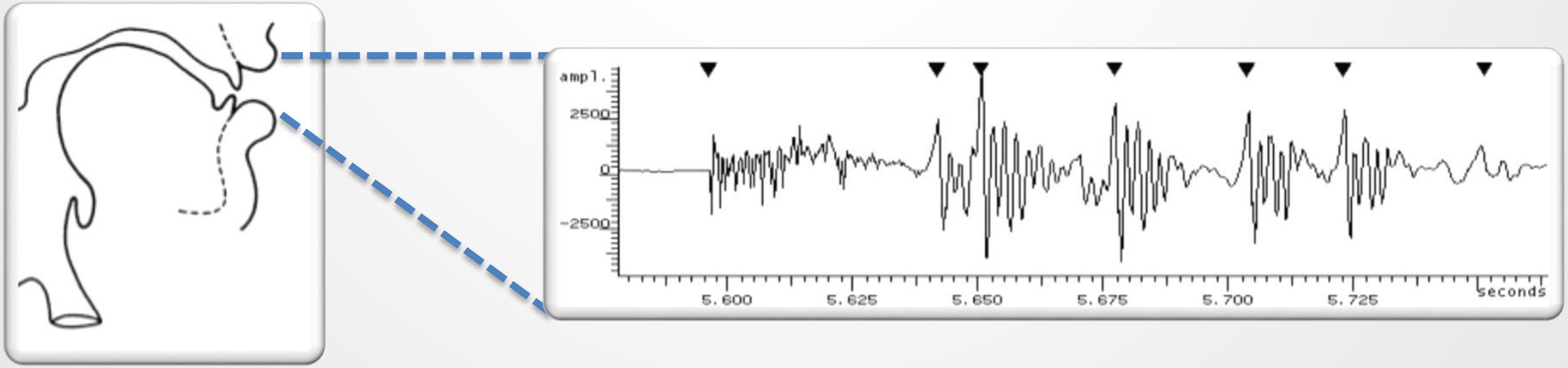




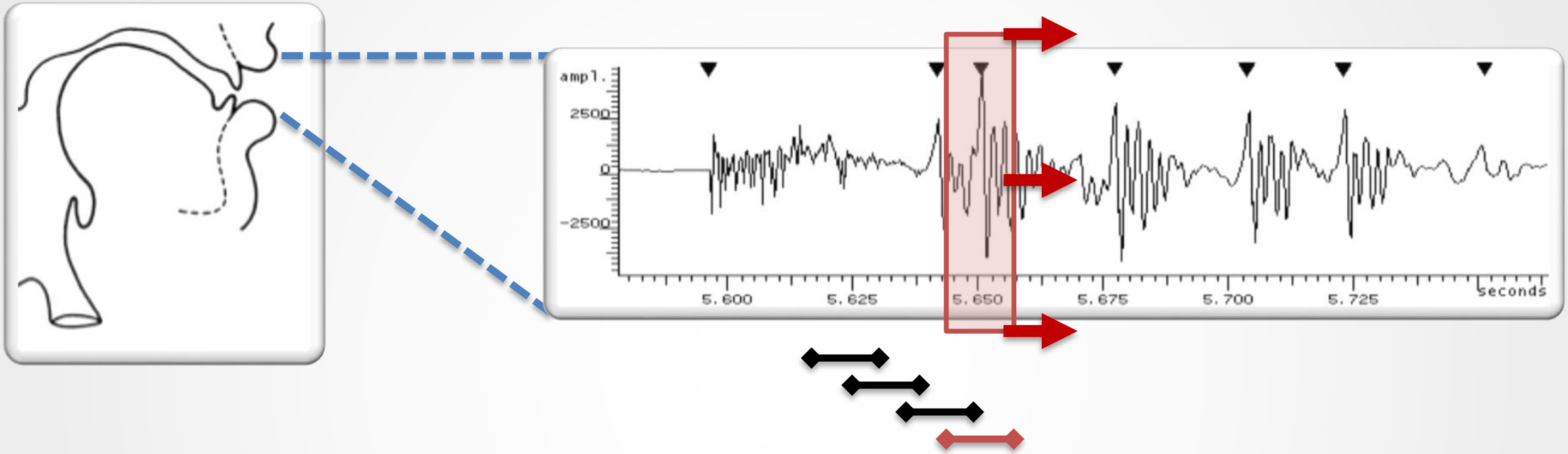
Extracting sinusoids from waveforms

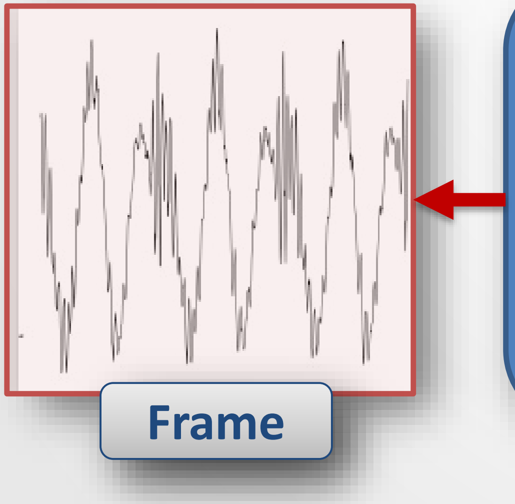
- As we will soon see, the relative **amplitudes** and **frequencies** of the sinusoids that combine in speech are **highly correlated** with the **speech units** being uttered.
    - ∴ If we could **separate** the waveform into its component sinusoids, it would help us **classify** the speech being uttered.
    - *But the shape of the signal changes over time*

      (it’s not a single repeating pattern)...



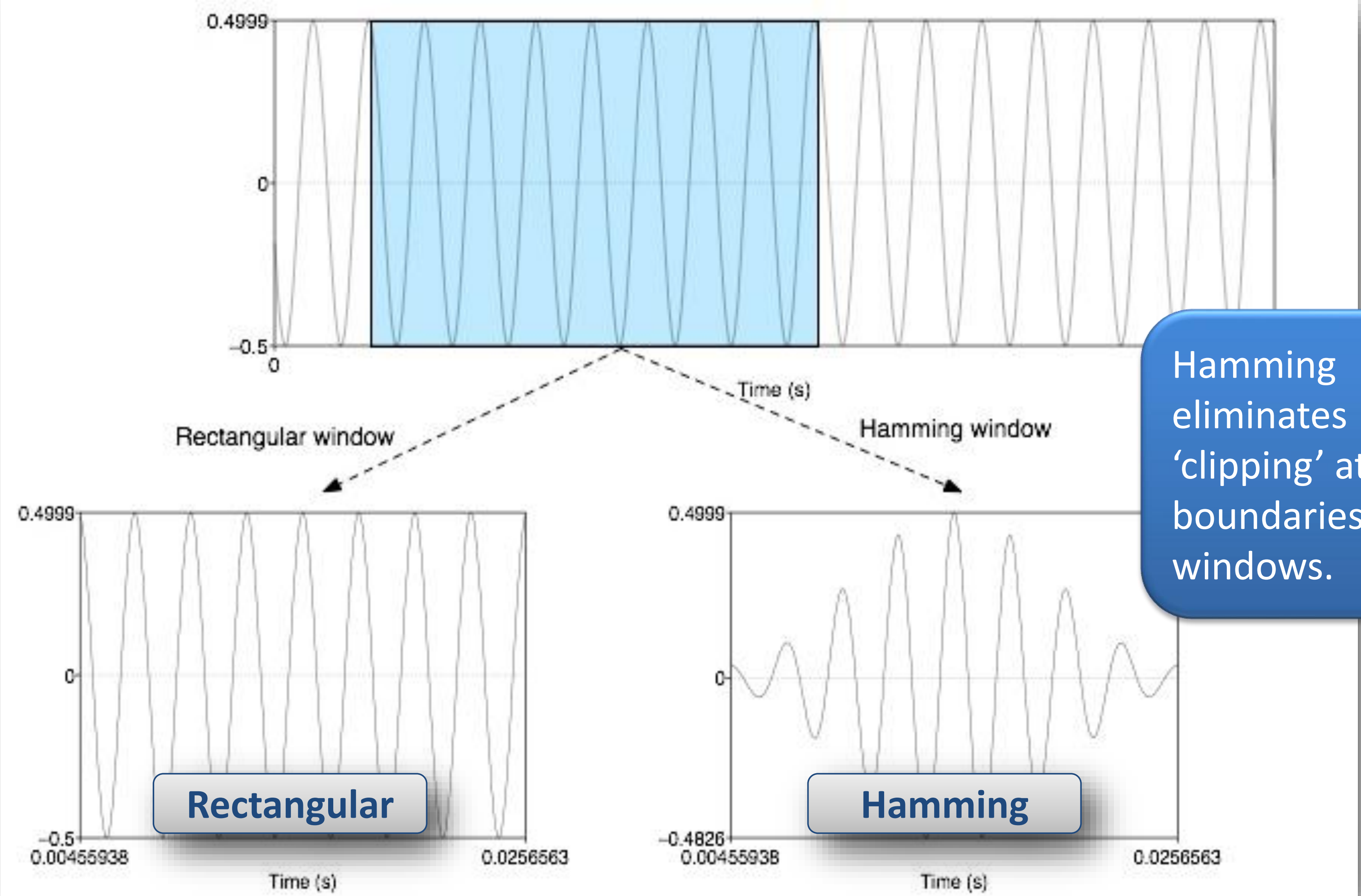
Short-time windowing





- Speech waveforms change drastically over time.
- We *move* a short analysis window (assumed to be time-invariant) across the waveform in time.
  - E.g. frame shift: 10-30 ms
  - E.g. frame length: 25-40 ms

Window types



Rectangular

Hamming

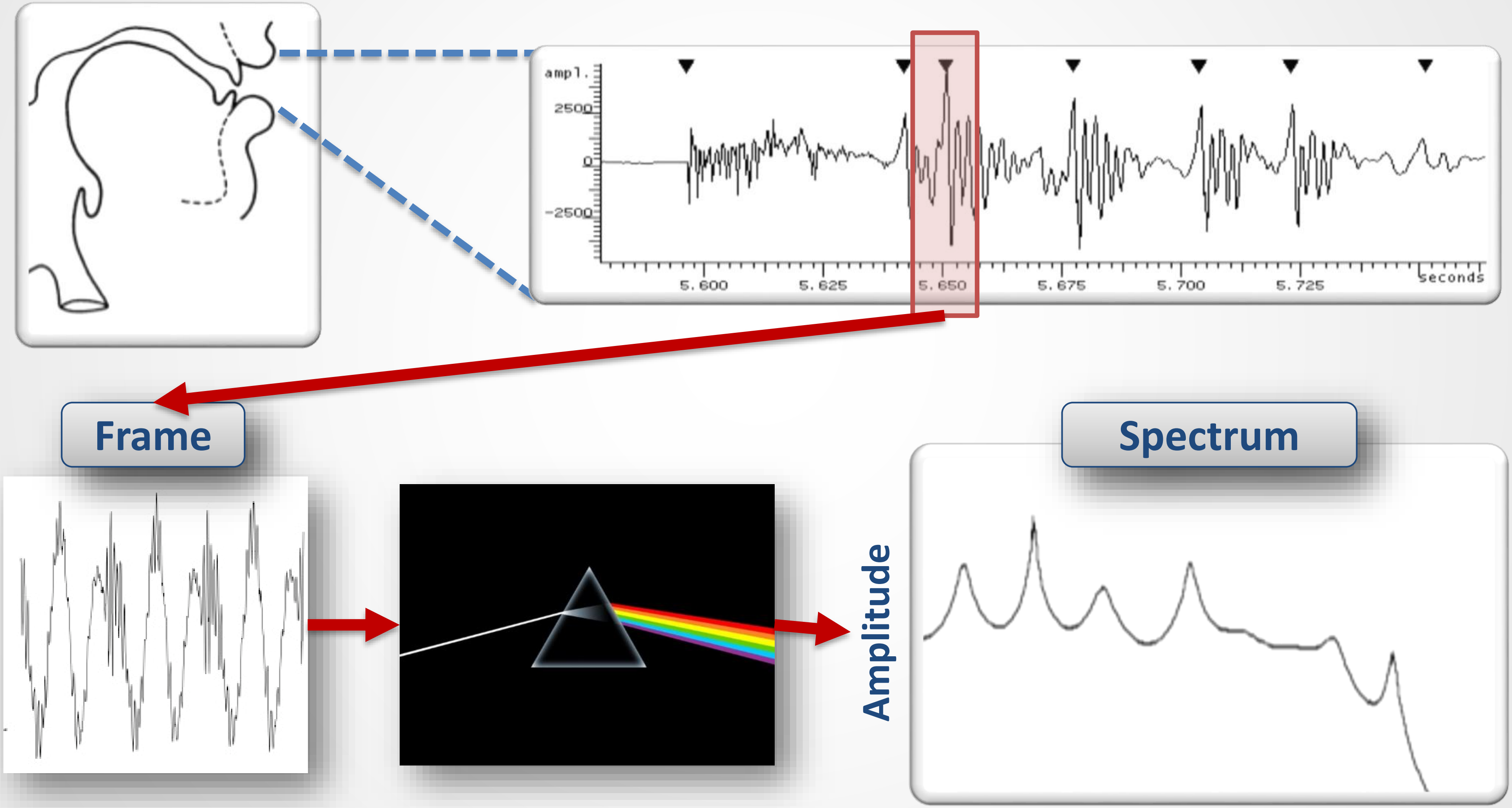
Hamming eliminates ‘clipping’ at the boundaries of windows.

Extracting a spectrum

White light

Any Colour
You Like
(track 8)

Extracting a spectrum in a window



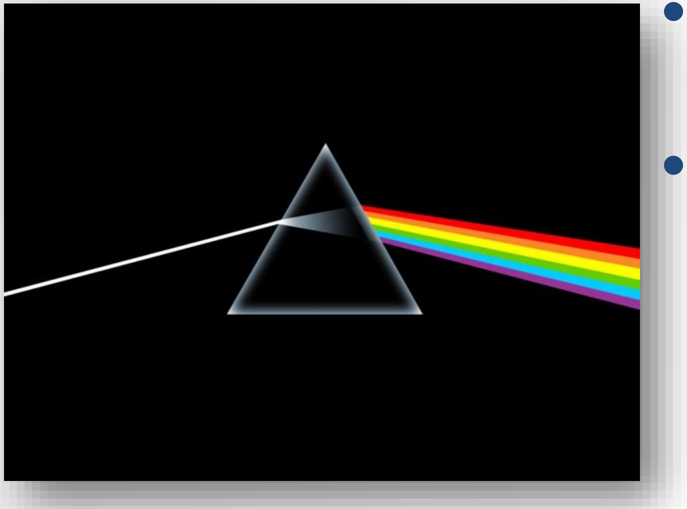
Frame

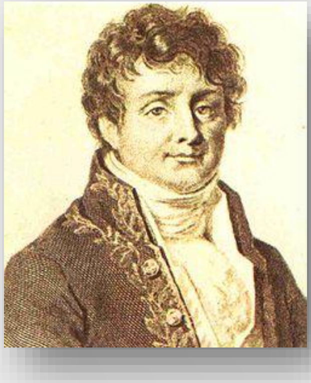
Spectrum

Amplitude

Frequency (Hz)

The continuous Fourier transform





Fun fact: Fourier instructed Champollion.

- **Input**: Continuous signal $x(t)$.
- **Output**: Spectrum $X(F)$

$$X(F) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi Ft}\,dt$$

- It's **invertible**, i.e., $x(t) = \int_{-\infty}^{\infty} X(F)e^{i2\pi Ft}\,dF$.
- It's **linear**, i.e., for $a, b \in \mathbb{C}$,

  if $h(t) = ax(t) + by(t)$,

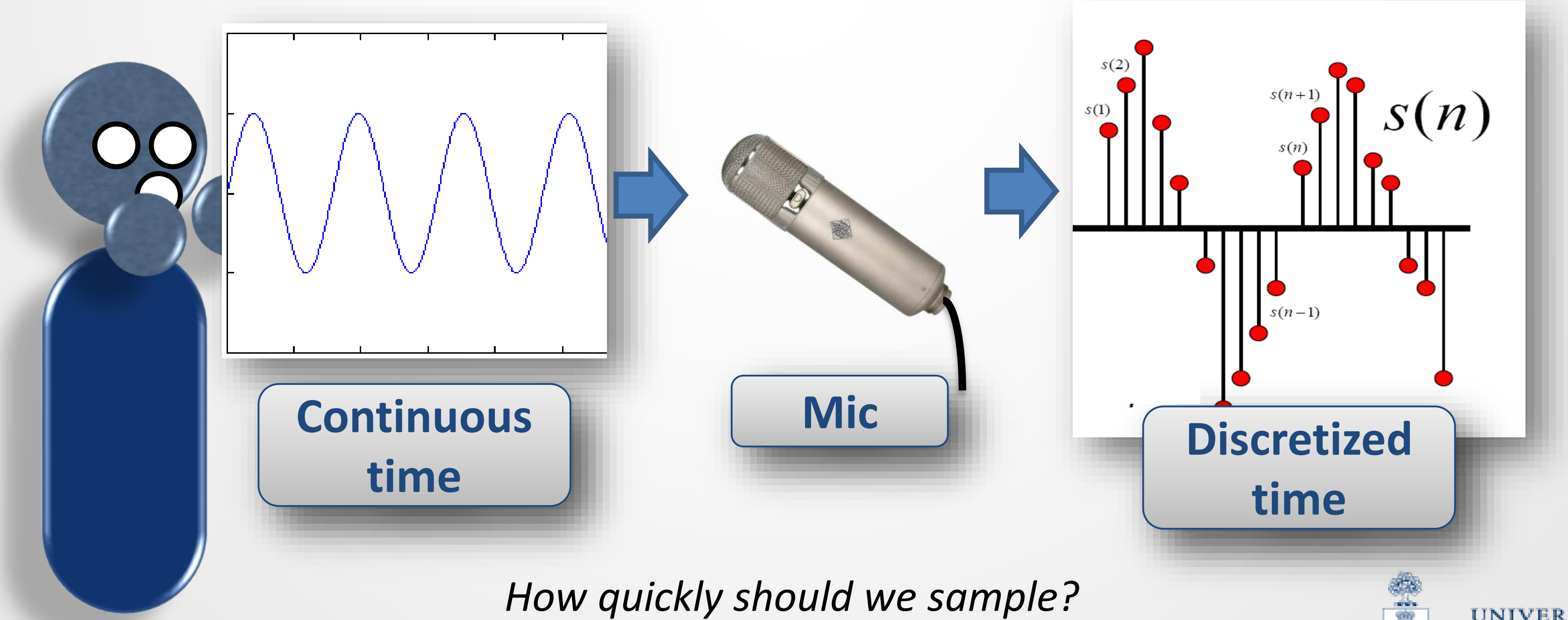
  then $H(F) = aX(F) + bY(F)$

…

It needs **continuous** input $x(t)$…

Discrete signal representation

- **Sampling**: *vbg.* measuring the amplitude of a signal at regular intervals.
  - e.g., 44.1 kHz (*CD*), 8 kHz (*telephone*).
  - These amplitudes are initially measured as **continuous** values at **discrete** time steps.



Continuous time

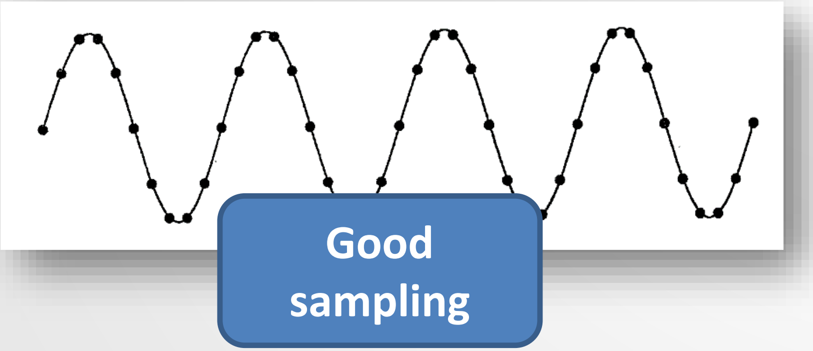
Mic

Discretized time

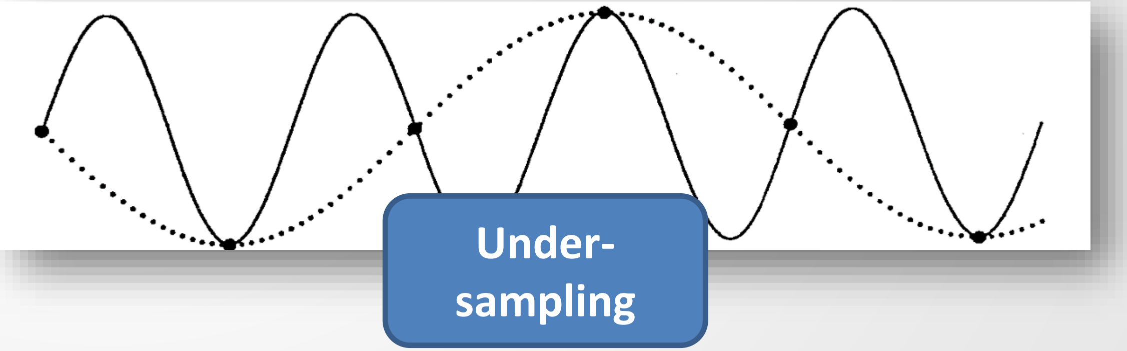
How quickly should we sample?

Discrete signal representation

- **Nyquist rate**: *n.* the **minimum** sampling rate necessary to preserve a signal's **maximum** frequency.
  - i.e., **twice** the maximum frequency, since we need $\geq$ 2 samples/cycle.
  - Human speech is very informative $\leq$ 4 kHz, ∴ At least 8 kHz sampling (16kHz the norm)

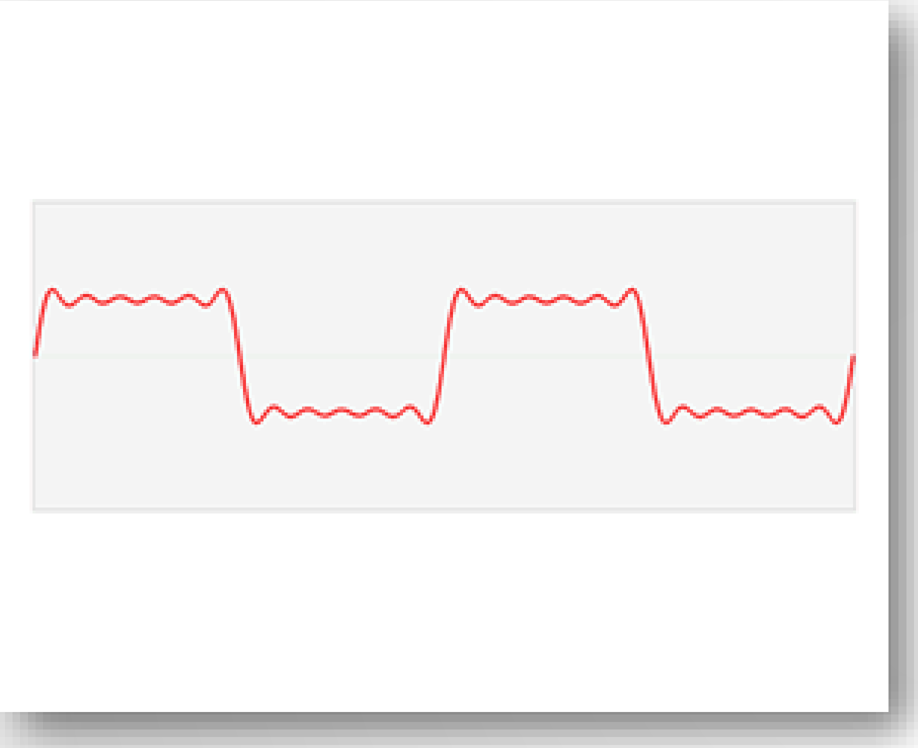


Good sampling



Under-sampling

Discrete Fourier transform (DFT)



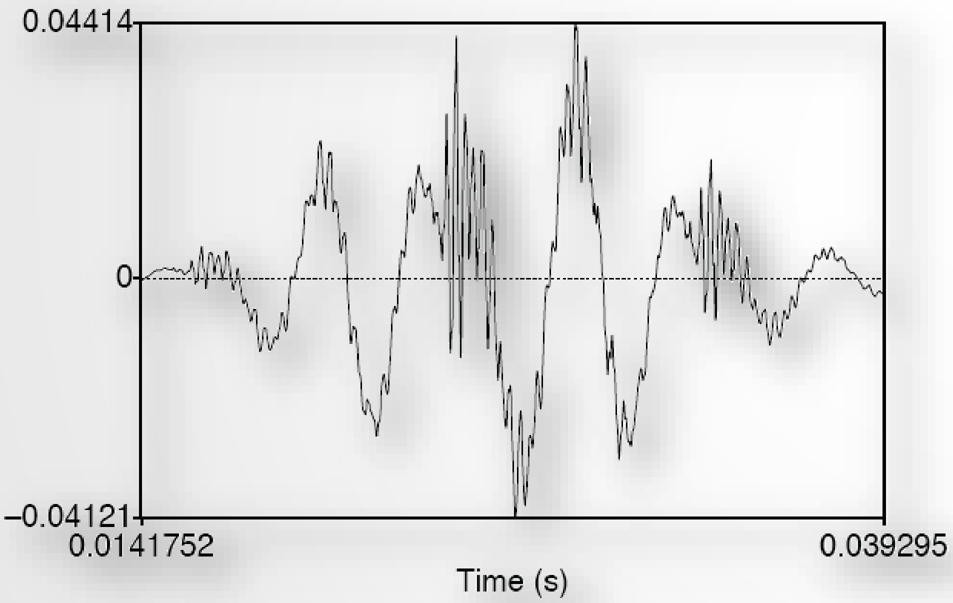
- **Input**: Windowed signal $x[0] \ldots x[N-1]$.
- **Output**: N complex numbers $X[k]$ $(k \in \mathbb{Z})$

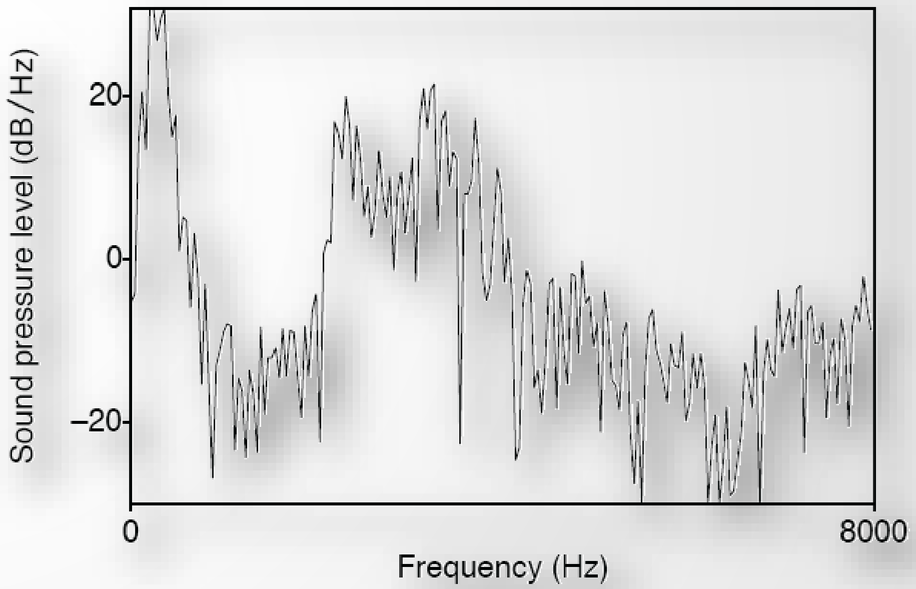
$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i2\pi k \frac{n}{N}}$$

- **Algorithm(s)**: the Fast Fourier Transform (FFT) with complexity $O(N \log N)$.

Discrete Fourier transform (DFT)

- Below is a 25 ms Hamming-windowed signal from /iy/ as in ‘bull sh***ee***p’, and its spectrum as computed by the DFT.



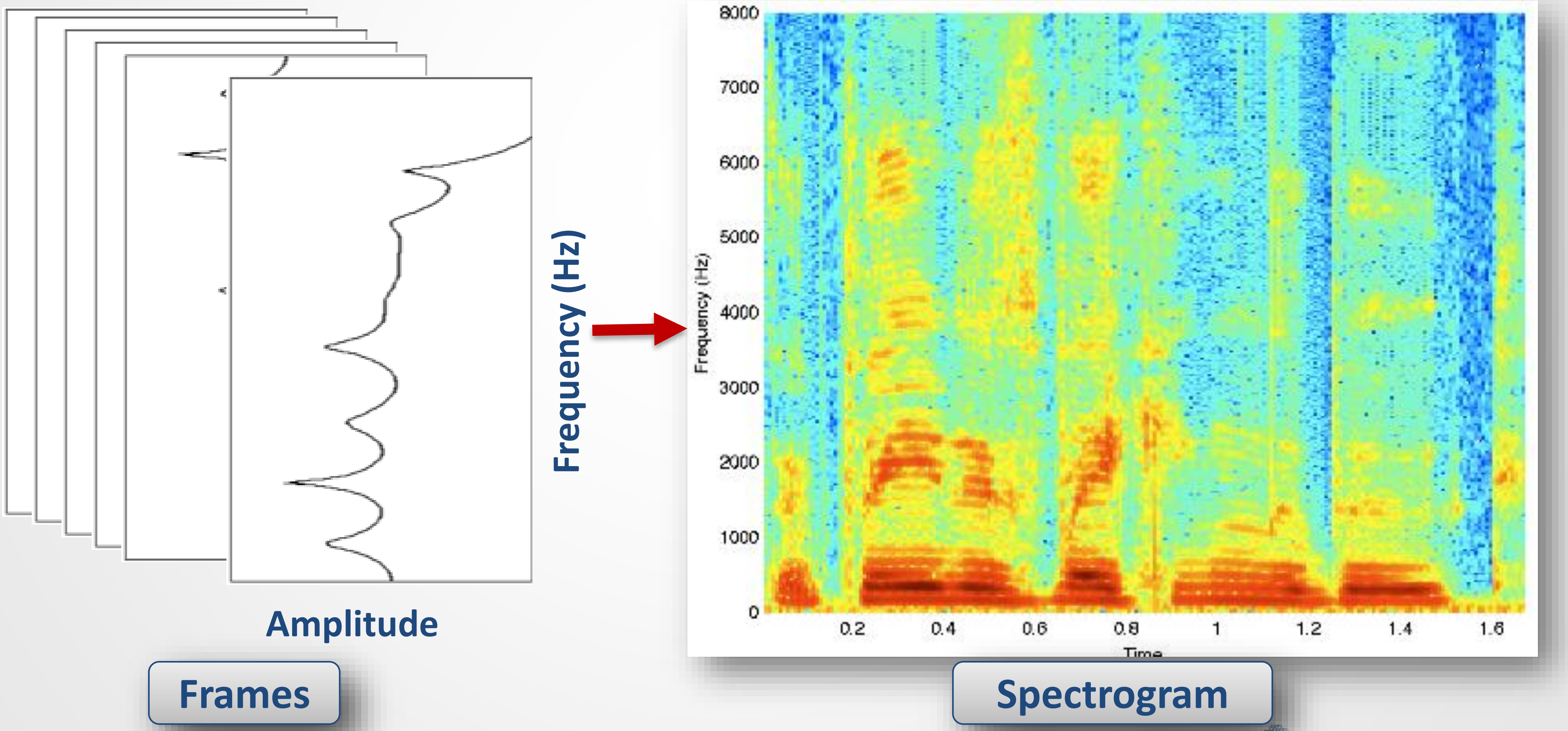


Recall: the Fourier transform is invertible

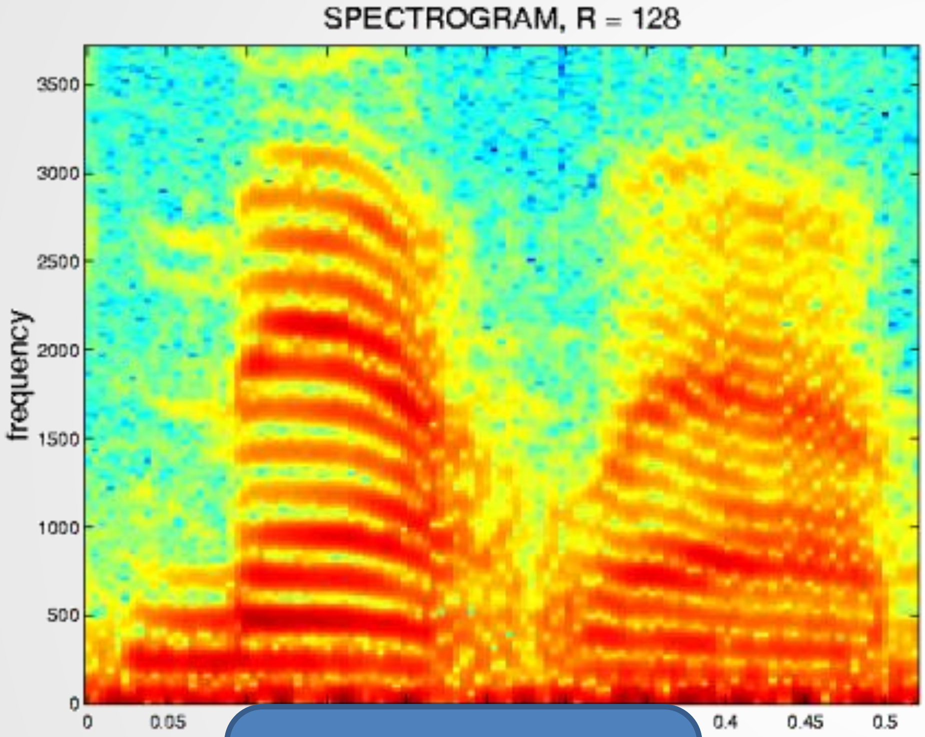
But this is all just for a small window…

Spectrograms

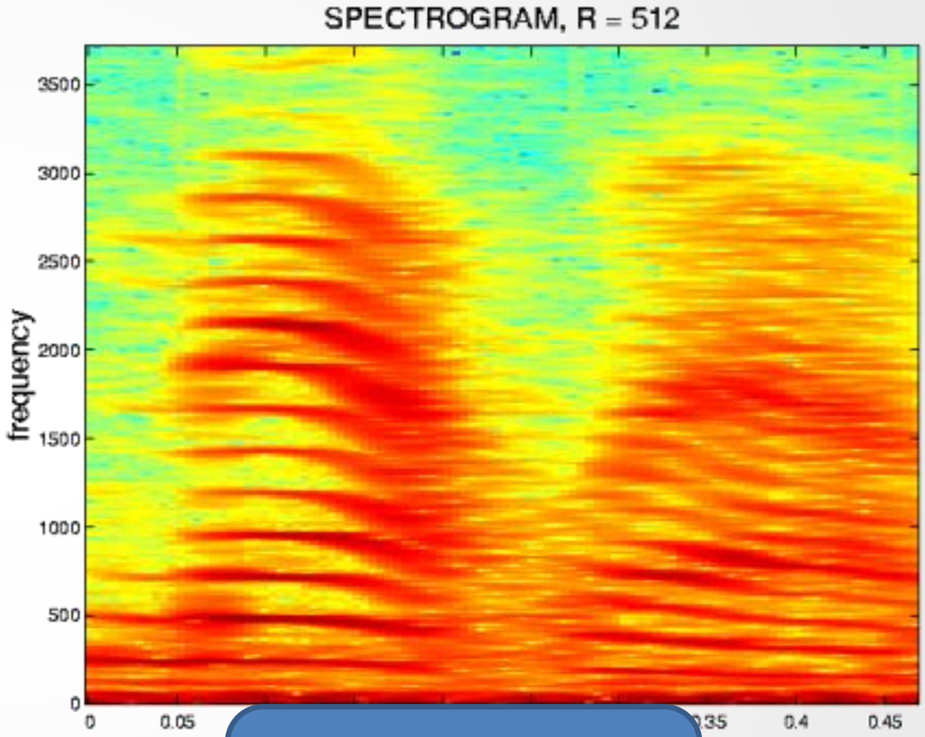
- **Spectrogram**: *n.* a 3D plot of **amplitude** and **frequency** over **time** (higher ‘redness’ → higher amplitude).



Effect of window length

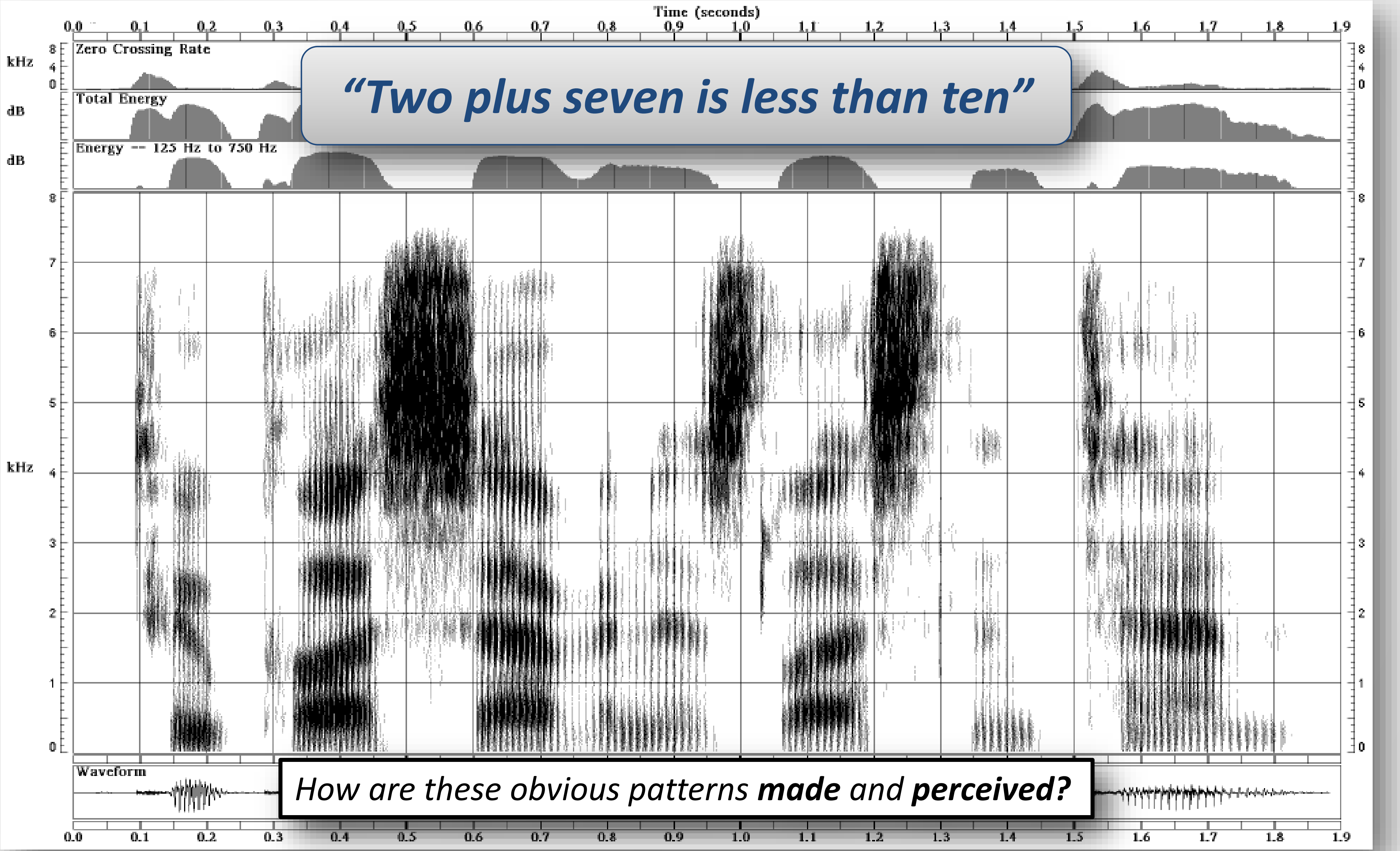


Wide-band
(better time
resolution)



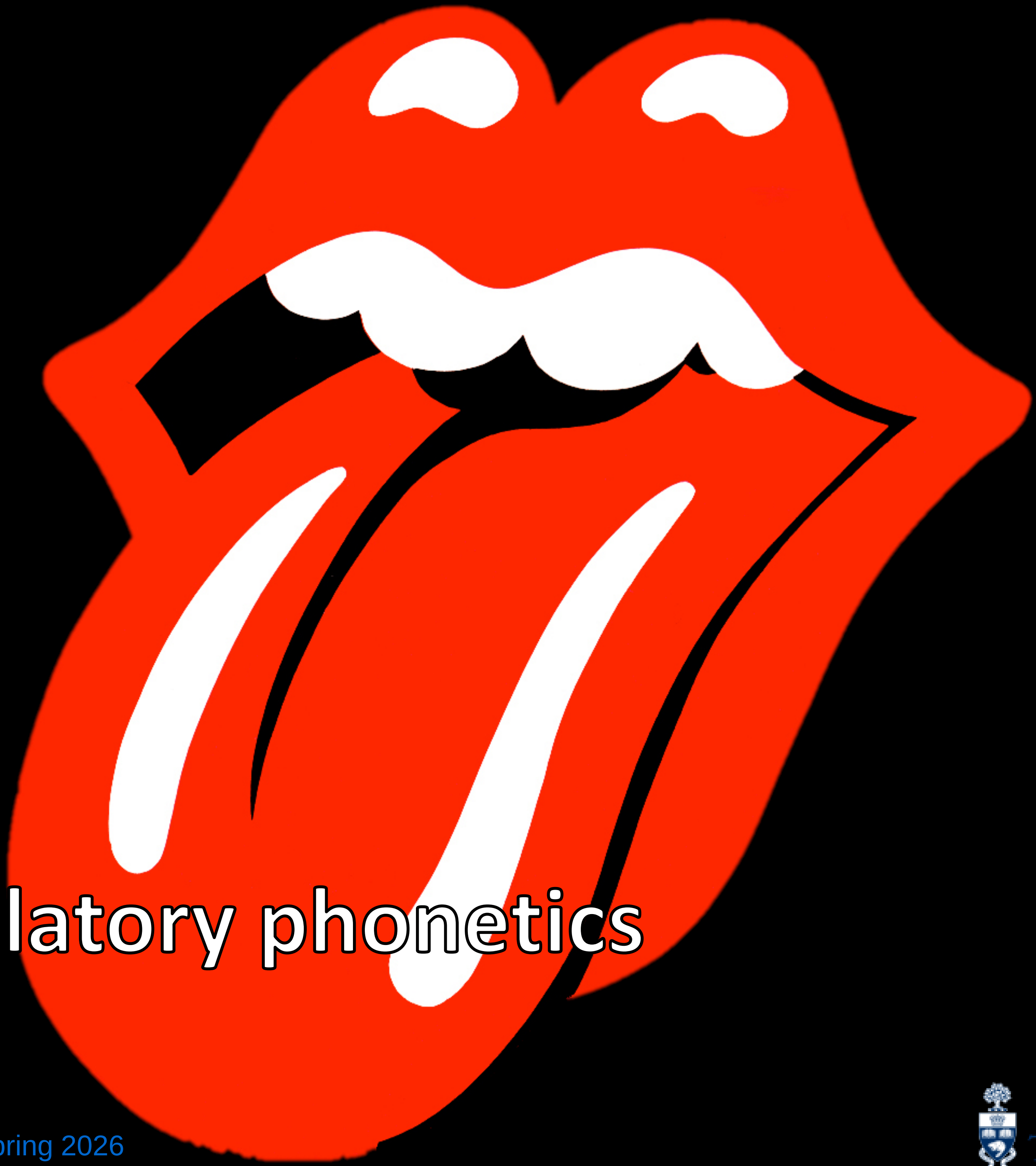
Narrow-band
(better frequency
resolution)

Spectrograms



“Two plus seven is less than ten”

*How are these obvious patterns **made** and **perceived?***



Articulatory phonetics

Sounds and transcriptions

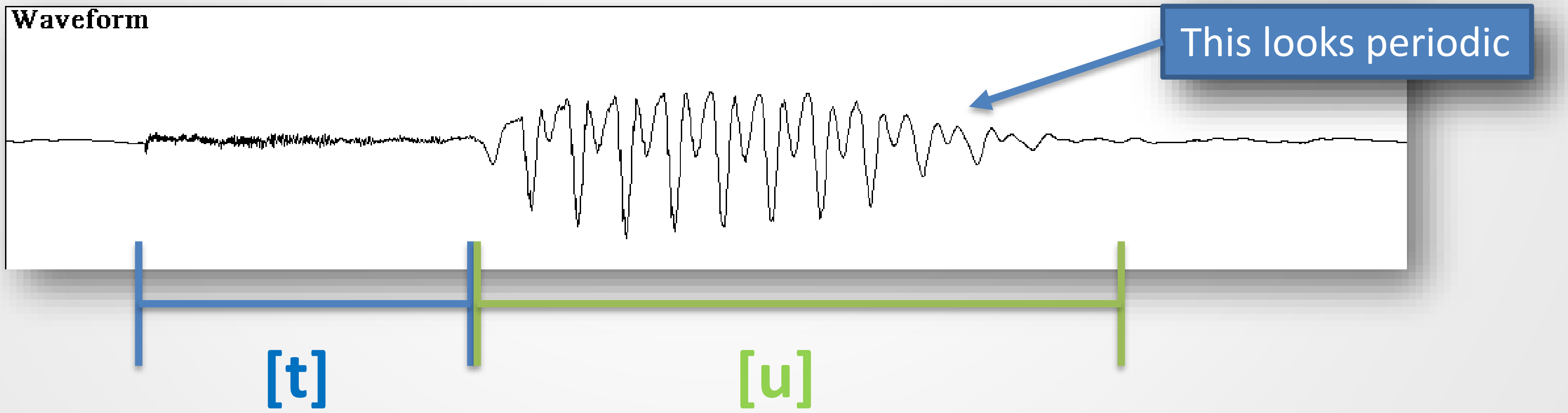
- We are often interested in the meaning of an utterance
- In English, we often transcribe utterances as word tokens
  - We write: <How to recognize speech>
- Is this “what was said?”
  - We might write instead: <How to wreck a nice beach>
  - We can transcribe (or even adopt) foreign words
    - 沙发 = <sofa>, not <sandy hair>
  - We can even transcribe brand new words
    - <delulu>
- We can instead transcribe “speech sounds”

Phones and phonetics

- **Phonetics** is the study of speech sounds
- A **phone** is a unit of speech
  - Denoted with square braces: [t], [t^h], [u]
  - Language-independent
- Phones which are perceived “similarly” are grouped into **phonemes**
  - Denoted with slashes: /t/, /u/
  - [t],[t^h] ↦ /t/
  - Language-dependent
- Transcriptions are often in-between:
  - [ˈt^{h}uː] ↦ [t^{h}u] ↦ [tu] ↦ /tu/
- We will be very loose with the distinction

Phonetic transcription

- Often, we assume that a **spoken utterance** can be **partitioned** into a **sequence** of **non-overlapping** phones.
  - Demarking the periods during which certain phones are being uttered is called **phonetic segmentation**
  - This approach has problems (e.g., when *exactly* does one phoneme end and another begin?), but it's useful for **classification**.



Phonetic alphabets

- There are several alphabets that categorize the sounds of speech.
  - The **International Phonetic Alphabet (IPA)** is popular, but it uses non-ASCII symbols.
  - The **TIMIT** phonetic alphabet will be used by **default** in this course.

  - Other popular alphabets include **ARPAbet**, **Worldbet**, and **OGIbet**, usually adding special cases.
    - E.g., [pcl] is the period of silence immediately before a [p].

TIMIT	IPA	e.g.
[iy]	[i^{y}]	*beat*
[ih]	[ɪ]	*bit*
[eh]	[ɛ]	*bet*
[ae]	[æ]	*bat*
[aa]	[ɑ]	*Bob*
[ah]	[ʌ]	*but*
[ao]	[ɔ]	*bought*
[uh]	[ʊ]	*book*
[uw]	[u]	*boot*
[ux]	~~[u]~~	*suit*
[ax]	[ə]	*about*

TIMITbet (incomplete)

Vowel	e.g.
[iy]	*beat*
[ih]	*bit*
[eh]	*bet*
[ae]	*Bat*
[aa]	*Bob*
[ah]	*But*
[ao]	*bought*
[uh]	*book*
[uw]	*boot*
[ux]	*suit*
[ax]	*about*

stop	e.g.
[b]	*Bilbo*
[d]	*dada*
[g]	*Gaga*
[p]	*Pippin*
[t]	*Toots*
[k]	*kick*

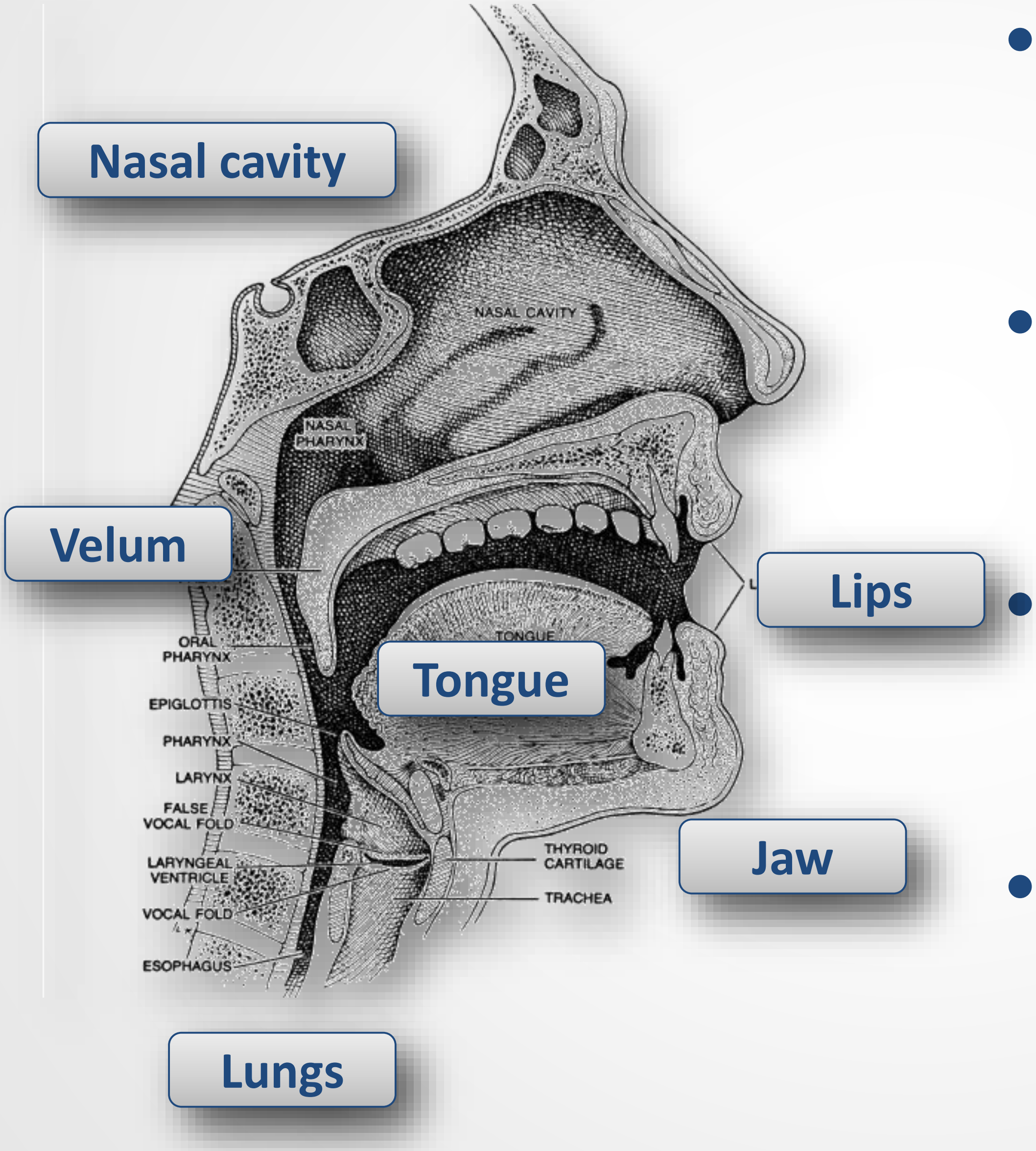
nasal	e.g.
[m]	*Mama*
[n]	*noon*
[ng]	*thing*

fricative	e.g.
[s]	*Sea*
[f]	*Frank*
[z]	*Zappa*
[th]	*this*
[sh]	*Ship*
[zh]	*azure*
[v]	*Vogon*
[dh]	*then*

…

(Incomplete)

The vocal tract



- Many physical structures are co-ordinated in the production of speech.
- Generally, sound is **generated** by passing air through the vocal tract.
- Sound is **modified** by constricting airflow in particular ways.
- We can classify phones by how they are **produced**

A taxonomy of phones

- Phones fall into two broad categories:
- **Vowels** are
  - Always periodic
  - Produced with relatively unobstructed airflow
  - Use tongue, lips, and jaw to produce **resonances** in vocal tract, in turn generating **formants**
- **Consonants** are
  - Mostly noisy (not nasals, semivowels)
  - Produced by obstructing airflow
  - Classified by the **place** and **manner** of primary obstruction, as well as **voicing**

Voicing and fundamental frequency

- **Voiced** phones are produced with vibrating **vocal folds**
  - The space between the folds is the **glottis**
- Vowels are generally voiced; consonants can be **unvoiced**
- F_0: *n.* (**fundamental frequency**), the rate of vibration (Hz)
  - Very indicative of speaker

	Avg F_0 (Hz)	Min F_0 (Hz)	Max F_0 (Hz)
Male	125	80	200
Female	225	150	350
Children	300	200	500

Vowels

- There are approximately **19** vowels in Canadian English, including **diphthongs** in which the articulators **move** over time.
- Vowels are distinguished primarily by their **formants**. (?)

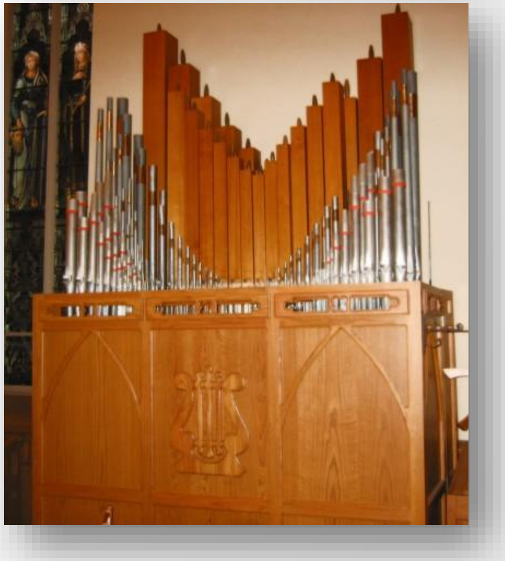
other	e.g.
[er]	*Bert*
[axr]	*butter*

diphthong	e.g.
[ey]	*bait*
[ow]	*boat*
[ay]	*bite*
[oy]	*boy*
[aw]	*bout*
[ux]	*suit*

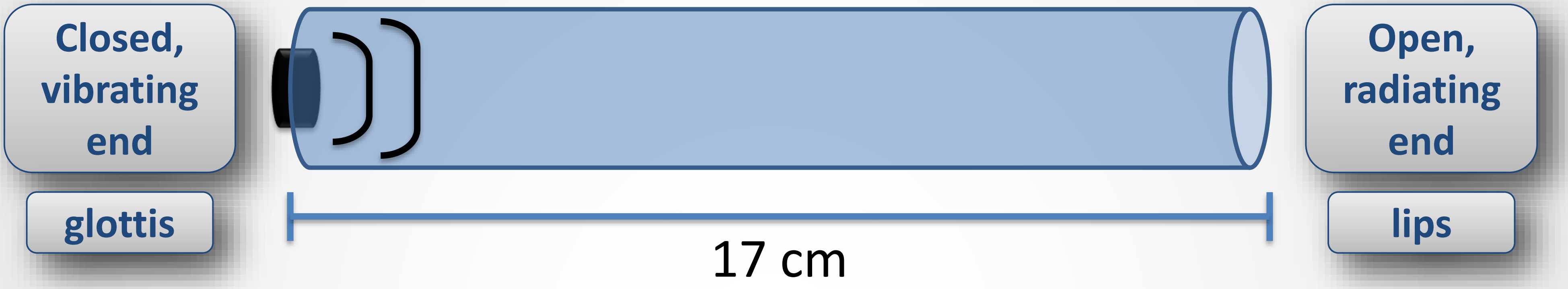
Mono-phthong	e.g.
[iy]	*beat*
[ih]	*bit*
[eh]	*bet*
[ae]	*bat*
[aa]	*Bob*
[ao]	*bought*
[ah]	*but*
[uh]	*book*
[uw]	*boot*
[ax]	*about*
[ix]	*roses*

Uniform tubes

- Formants and resonances can be approximated with tubes
- Many **musical instruments** are based on the idea of uniform (or, in many cases, bent) tubes.
- **Longer** tubes produce ‘**deeper**’ sounds (lower frequencies).
    - A tube ½ the length of another will be 1 octave higher.



The uniform tube



Closed, vibrating end

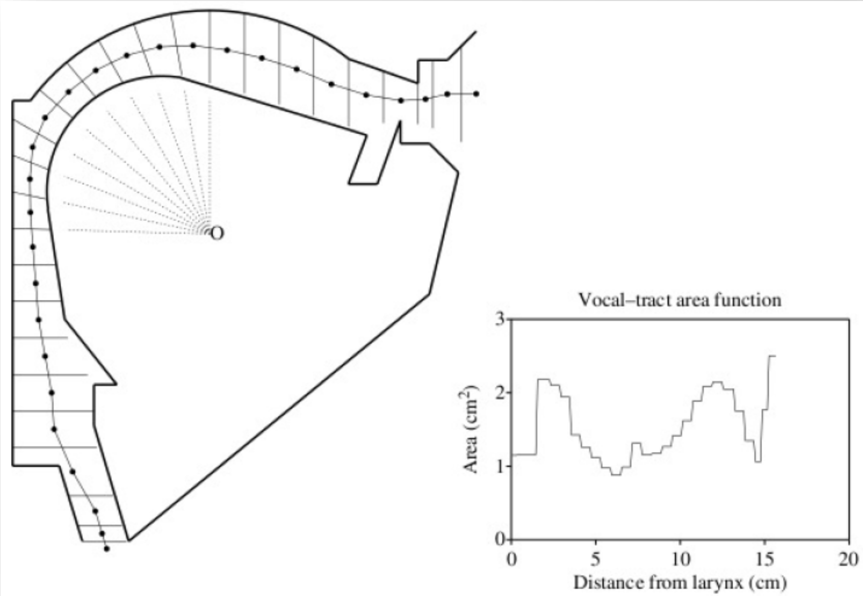
glottis

17 cm

Open, radiating end

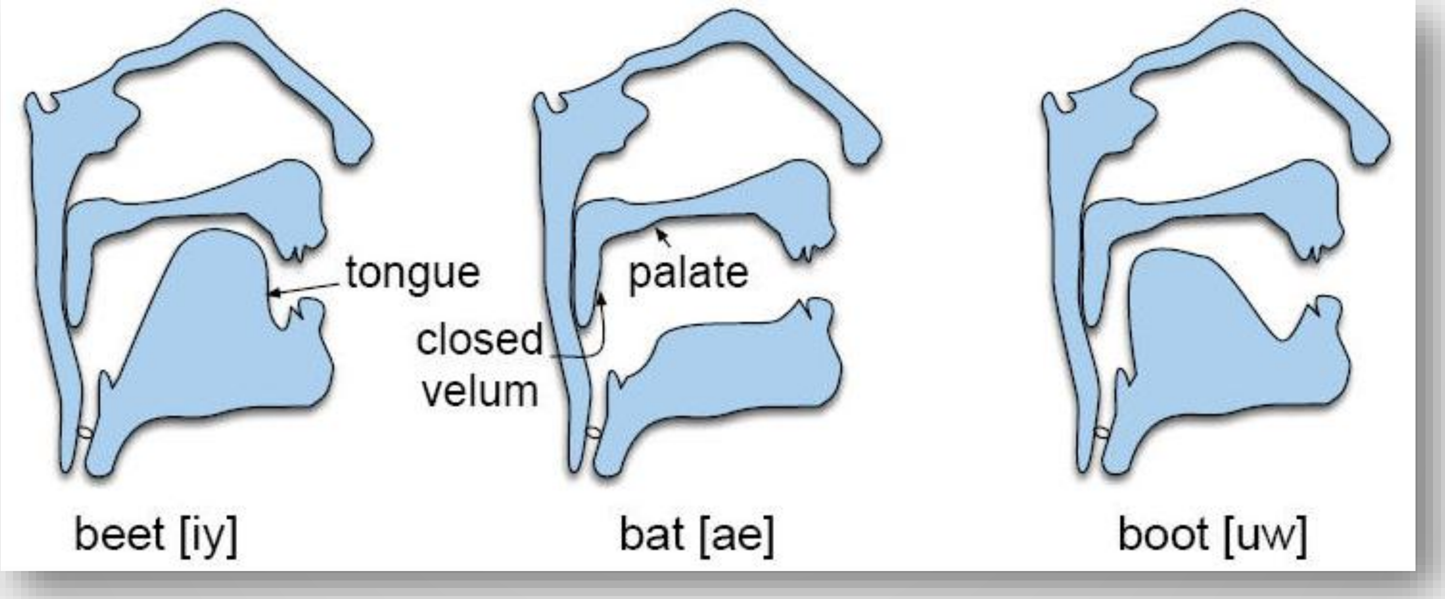
lips

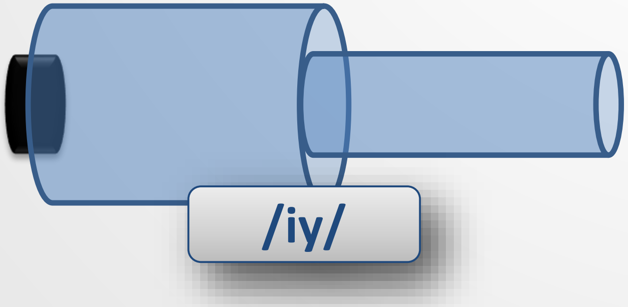
- The positions of the tongue, jaw, and lips change the **shape** and **cross-sectional** area of the vocal tract.

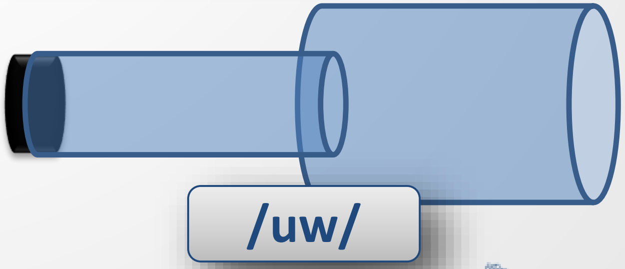


Vowels as concatenated tubes

- The vocal tract can be modelled as the concatenation of dozens, hundreds, or thousands of tubes.

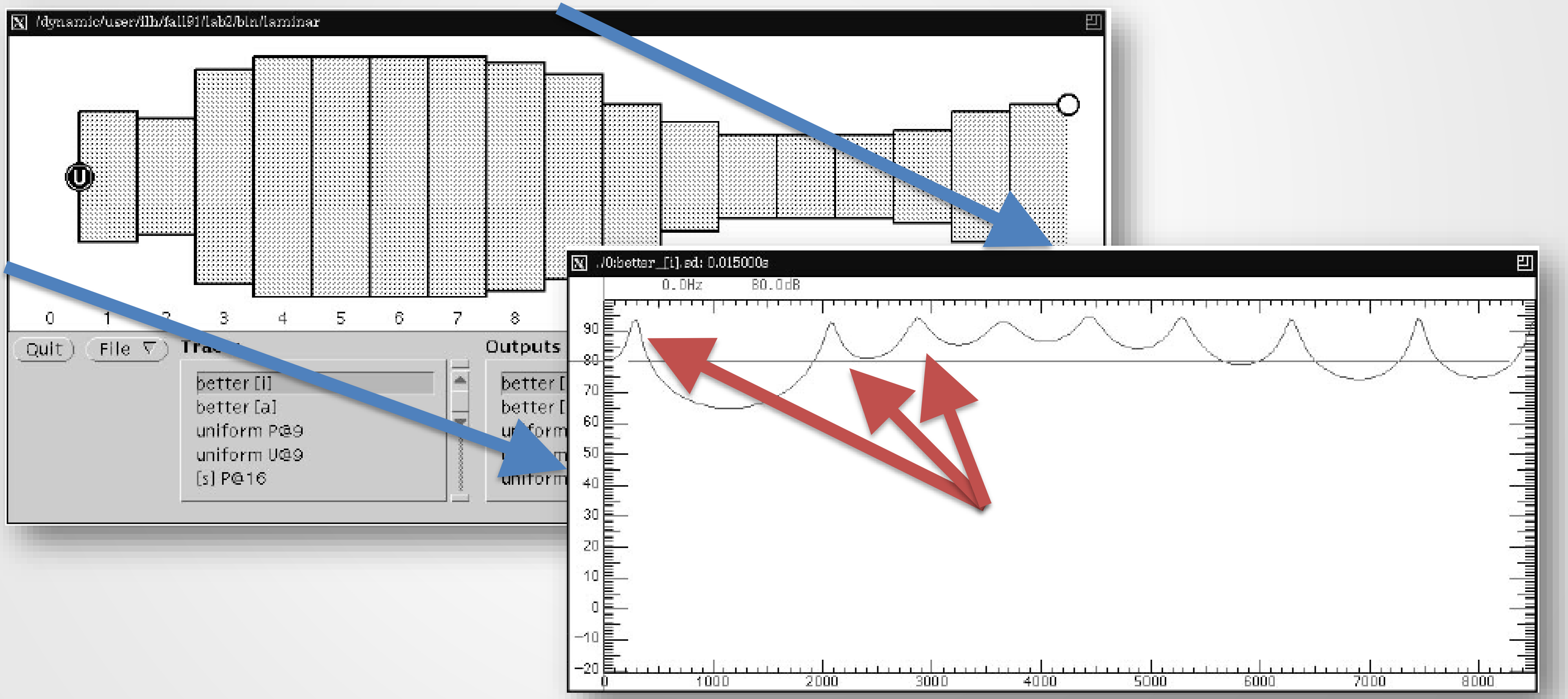






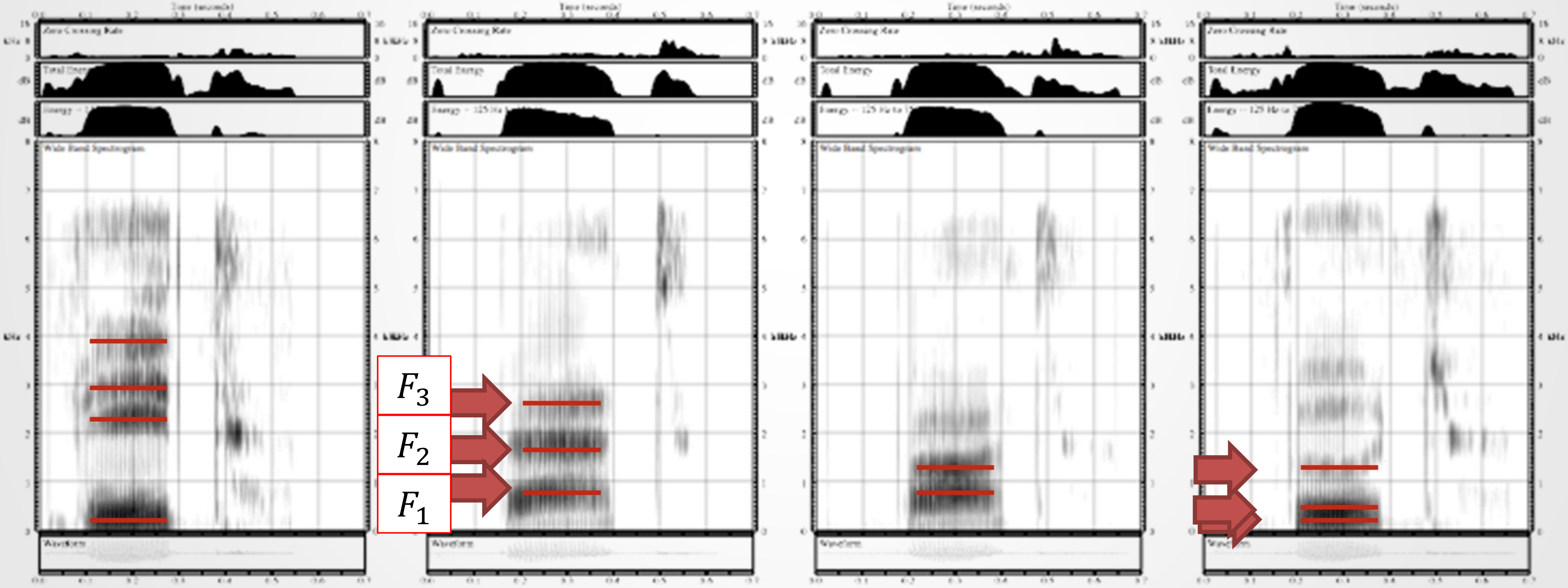
Waves in concatenated tubes

- Reflections at tube boundaries produce resonances which amplify certain frequencies



Formants and vowels

- **Formant**: *n.* A concentration of energy within a frequency band. Ordered from low to high bands (e.g., F_1, F_2, F_3).



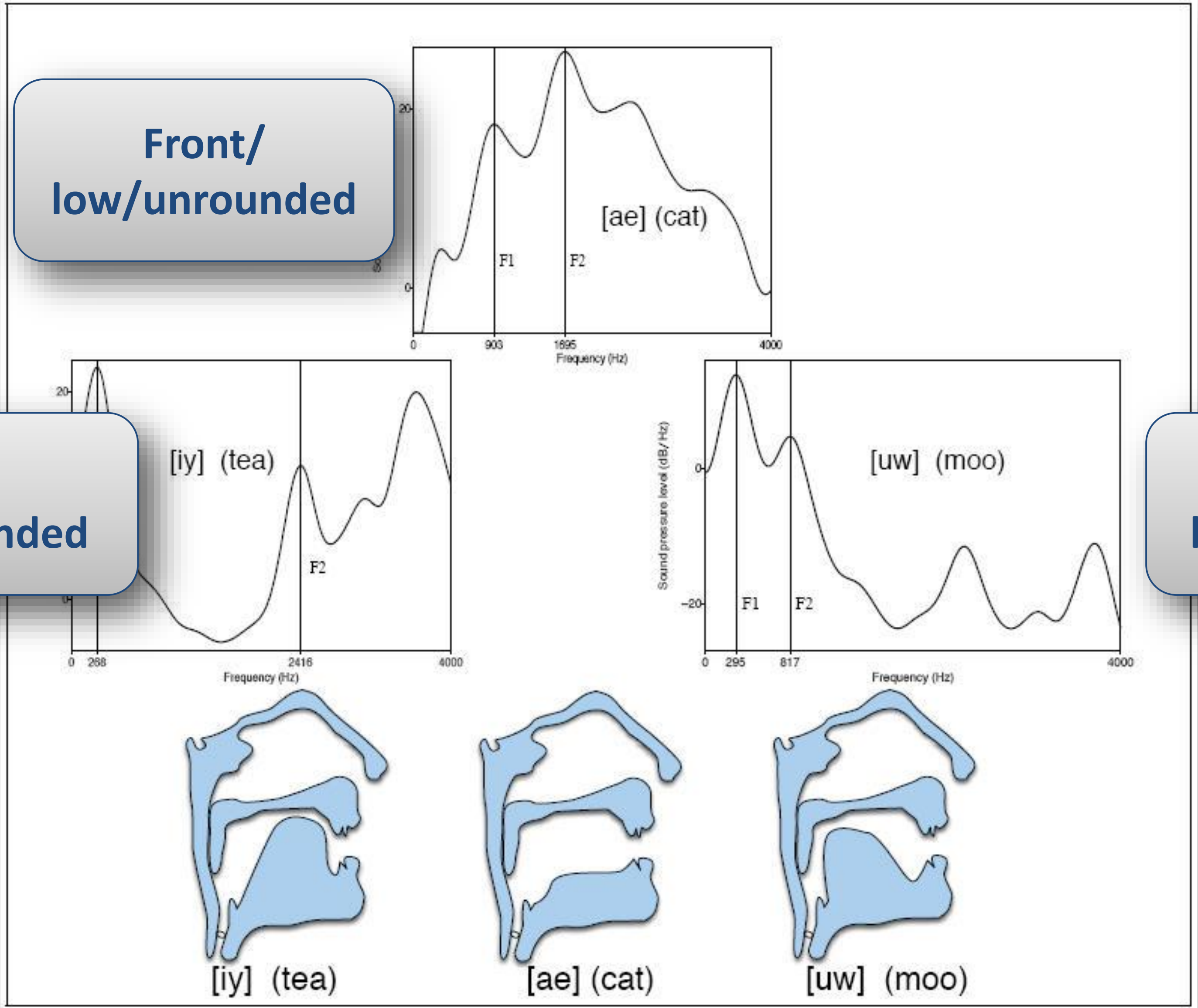
beet /biʸt/

bat /bæt/

bott /bat/

boot /but/

Tongue, lips, and formants

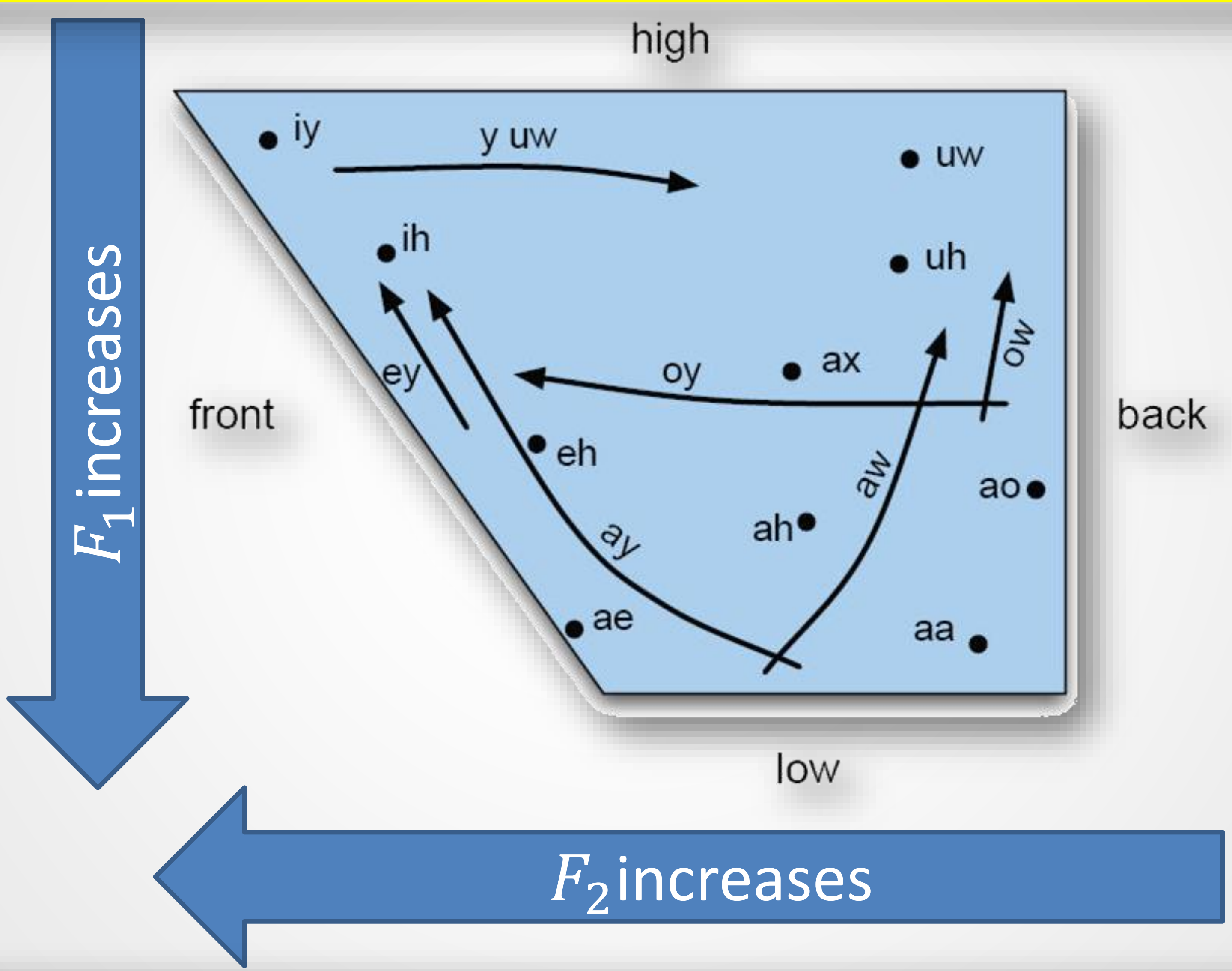


Front/
low/unrounded

Front/
high/unrounded

Back/
high/rounded

The vowel trapezoid

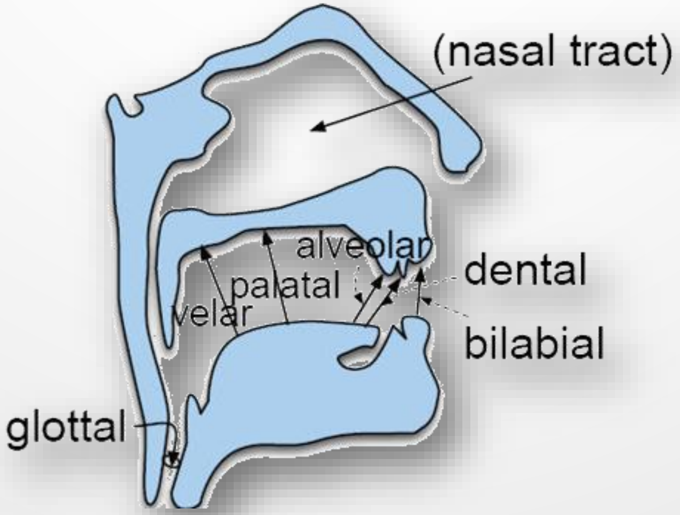


Manner of articulation

- Consonants are classified by **place** and **manner** of obstruction
- For manner:
    - **Fricatives:** **noisy**, with air passing through a tight constriction (e.g., ‘**_sh_**_i_**_f_**_t_’).
    - **Stops/plosives:** **complete** vocal tract constriction and burst of energy (e.g., ‘**_p_**_a_**_p_**_a_’).
    - **Nasals**: air passes through the **nasal** cavity (e.g., ‘**_m_**_a_**_m_**_a_’).
    - **Semivowels**: similar to vowels, but typically with more constriction (e.g., ‘**_w_**_a_**_ll_**’).
    - **Affricates**: Alveolar stop followed by fricative.
    - **Taps:** Quick collision of articulators (‘butter’)

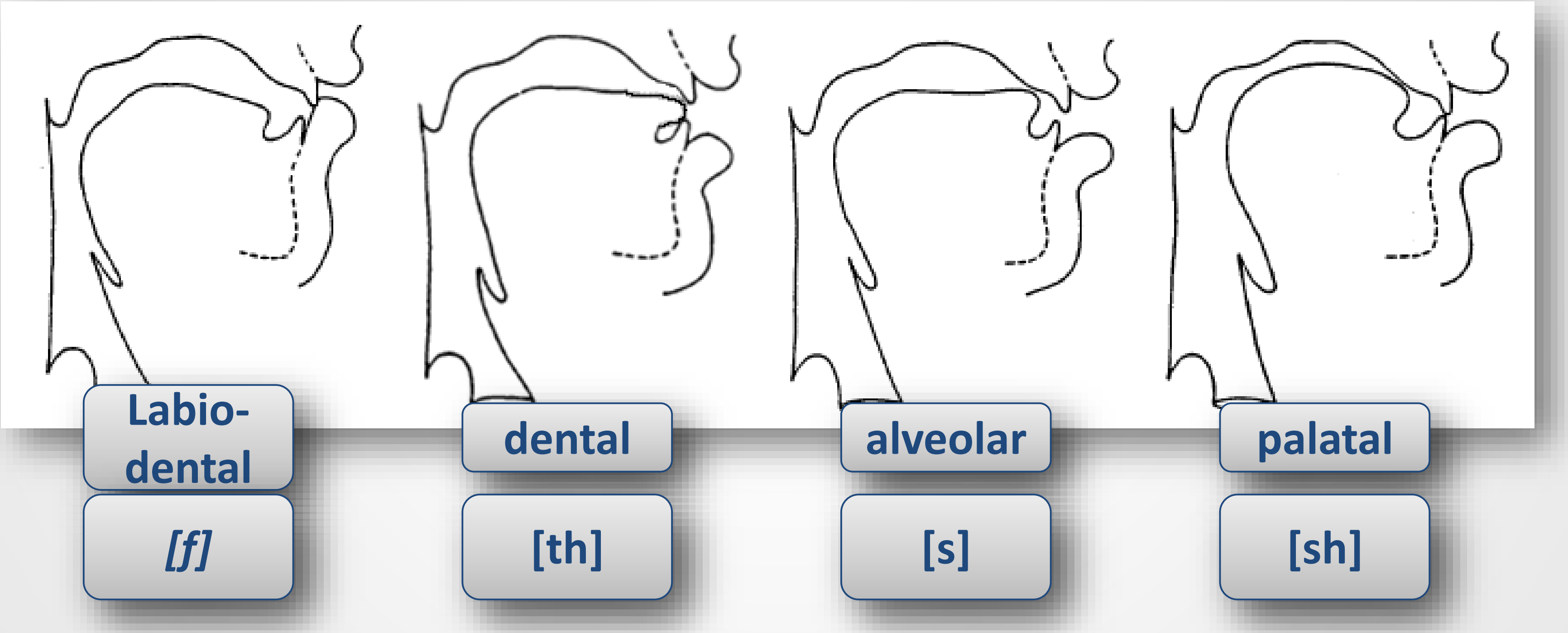
Place of articulation

- The **location** of the *primary constriction* can be:
  - **Alveolar**: constriction near the alveolar ridge (e.g., [*t*])
  - **Bilabial**: touching of the lips together (e.g., [*m*], [*p*])
  - **Dental**: constriction of/at the teeth (e.g., [*th*])
  - **Labiodental**: constriction between lip and teeth (e.g., [*f*])
  - **Velar**: constriction at or near the velum (e.g., [*k*]).
  - **Glottal:** constriction of the glottis ([q])

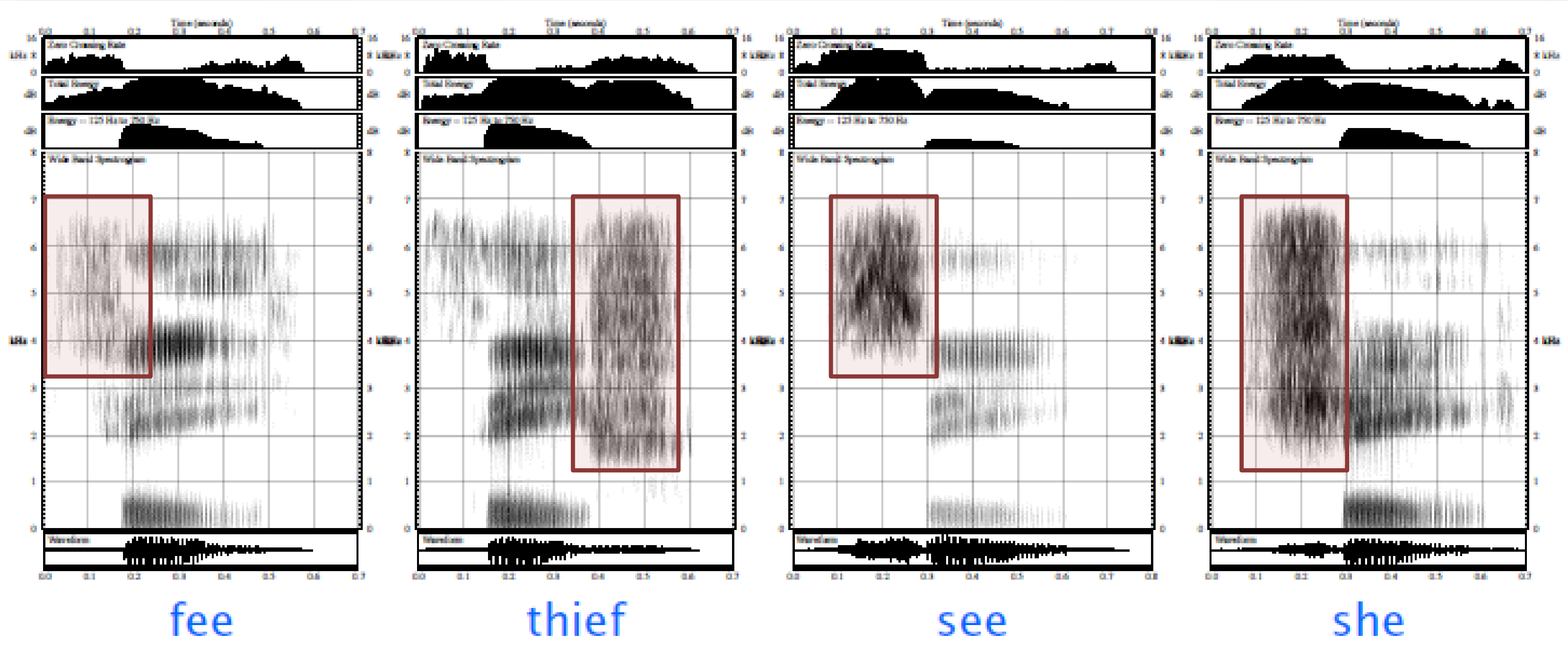


Fricatives

- **Fricatives** are caused by acoustic turbulence at a **narrow constriction** whose position determines the sound.

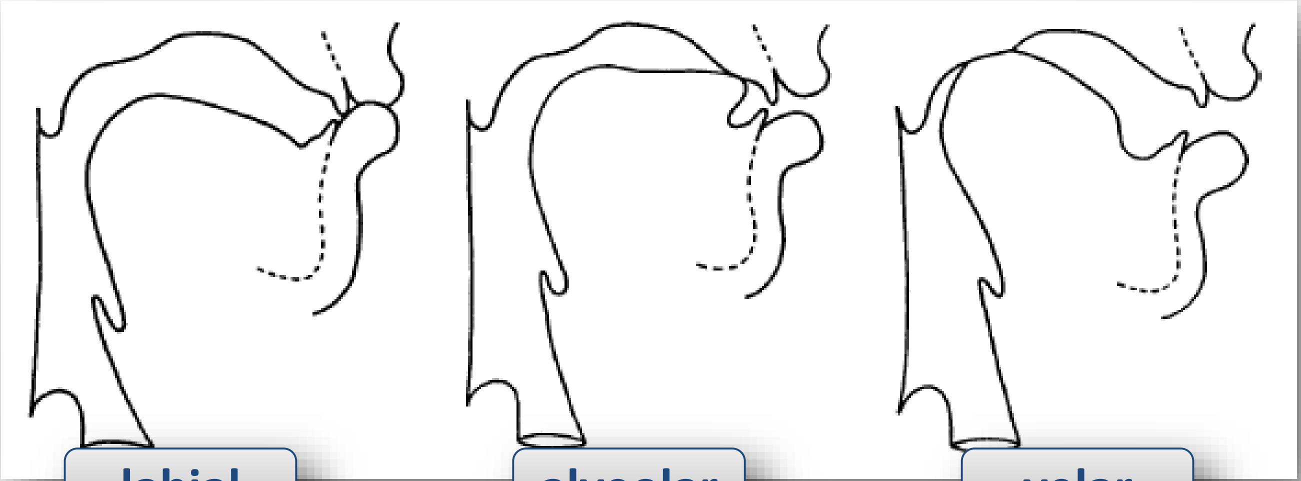


Unvoiced fricatives



Plosives (3/6)

- **Plosives** build pressure behind a **complete closure** in the vocal tract.
- A **sudden release** of this constriction results in **brief noise**.



labial

[b]

alveolar

[d]

velar

[g]

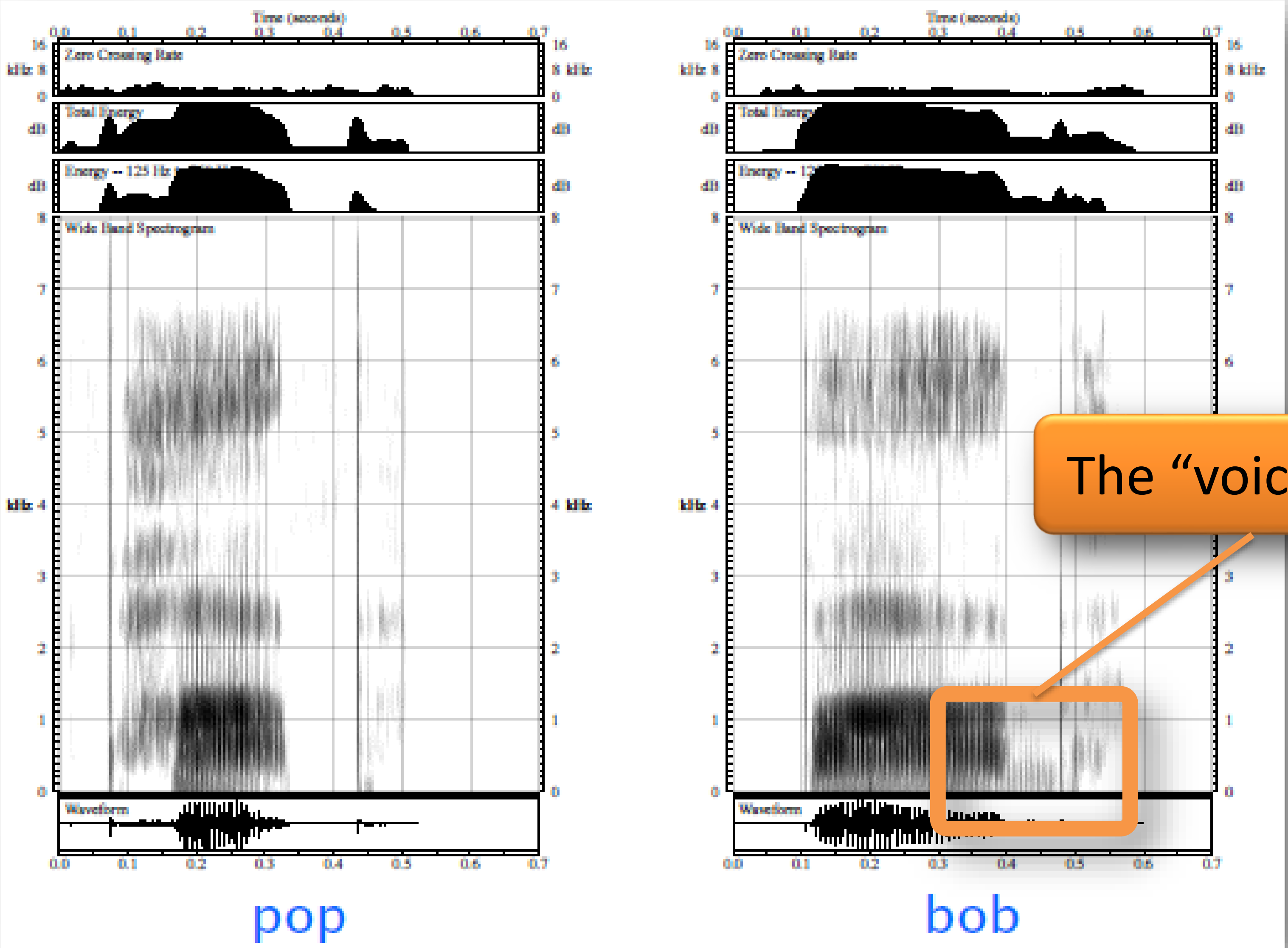
Plosives

- **Plosives** have three places of articulation:

	Unvoiced		Voiced	
Labial	[p]	*porpoise*	[b]	*baboon*
Alveolar	[t]	*tort*	[d]	*dodo*
Velar	[k]	*kick*	[g]	*Google*

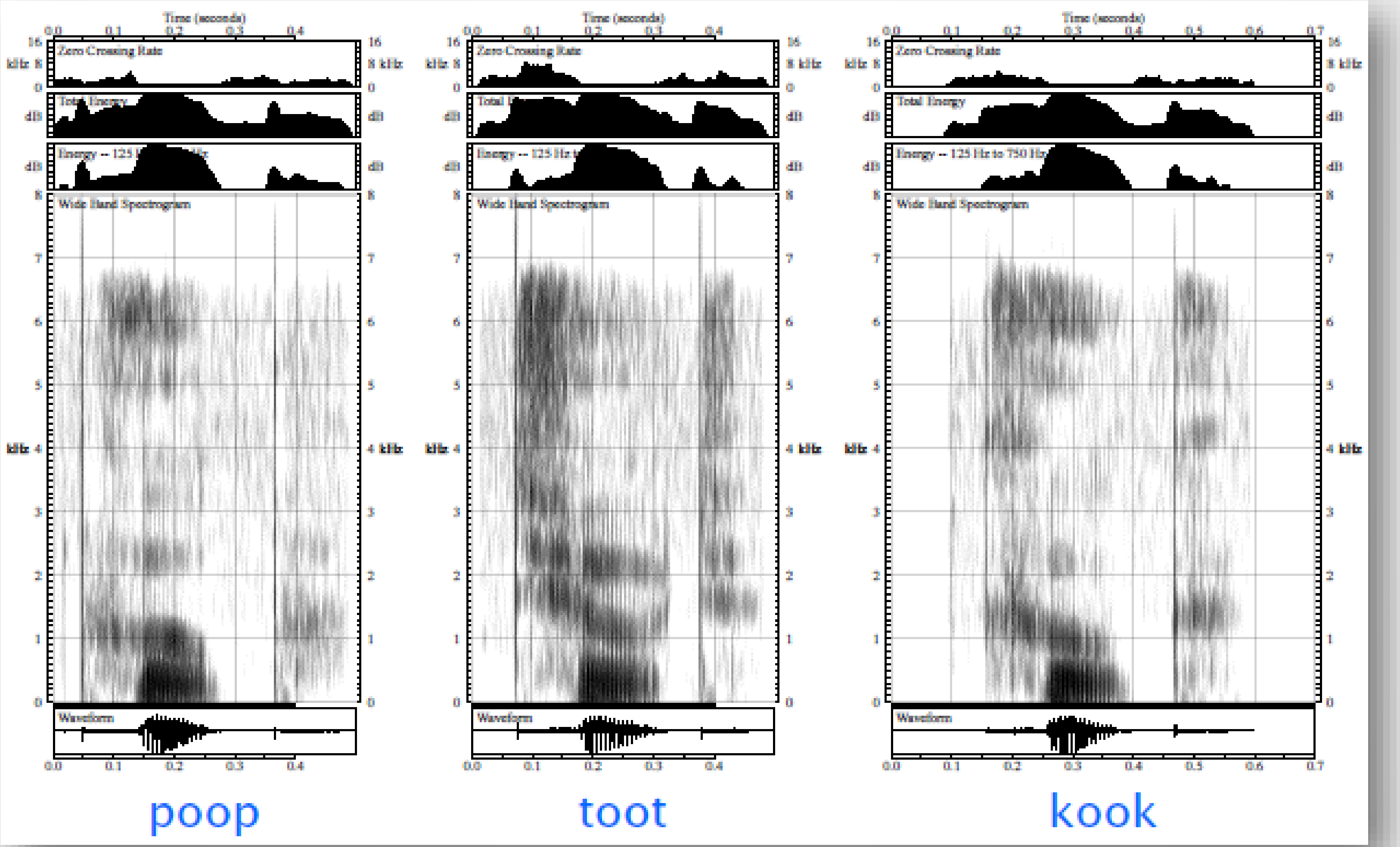
- **Voiced** stops are usually characterized by a “**voice bar**” during closure, indicating the vibrating glottis.
- Formant **transitions** are very **informative** in classification.

Voicing in plosives



The “voice bar”

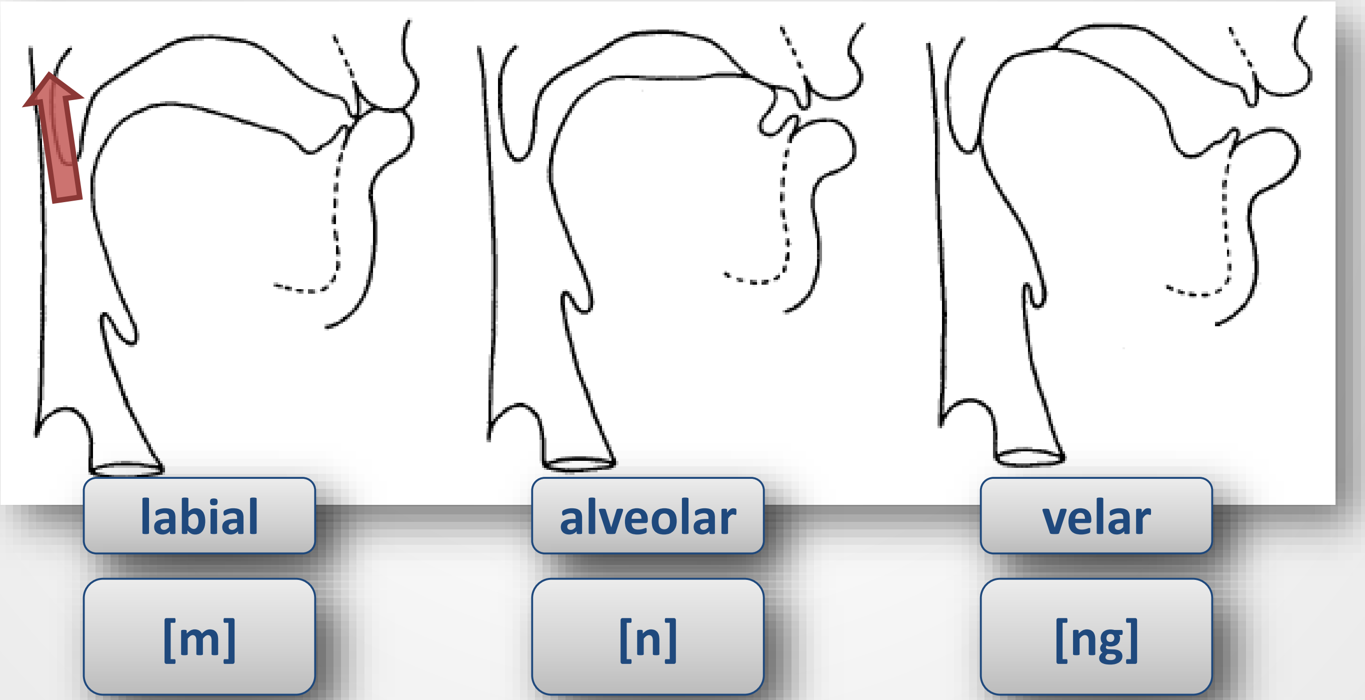
Formant transitions in plosives



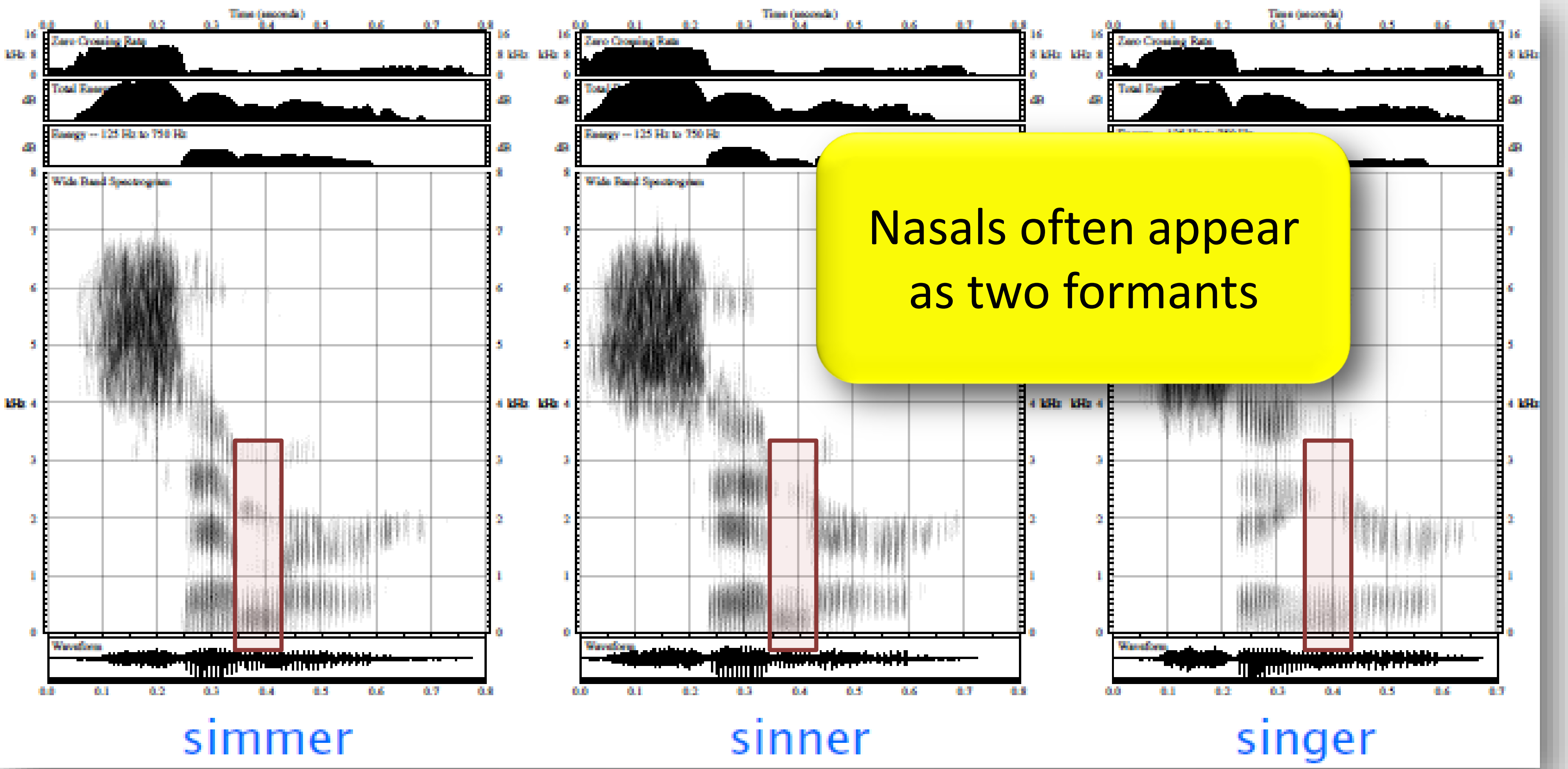
- Despite a **common** vowel, the **motion** of F_2 and F_3 into (and out of) the vowel helps identify the plosive.

Nasals

- **Nasals** involve lowering the velum so that air passes through the **nasal cavity**.
- **Closures** in the oral cavity (at same positions as plosives) change the resonant characteristics of the nasal sonorant.



Formant transitions among nasals



simmer

sinner

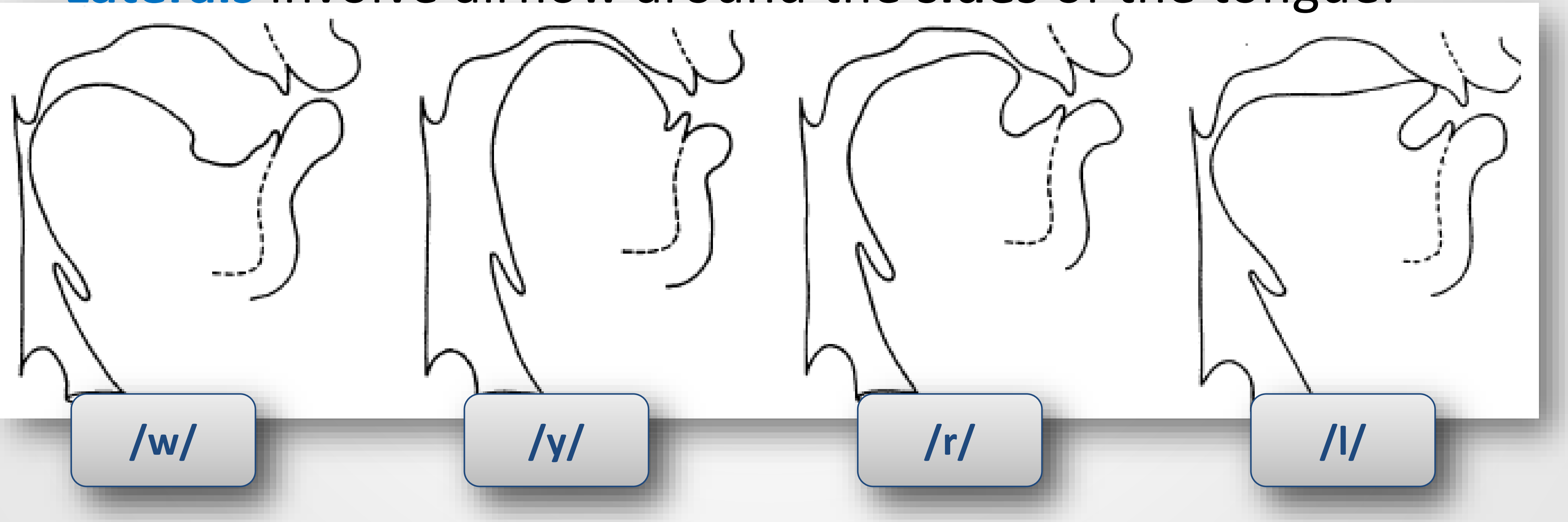
singer

Nasals often appear as two formants

- Despite a common vowel, the motion of F_2 and F_3 before and after each nasal helps to identify it.

Semivowels (5/6)

- **Semivowels** have both consonantal and vocalic realizations, involve constriction in the vocal tract, but exhibit **less turbulence**.
  - They also involve slower articulatory motion.
- **Laterals** involve airflow around the **sides** of the tongue.



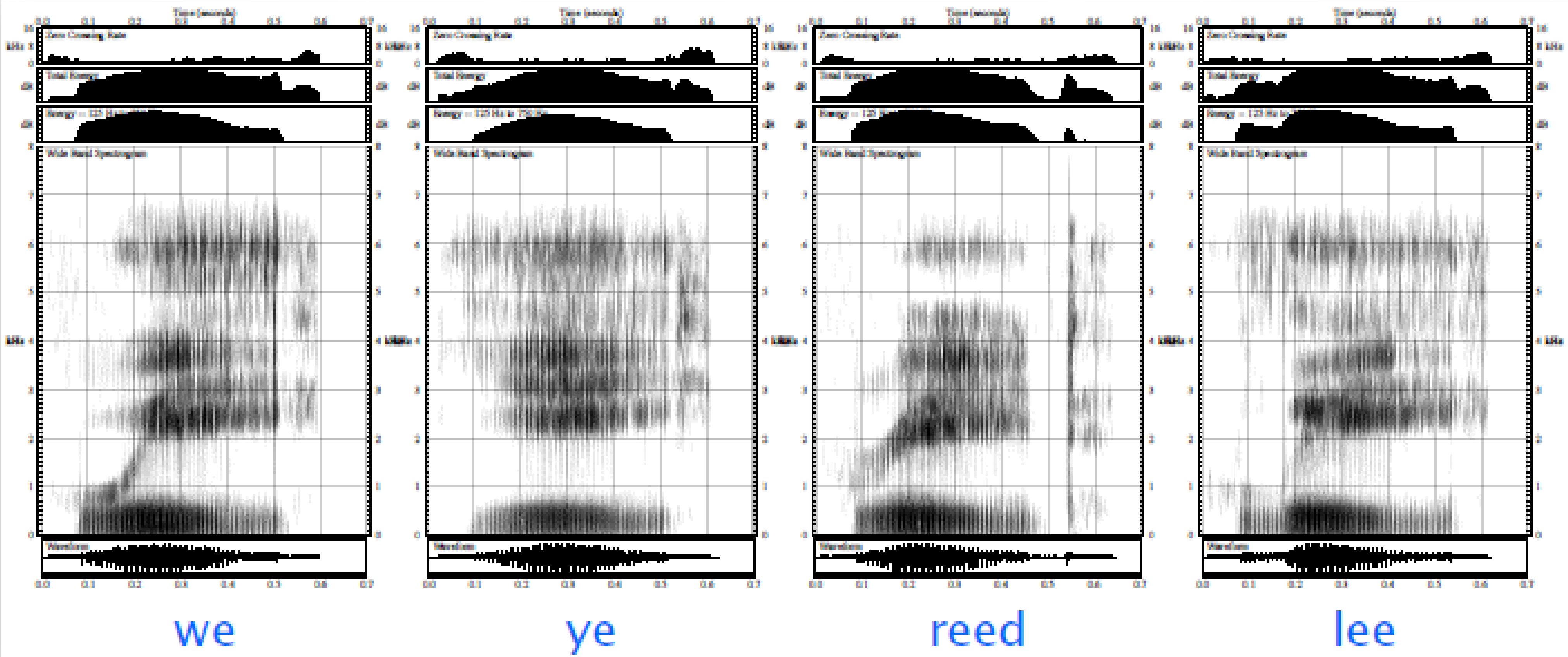
Semivowels

- Semivowels are often sub-classified as glides or liquids.

	Semivowel		Nearest vowel
Glides	[w]	**W**o**w**	[uw]
	[y]	**y**o**y**o	[iy]
Liquids	[r]	**r**ea**r**	[er]
	[l]	**L**u**l**u	[ow]

- Glides and liquids are more constricted versions of corresponding vowels.
  - Similar formants, though generally weaker.

Semivowels



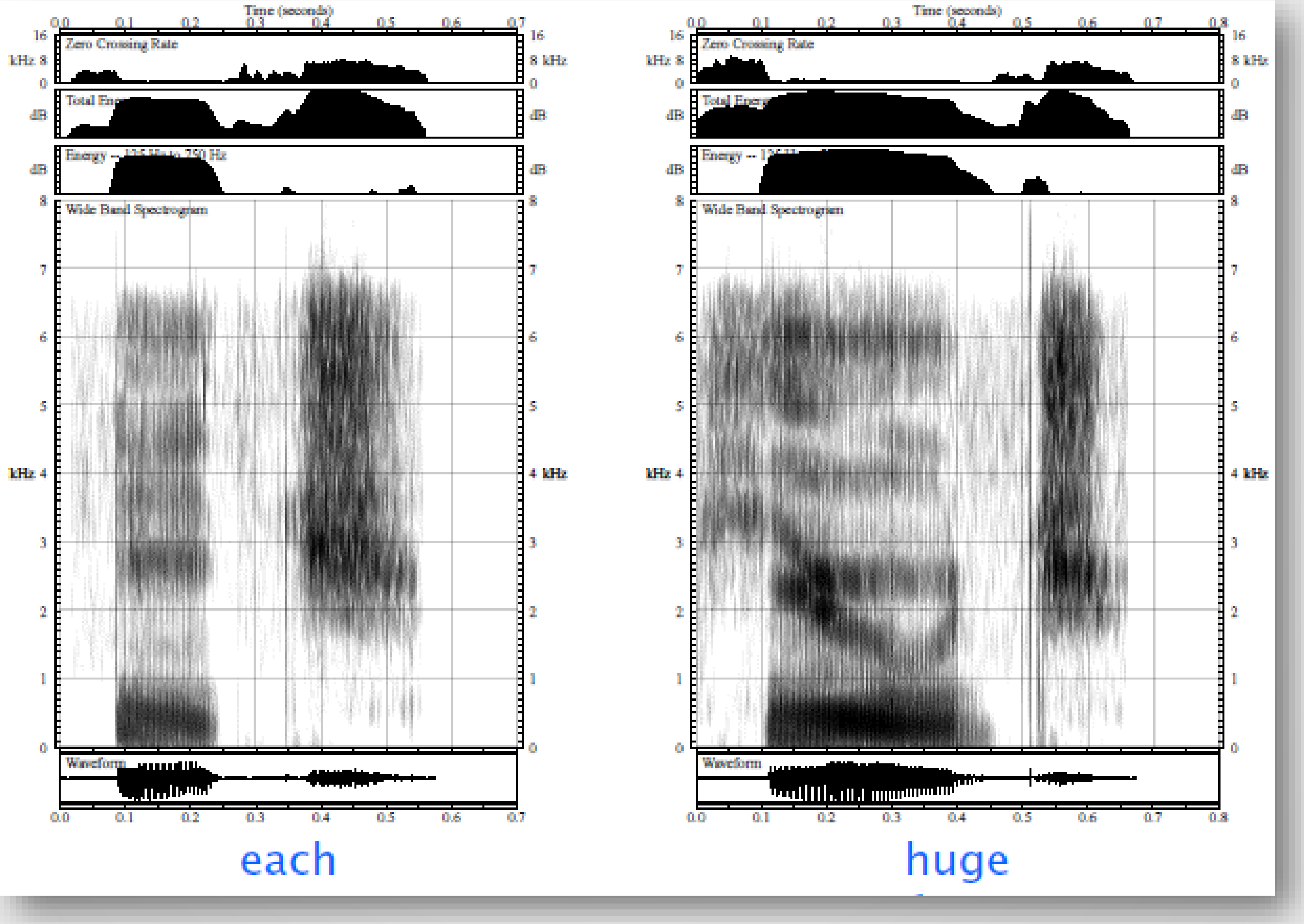
- Note the drastic formant transitions which are more typical of semivowels.

Affricates and aspirants

- Two common **affricates** in English are: [jh] (voiced; e.g., _**j**u**dge**_) and [ch] (unvoiced; e.g., _**ch**ur**ch**_).
  - These involve an **palatal stop** followed by a **fricative**.
  - Voicing in [jh] is normally indicated by voice bars, as with plosives.

- There’s only one **aspirant** in Canadian English: [h] (e.g., _**h**at_)
  - This involves turbulence generated at the **glottis**,
  - In Canadian English, there is **no** constriction in the vocal tract.

Affricates and aspirants



each

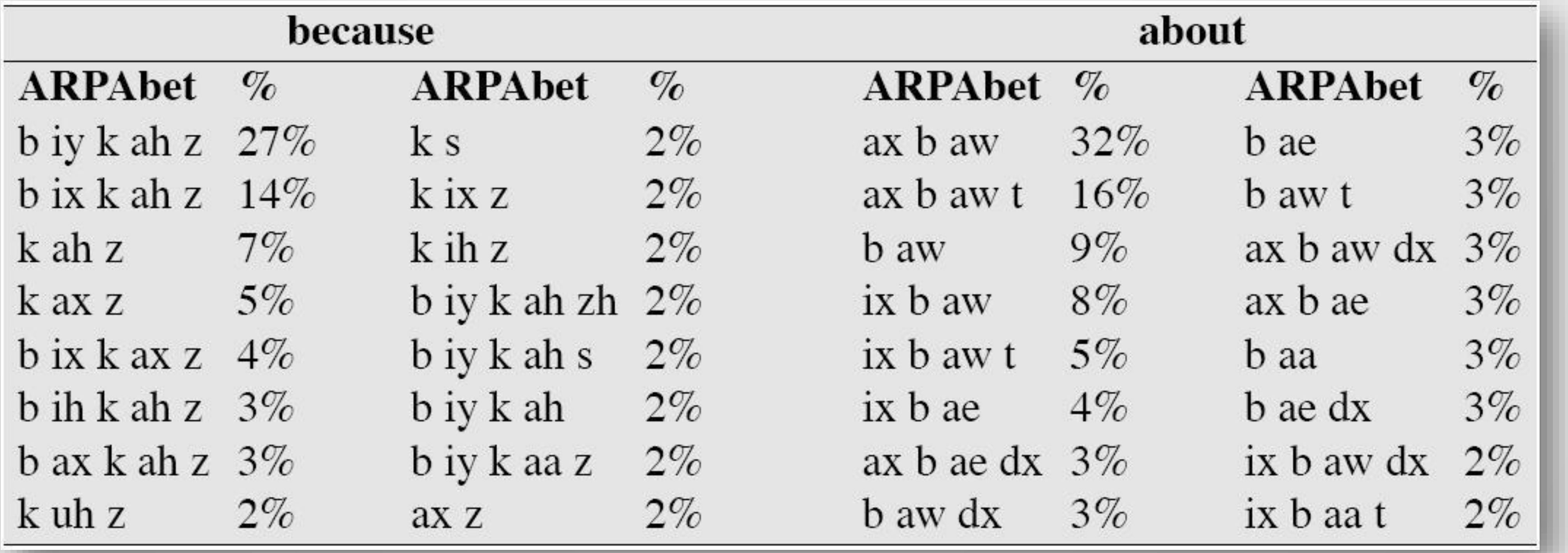
huge

Other topics in phonetics

- The grouping of phones into **syllables**
  - Consisting of a vowel (**nucleus**), and optionally preceding (**onset**) and succeeding (**coda**) consonants
  - Only certain sequences are permissible in English
  - Syllables may be made more **prominent** via pitch, duration, or loudness
- The **prosody**, or intonation and rhythm, of an utterance
  - Prominence can also indicate phrase boundaries
  - Gradual F0 movement (**tune**) can indicate a question or statement
- **Duration**
- These are especially important to **text-to-speech synthesis**

Alternative pronunciations

- **Pronunciations** of words can vary significantly, but with observable **frequencies**.
  - The **Switchboard** corpus is a phonetically annotated database of speech recorded in telephone conversations.



because				about			
ARPAbet	**%**	**ARPAbet**	**%**	**ARPAbet**	**%**	**ARPAbet**	**%**
b iy k ah z	27%	k s	2%	ax b aw	32%	b ae	3%
b ix k ah z	14%	k ix z	2%	ax b aw t	16%	b aw t	3%
k ah z	7%	k ih z	2%	b aw	9%	ax b aw dx	3%
k ax z	5%	b iy k ah zh	2%	ix b aw	8%	ax b ae	3%
b ix k ax z	4%	b iy k ah s	2%	ix b aw t	5%	b aa	3%
b ih k ah z	3%	b iy k ah	2%	ix b ae	4%	b ae dx	3%
b ax k ah z	3%	b iy k aa z	2%	ax b ae dx	3%	ix b aw dx	2%
k uh z	2%	ax z	2%	b aw dx	3%	ix b aa t	2%