Chapter 13: Speech Perception

Overview of Questions

- Can computers perceive speech as well as humans?
- Why does an unfamiliar foreign language often sound like a continuous stream of sound, with no breaks between words?
- Does each word that we hear have a unique pattern of air pressure changes associated with it?
- Are there specific areas in the brain that are responsible for perceiving speech?

Can computers perceive speech as well as humans?

The Speech Stimulus

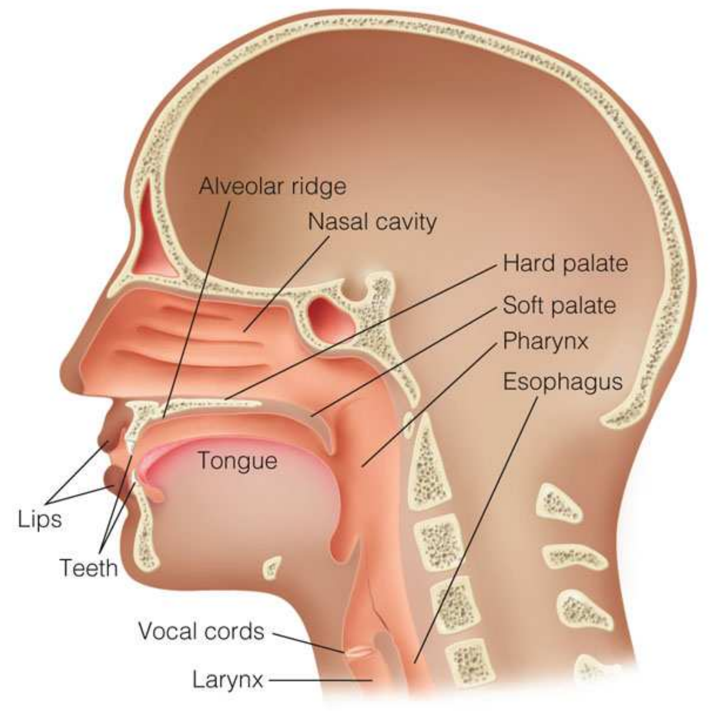
- Phoneme - smallest unit of speech that changes meaning in a word
  - In English there are 47 phonemes:
    - 13 major vowel sounds
    - 24 major consonant sounds
  - Number of phonemes in other languages varied—11 in Hawaiian and 60 in some African dialects

Table 13.1 *Major consonants and vowels of English and their phonetic symbols*

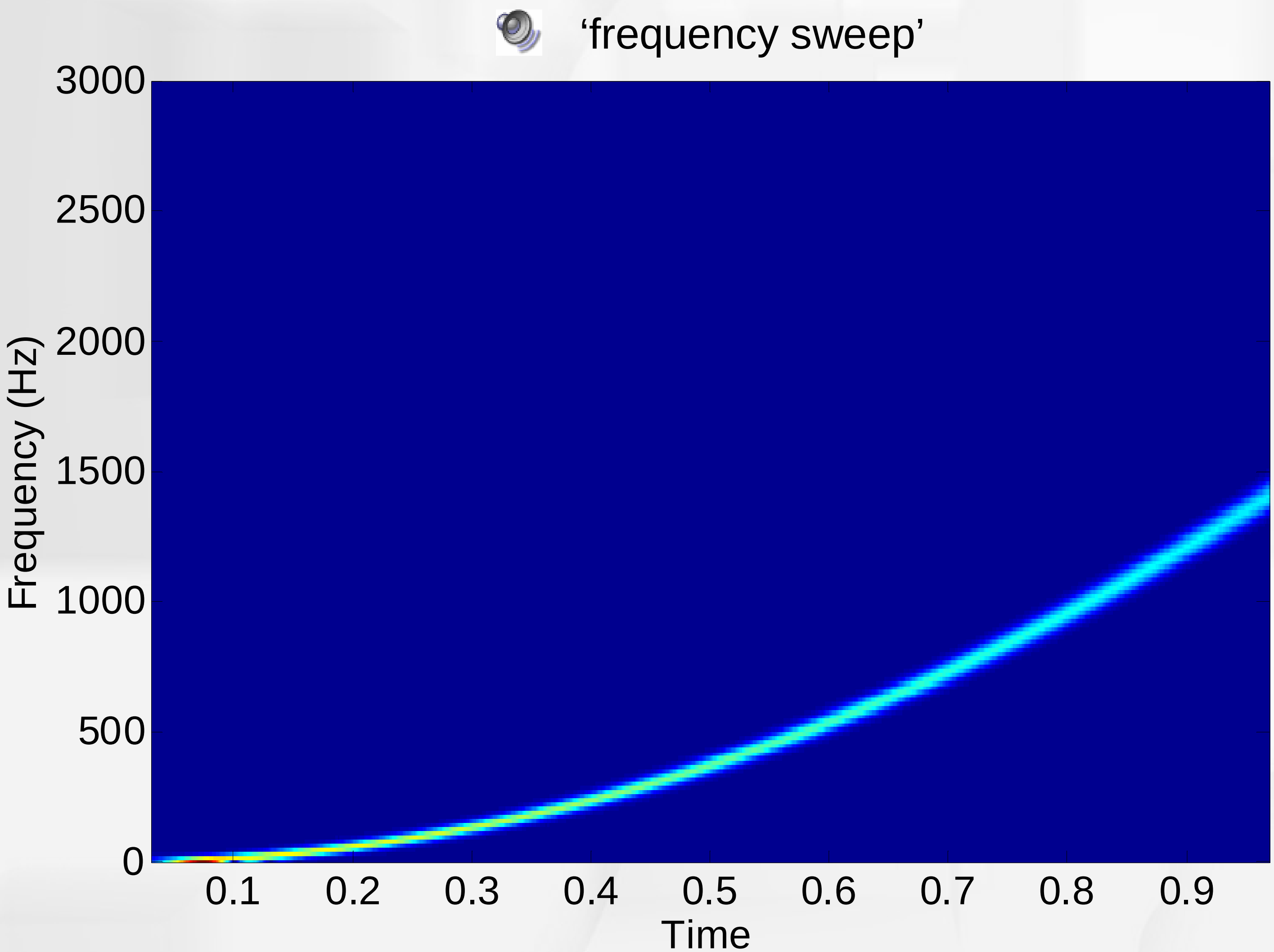
Consonants				Vowels	
p	*pull*	s	*sip*	i	h*ee*d
b	*bull*	z	*zip*	I	h*i*d
m	*man*	r	*rip*	e	b*ai*t
w	*will*	š	*sh*ould	ɛ	h*ea*d
f	*fill*	ž	plea*s*ure	æ	h*a*d
v	*vet*	č	*ch*op	u	wh*o'*d
θ	*th*igh	ǰ	*g*yp	U	p*u*t
ŏ	*th*y	y	*y*ip	ʌ	b*u*t
t	*t*ie	k	*k*ale	o	b*oa*t
d	*d*ie	g	*g*ale	ɔ	b*ough*t
n	*n*ear	h	*h*ail	a	h*o*t
l	*l*ear	ŋ	si*ng*	ə	sof*a*
				ɨ	m*a*ny

The Acoustic Signal

- Produced by air that is pushed up from the lungs through the vocal cords and into the vocal tract
- Vowels are produced by vibration of the vocal cords and changes in the shape of the vocal tract

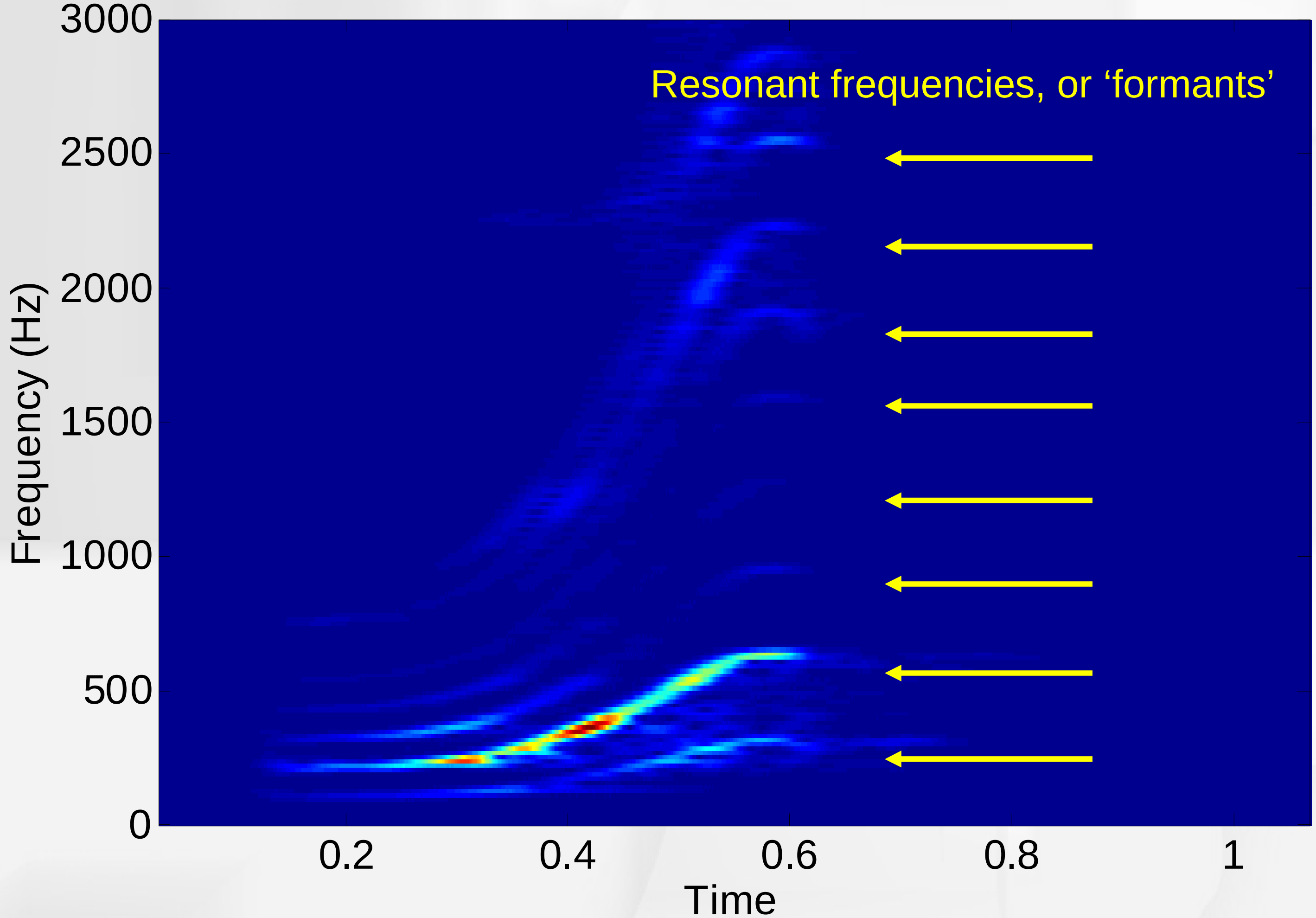


The Sound Spectrogram



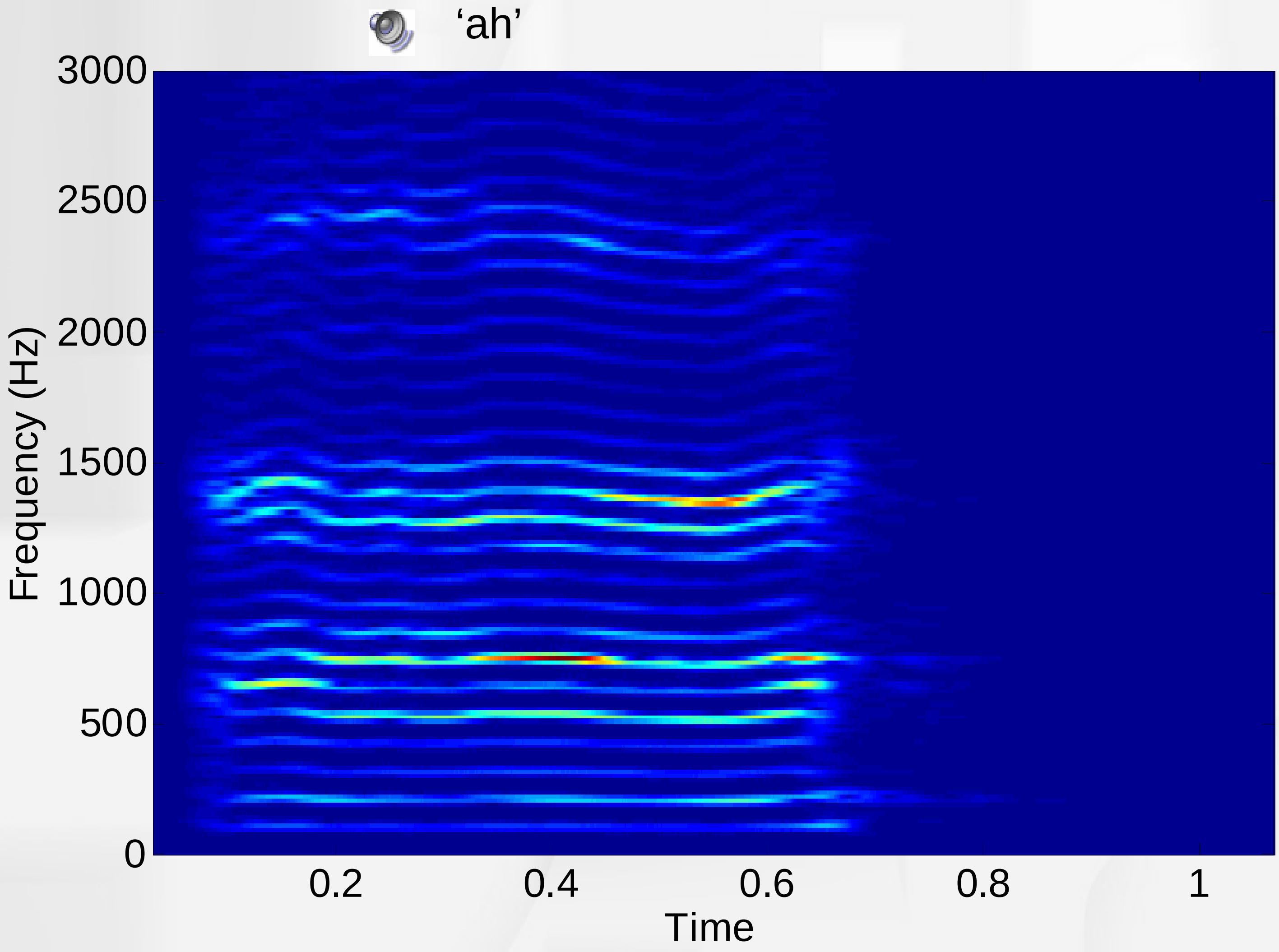
The Sound Spectrogram

my (lame) attempt at a 'frequency sweep'



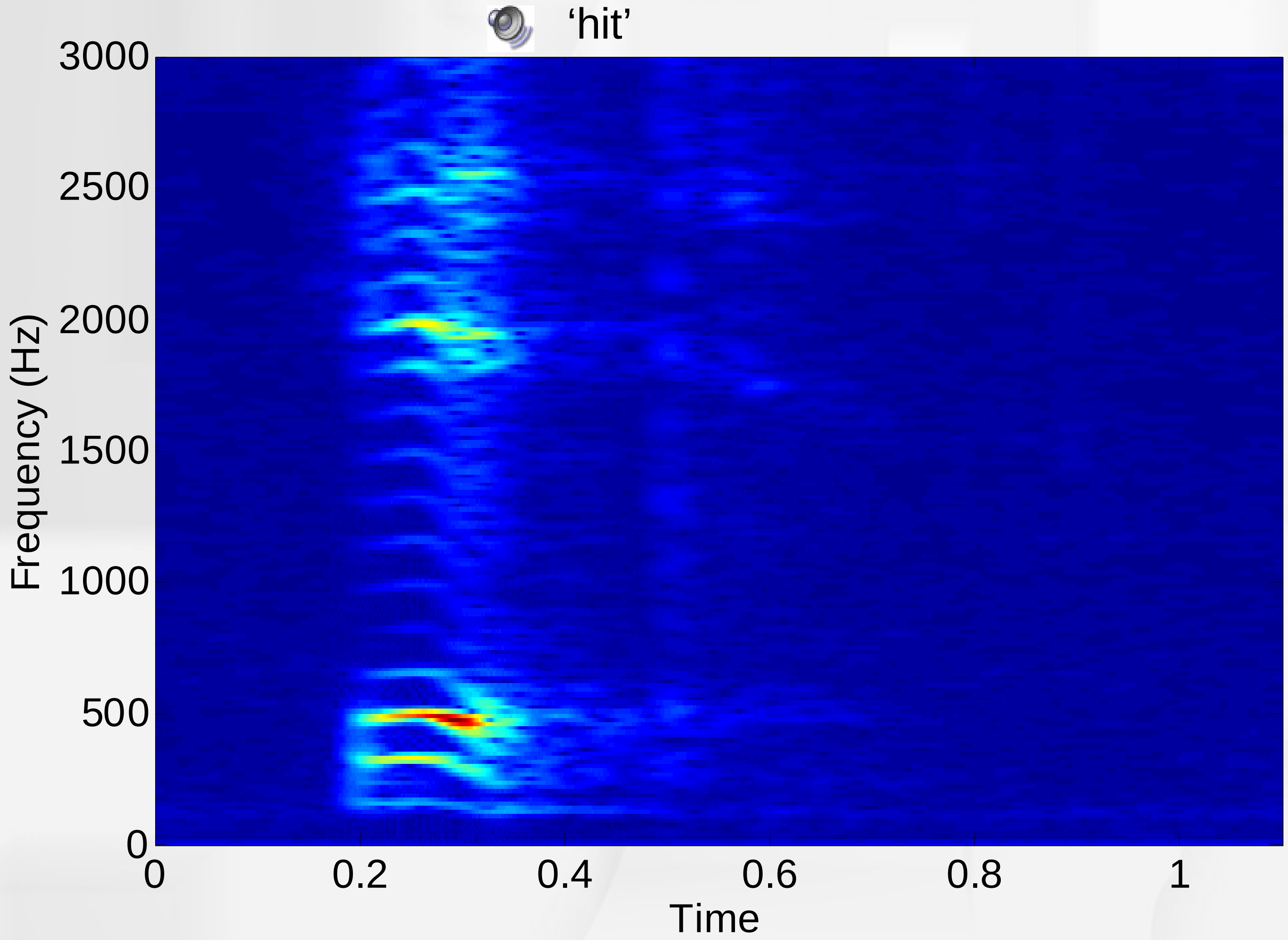
Vowel sounds are caused by a resonant frequency of the vocal cords and produce peaks in pressure at a number of frequencies called *formants*

The first formant has the lowest frequency, the second has the next highest, etc.



The Acoustic Signal

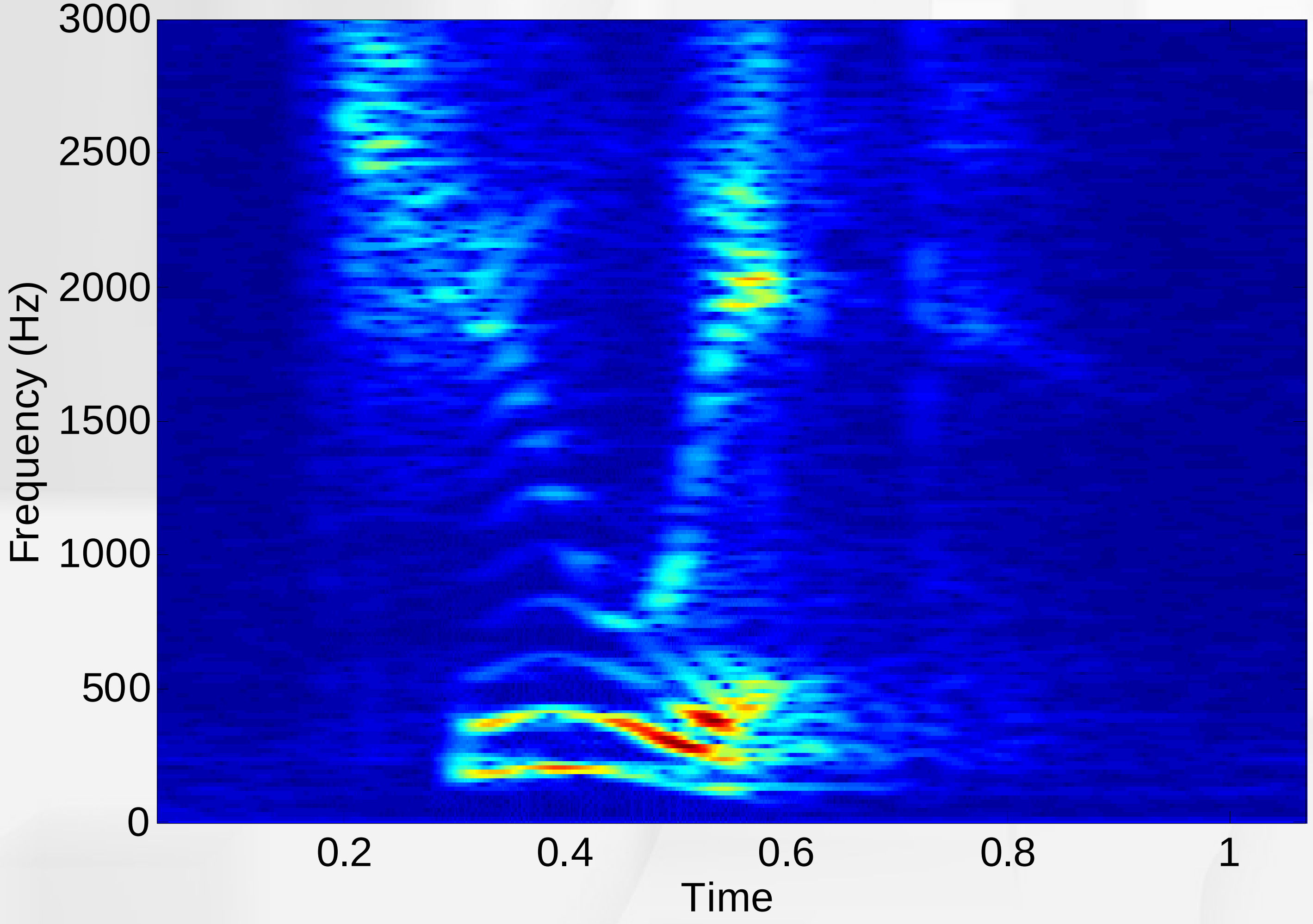
- Consonants are produced by a constriction of the vocal tract



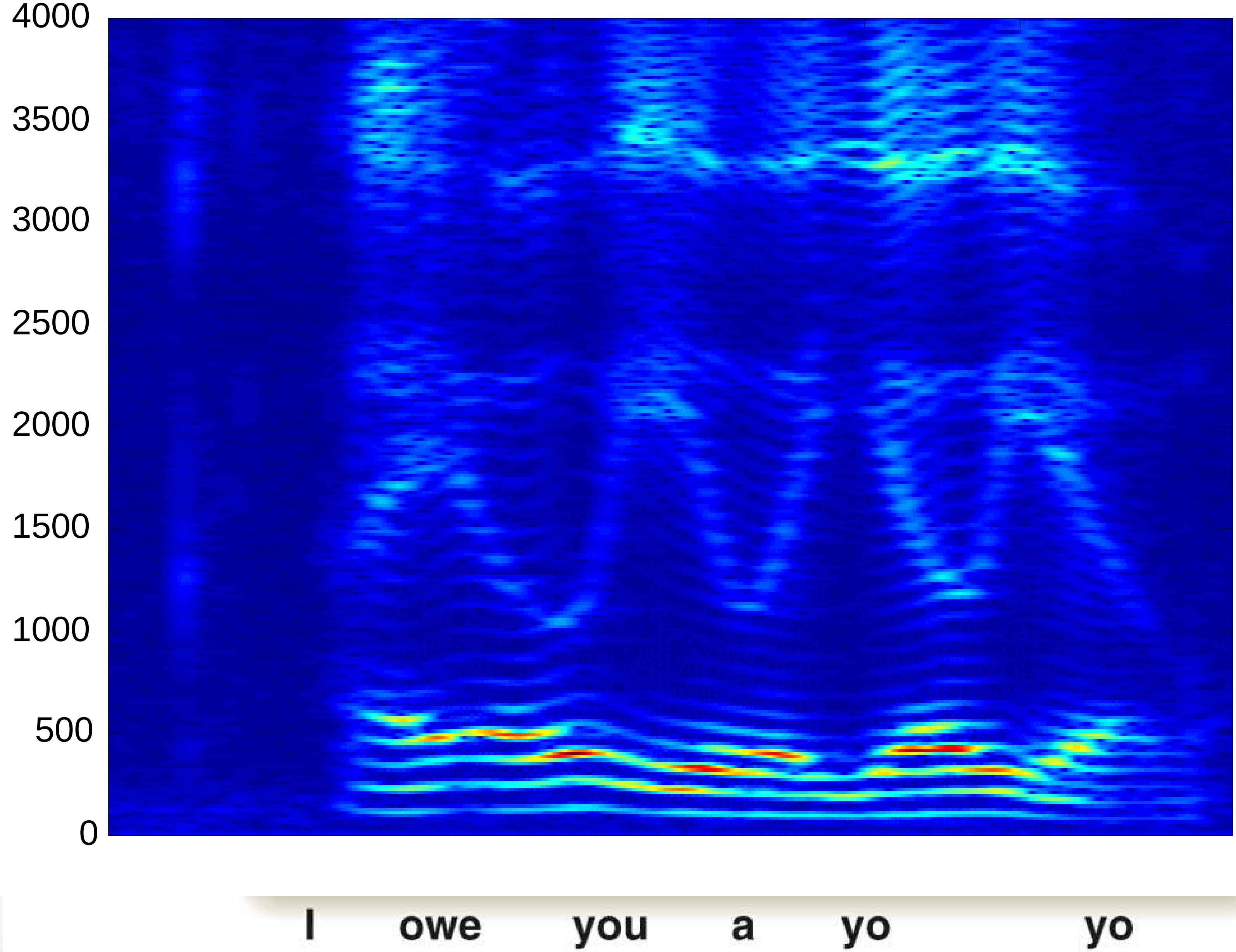
The segmentation problem:

There are no physical breaks in the continuous acoustic signal.

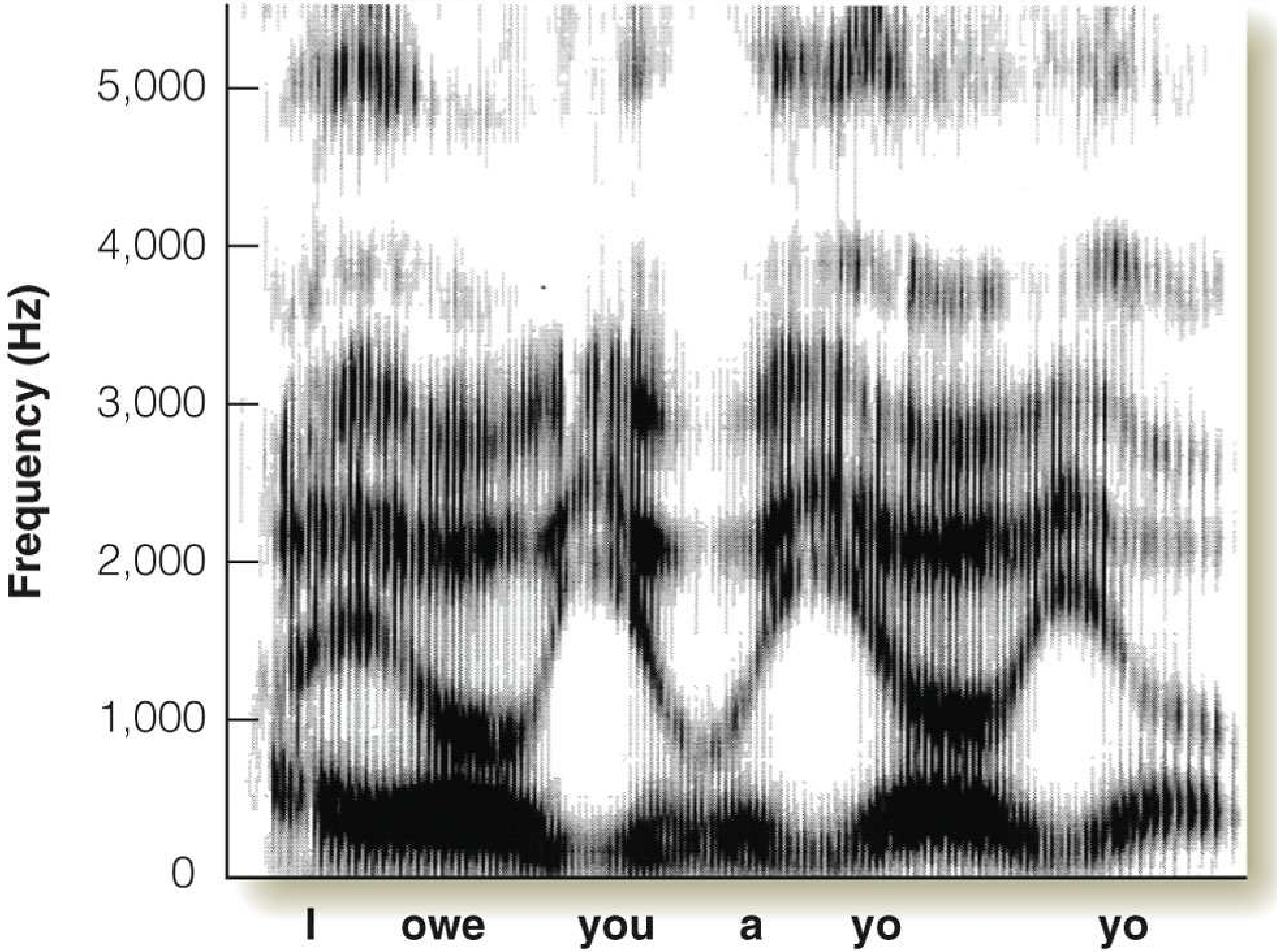
‘chew it’



The segmentation problem



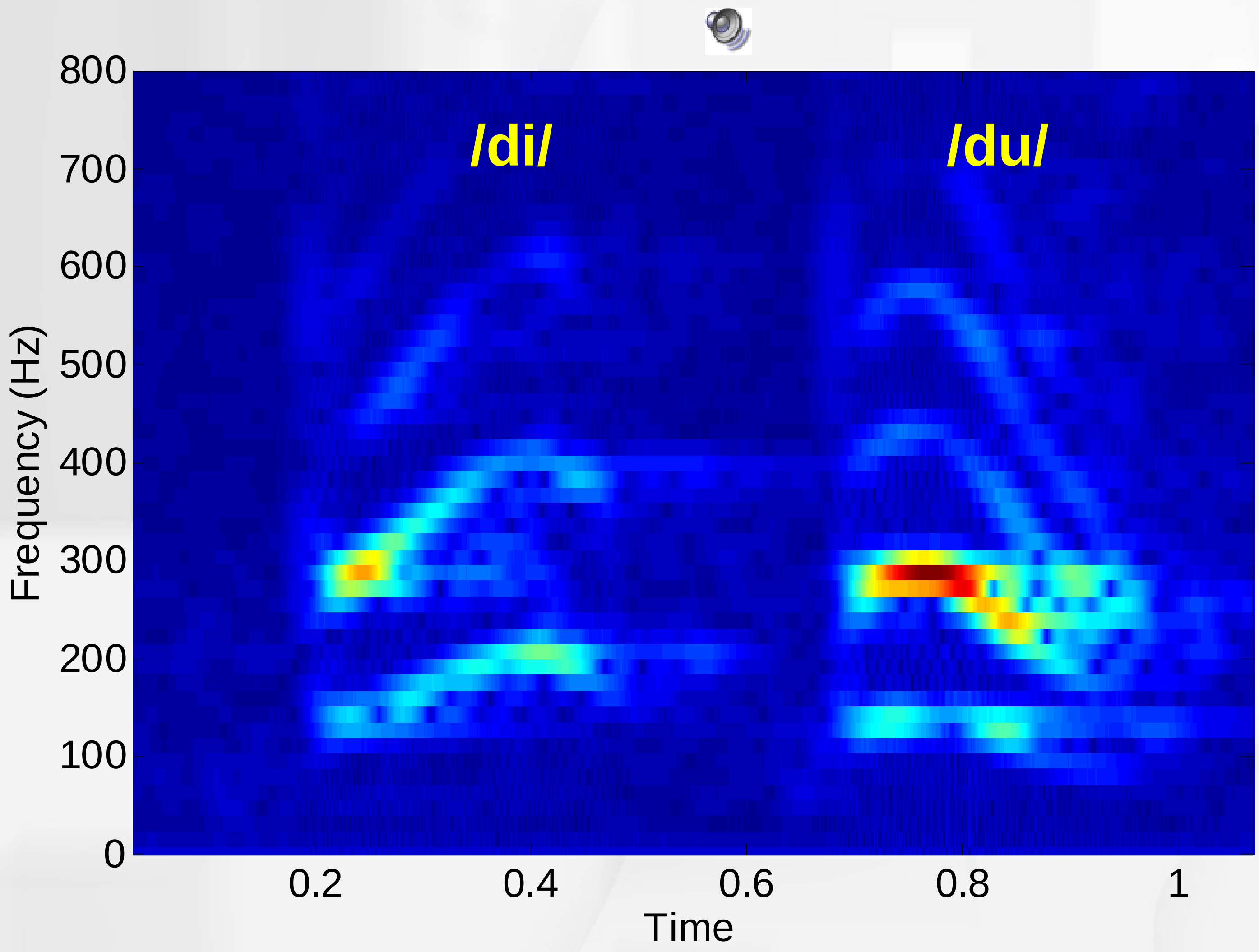
The segmentation problem



The variability problem

There is no simple correspondence between the acoustic signal and individual phonemes

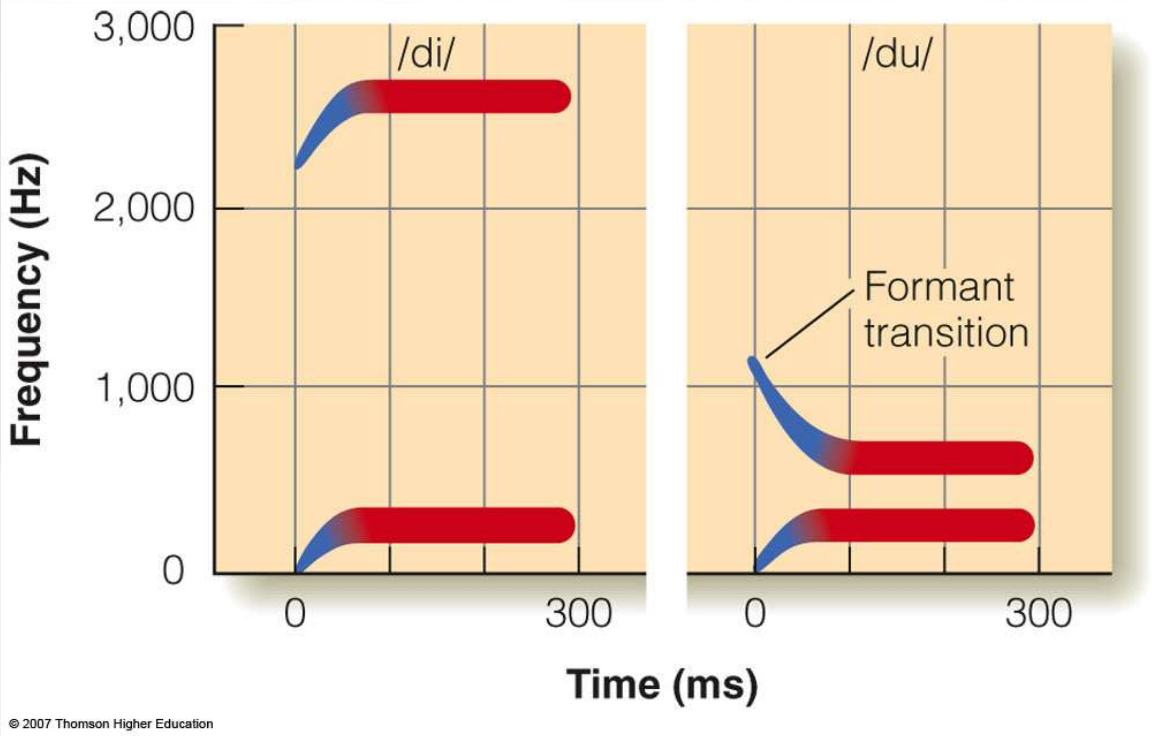
Coarticulation - overlap between articulation of neighboring phonemes



The variability problem

There is no simple correspondence between the acoustic signal and individual phonemes

1) Coarticulation - overlap between articulation of neighboring phonemes

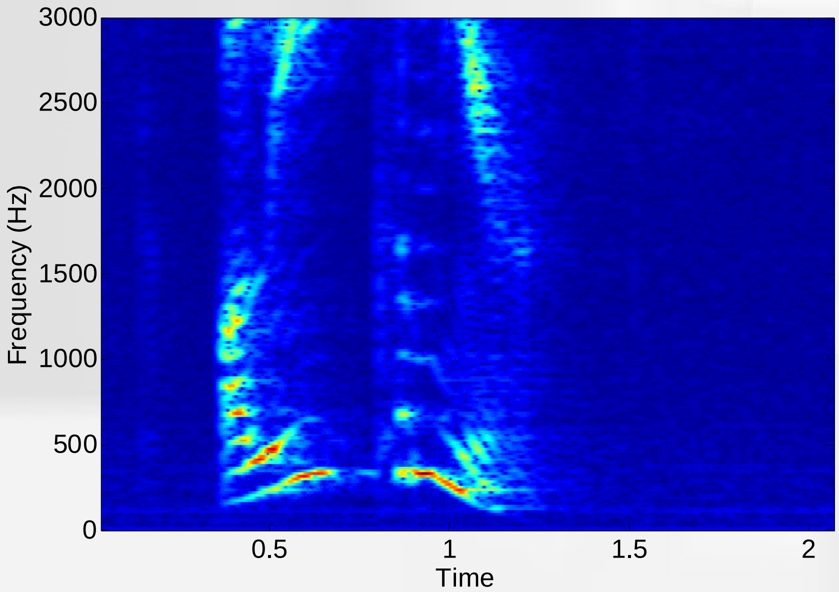


The variability problem

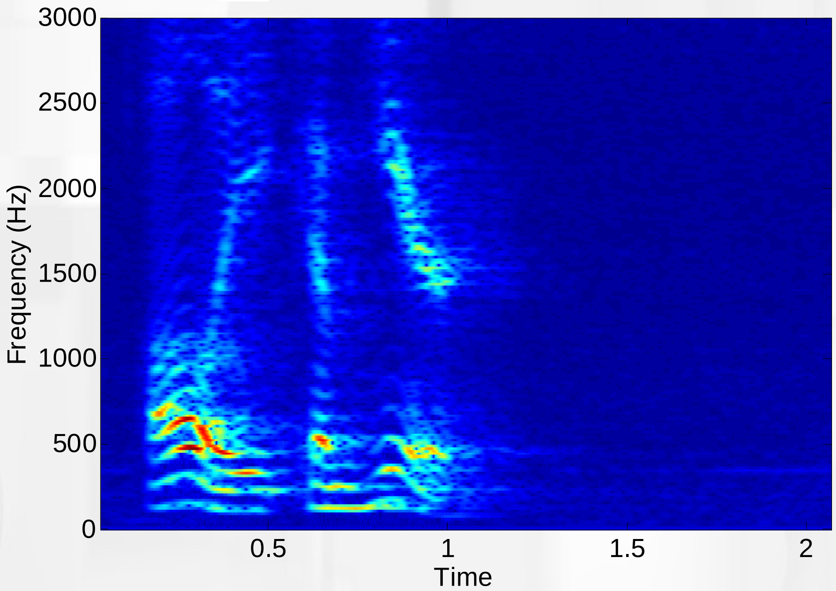
2) Variability across different speakers:

Speakers differ in pitch, accent, speed in speaking, and pronunciation

‘Ollie come here’ (Ione)

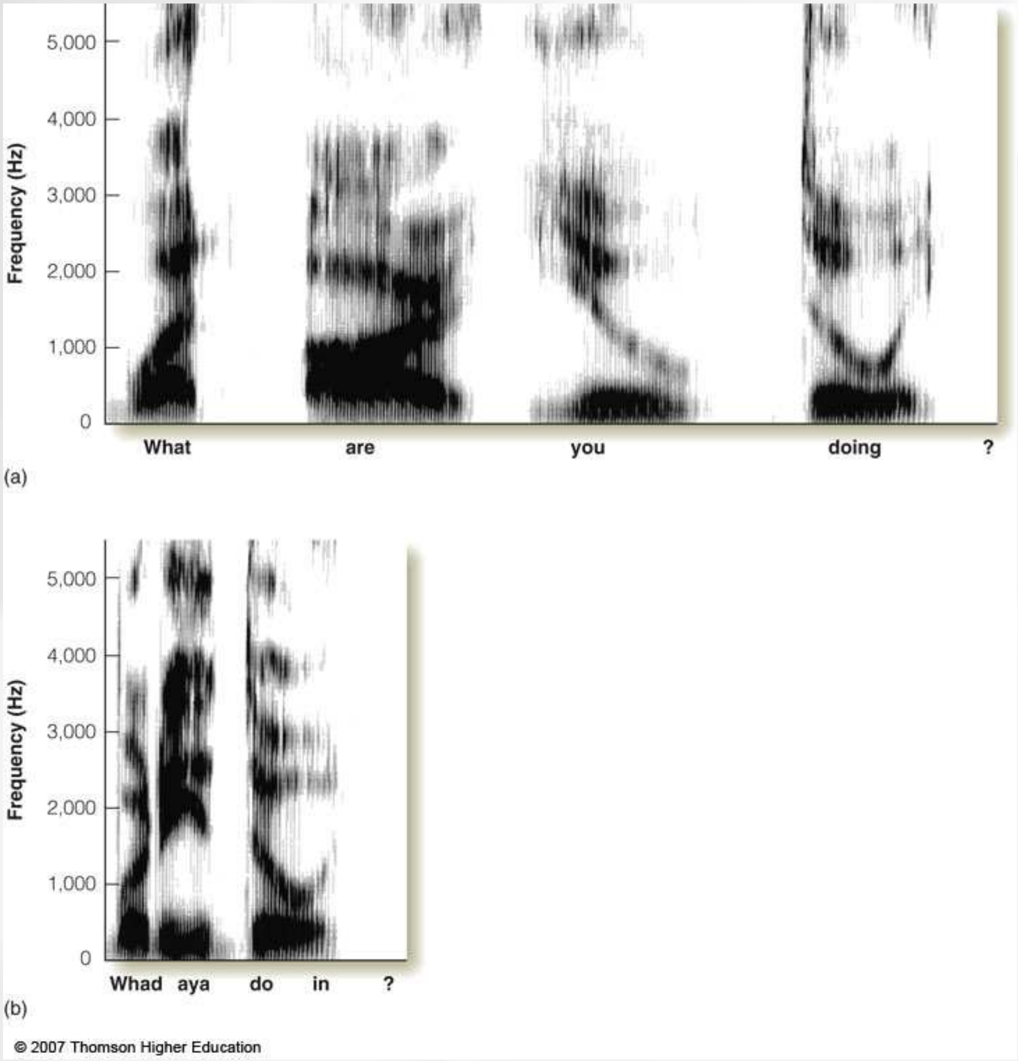


‘Ollie come here’ (Geoff)



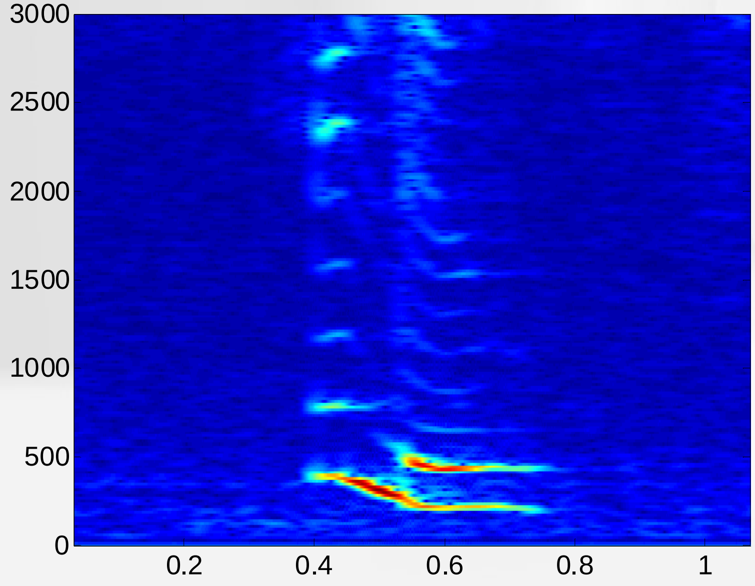
The variability problem

3) Different pronunciations have the same meaning, but very different spectrograms

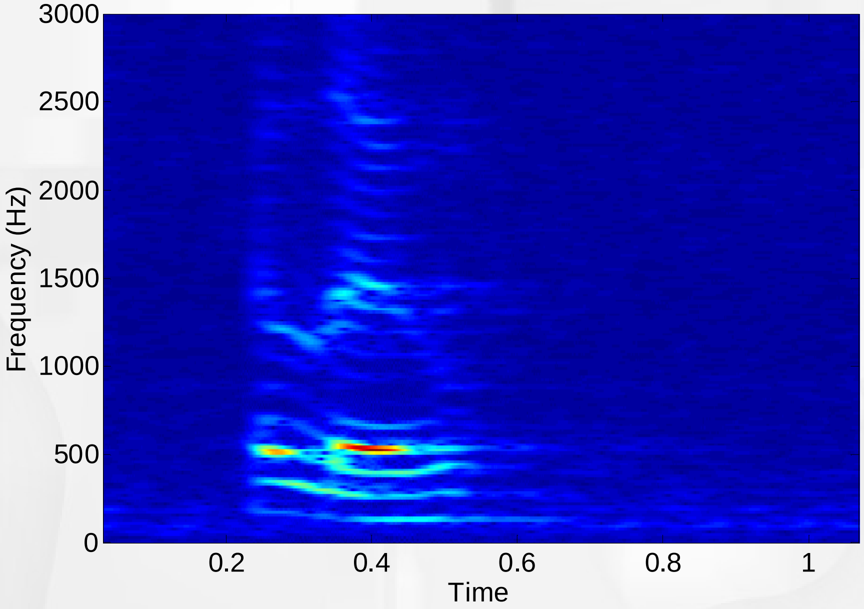


But there are some ‘invariances’ in speech perception.

‘hello’ (Ione)



‘hello’ (Geoff)



These spectrograms look similar.

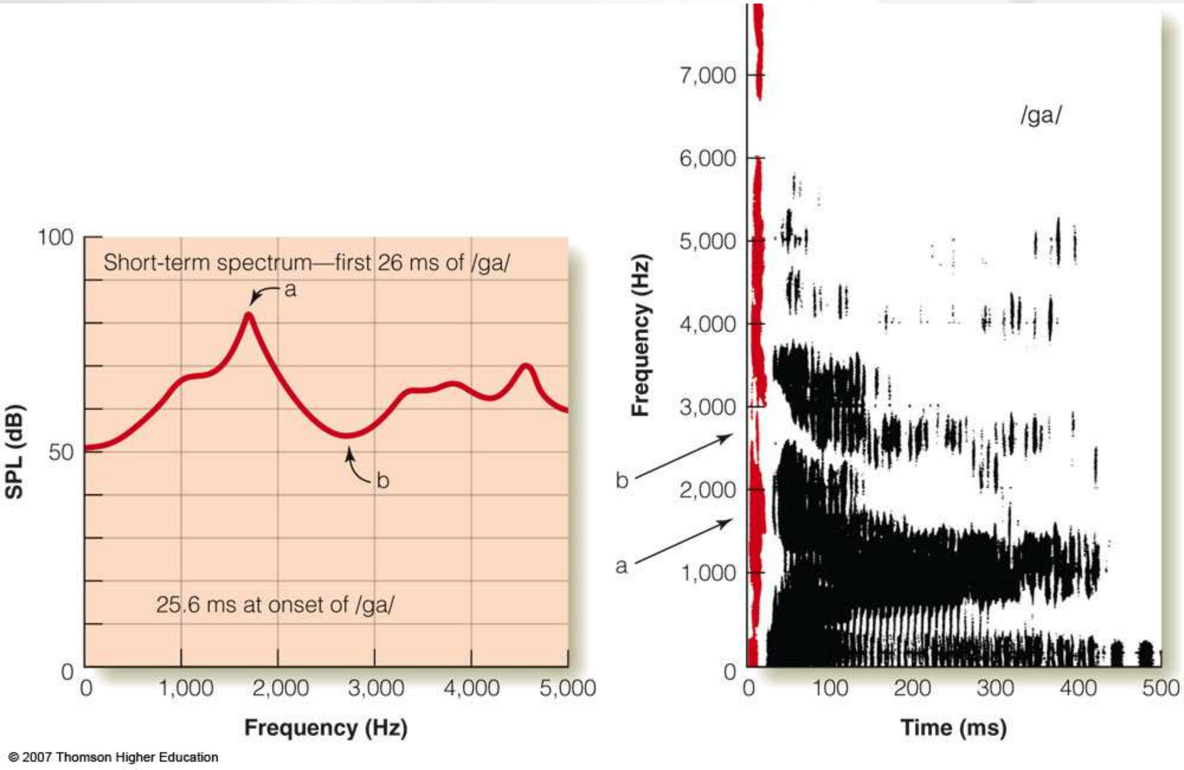
Invariant acoustic cues:

Some features of phonemes remain constant

Short-term spectrograms are used to investigate invariant acoustic cues.

Sequence of short-term spectra can be combined to create a running spectral display.

From these displays, there have been some invariant cues discovered



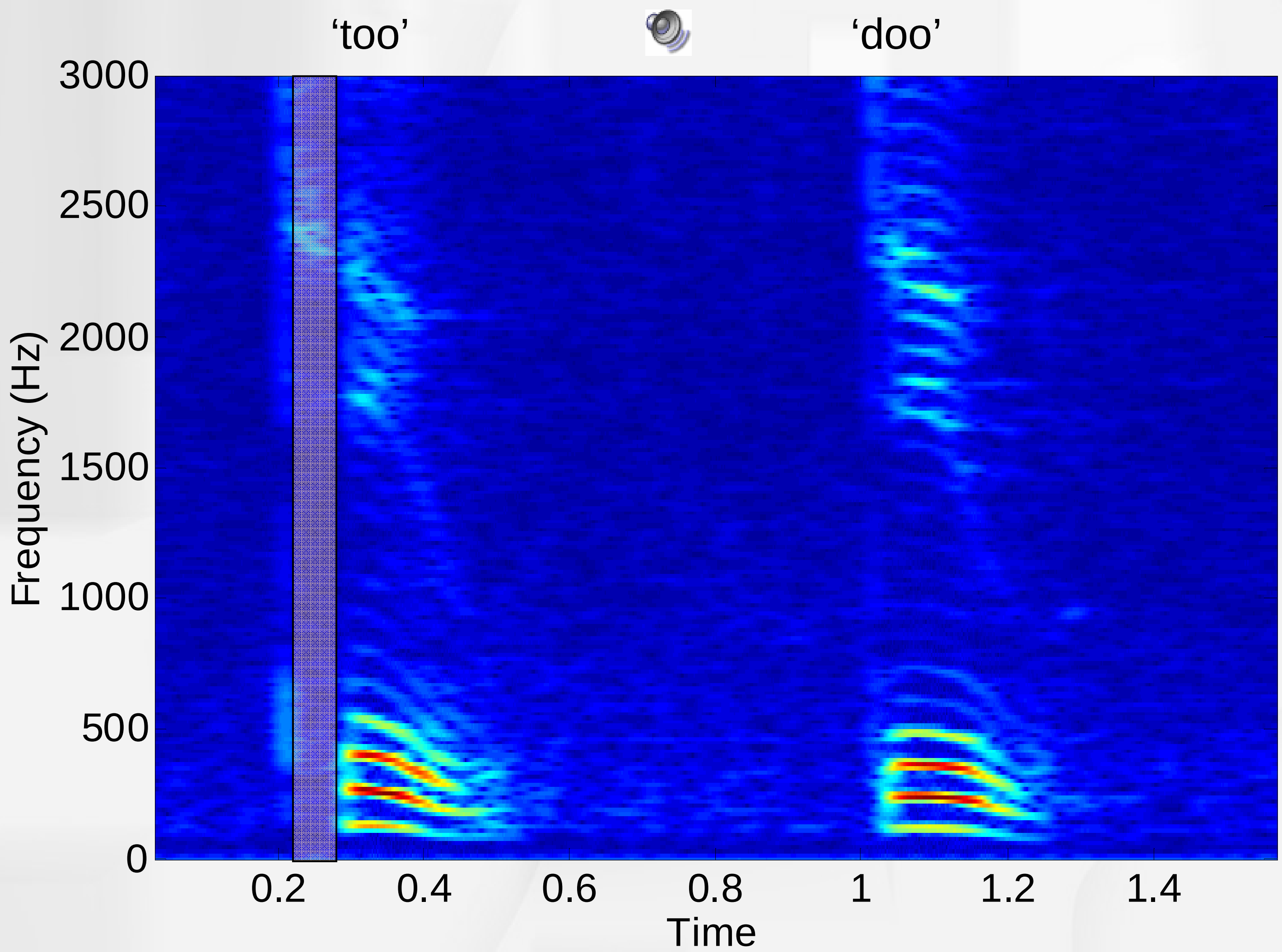
Categorical Perception

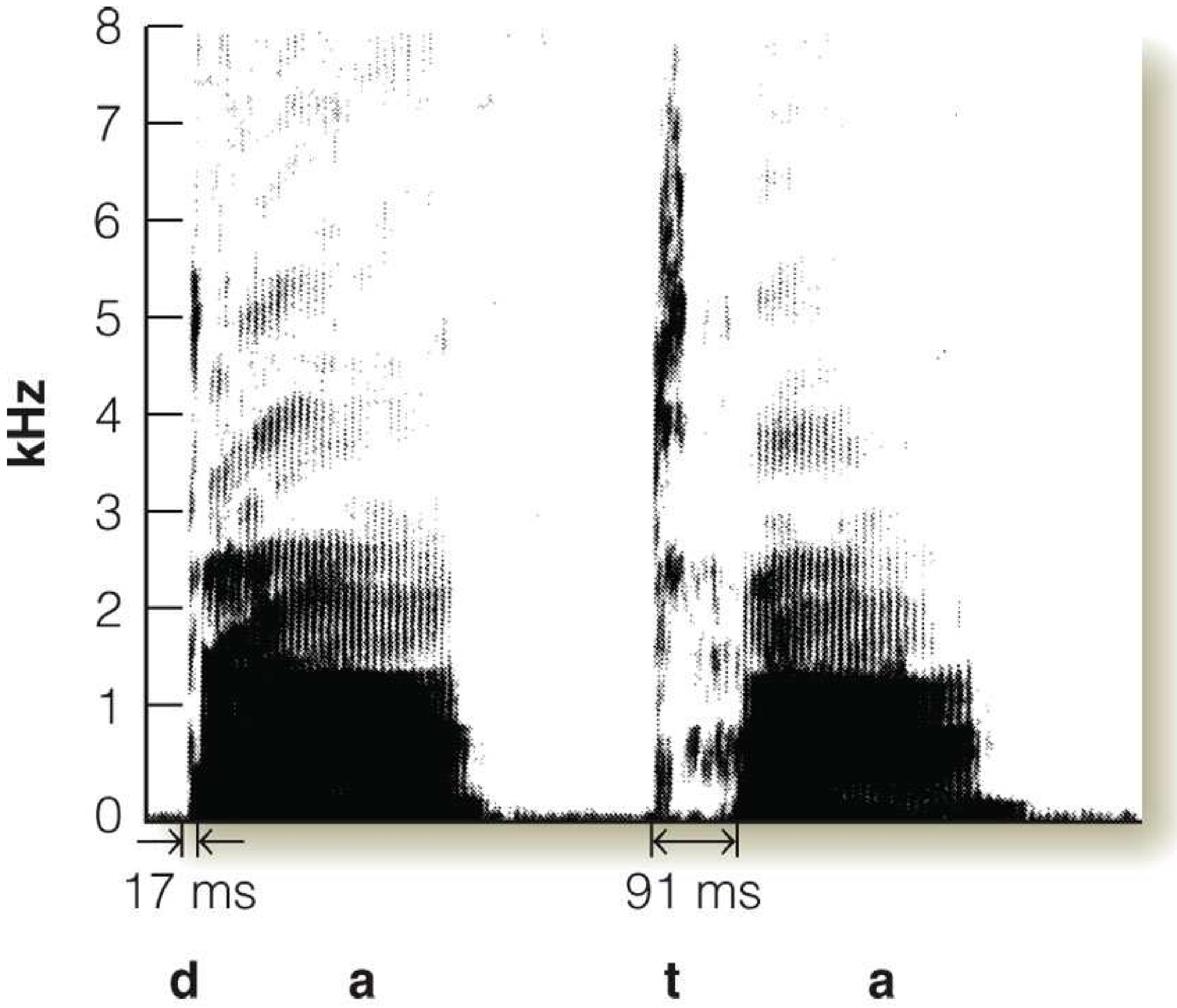
- This occurs when a wide range of acoustic cues results in the perception of a limited number of sound categories
- An example of this comes from experiments on voice onset time (VOT) - time delay between when a sound starts and when voicing begins
  - Stimuli are **da** (VOT of 17ms) and **ta** (VOT of 91ms)

Voice onset time (VOT)

Delay between when the sound begins and the onset of vocal cords.

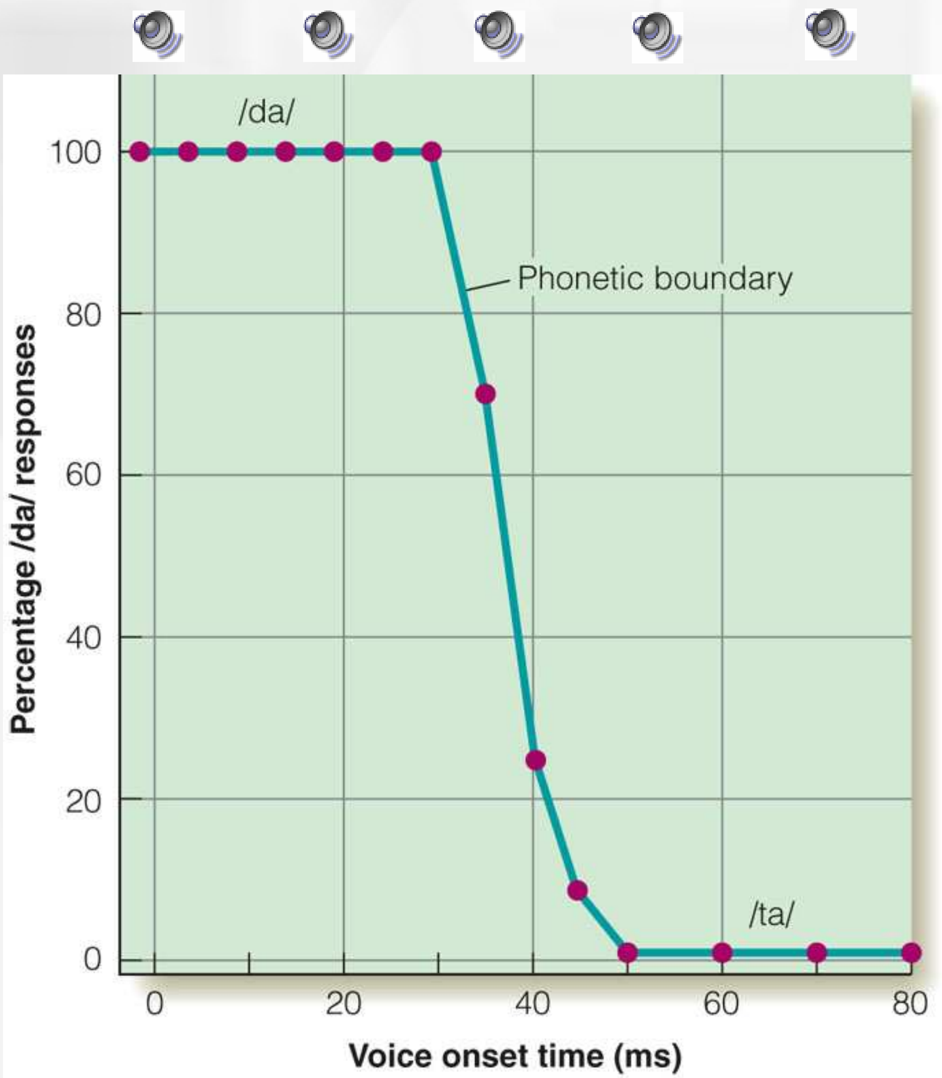
Distinguishes between ‘ta’ vs. ‘da’, and ‘pa’ vs. ‘pa’.

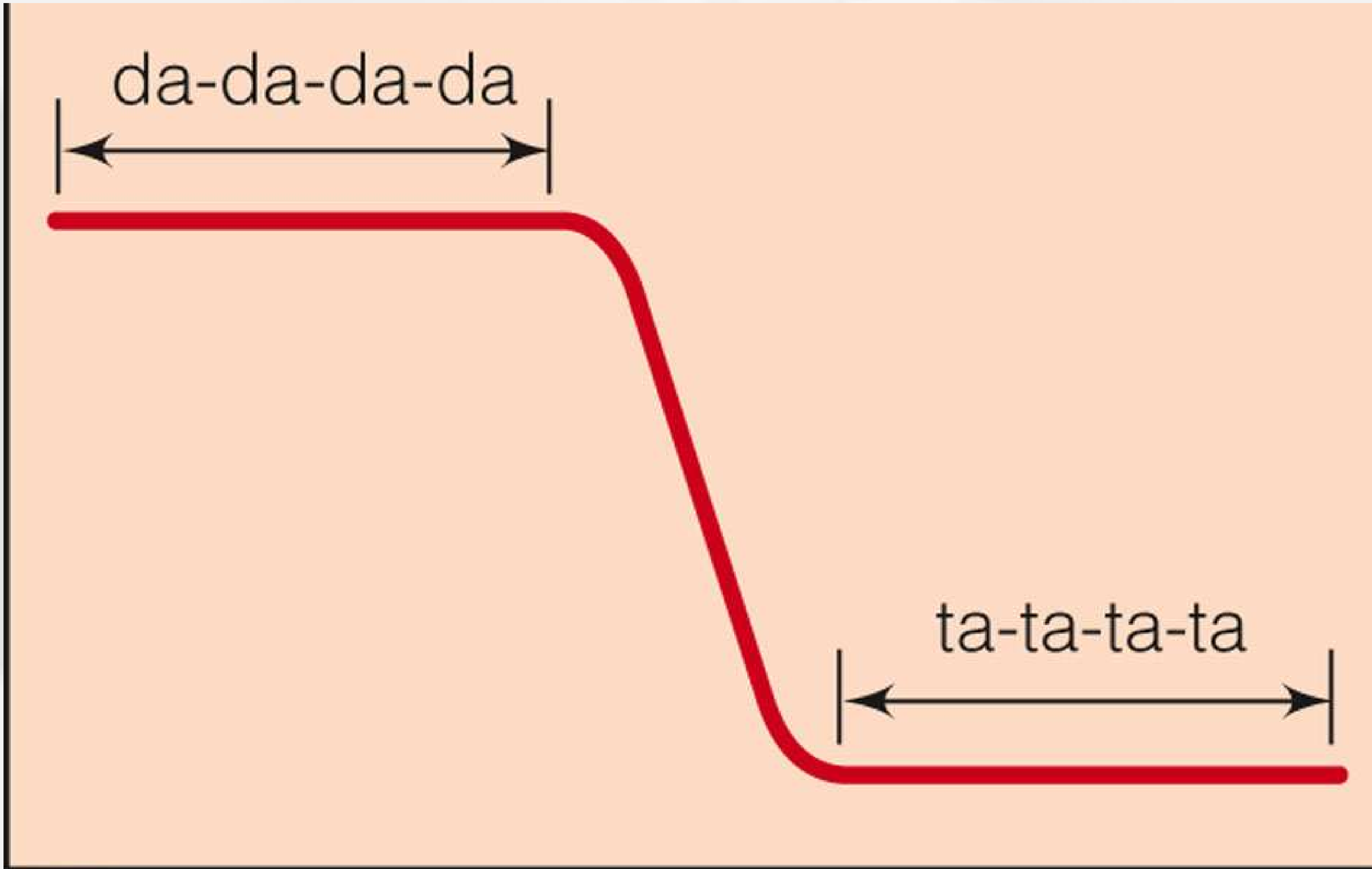



8
7
6
5
4
3
2
1
0
kHz
17 ms
91 ms
d
a
t
a

‘Categorical perception’

Despite the continuous variation of VOT, we only hear one phoneme or the other.





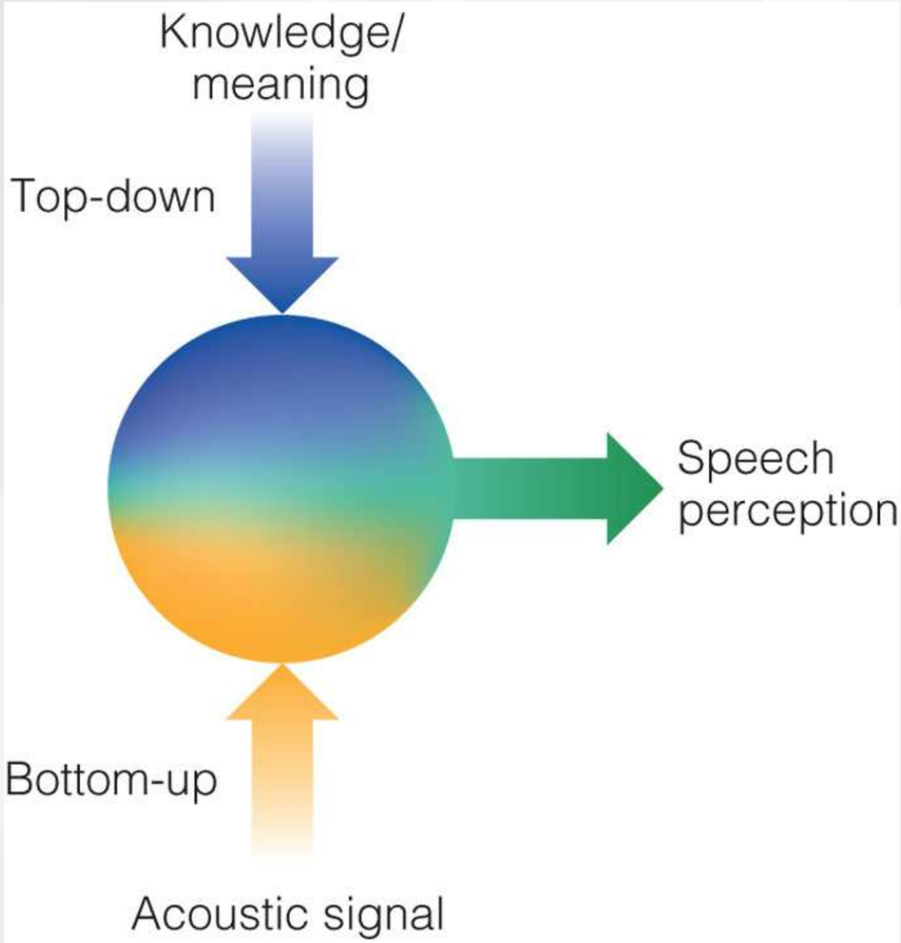
Voice onset time

Speech Perception is Multimodal

- Auditory-visual speech perception
  - The McGurk effect
    - Visual stimulus shows a speaker saying “ga-ga”
    - Auditory stimulus has a speaker saying “ba-ba”
    - Observer watching and listening hears “da-da”, which is the midpoint between “ga” and “ba”
    - Observer with eyes closed will hear “ba”

Cognitive Dimensions of Speech Perception

- Top-down processing, including knowledge a listener has about a language, affects perception of the incoming speech stimulus
- Segmentation is affected by context and meaning
  - I scream you scream we all scream for ice cream



Meaning and Phoneme Perception

- Experiment by Turvey and Van Gelder
  - Short words (sin, bat, and leg) and short nonwords (jum, baf, and teg) were presented to listeners
  - The task was to press a button as quickly as possible when they heard a target phoneme
  - On average, listeners were faster with words (580 ms) than non-words (631 ms)

Meaning and Phoneme Perception

- Experiment by Warren
  - Listeners heard a sentence that had a phoneme covered by a cough
  - The task was to state where in the sentence the cough occurred
  - Listeners could not correctly identify the position and they also did not notice that a phoneme was missing -- called the *phonemic restoration effect*

Meaning and Word Perception

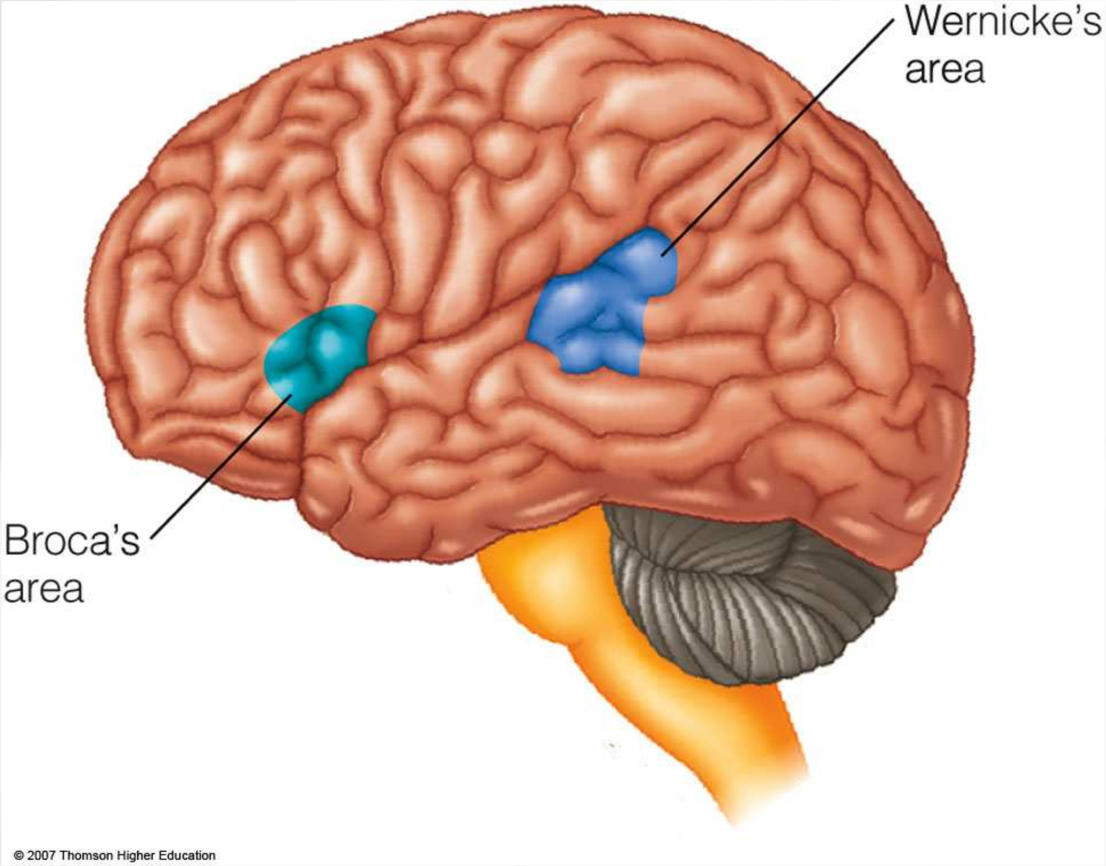
- Experiment by Miller and Isard
  - Stimuli were three types of sentences:
    - Normal grammatical sentences
    - Anomalous sentences that were grammatical
    - Ungrammatical strings of words
  - Listeners were to *shadow* (repeat aloud) the sentences as they heard them through headphones

- Results showed that listeners were
  - 89% accurate with normal sentences
  - 79% accurate for anomalous sentences
  - 56% accurate for ungrammatical word strings
  - Differences were even larger if background noise was present

Speech Perception and the Brain

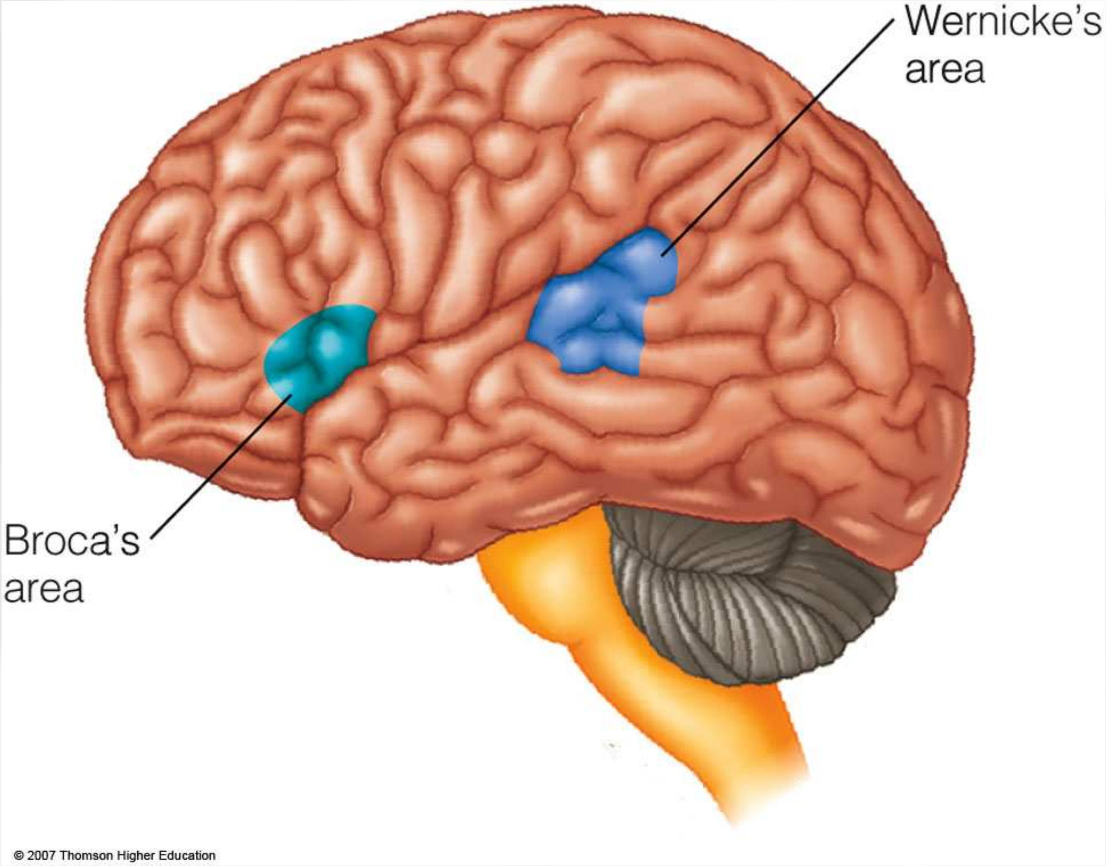
- **Broca's aphasia** - individuals have damage in Broca's area (in frontal lobe)
  - Labored and stilted speech and short sentences but they understand others

Affected people often omit small words such as "is," "and," and "the."



Wernicke’s aphasia - individuals have damage in Wernicke’s area (in temporal lobe)

Speak fluently but the content is disorganized and not meaningful
They also have difficulty understanding others



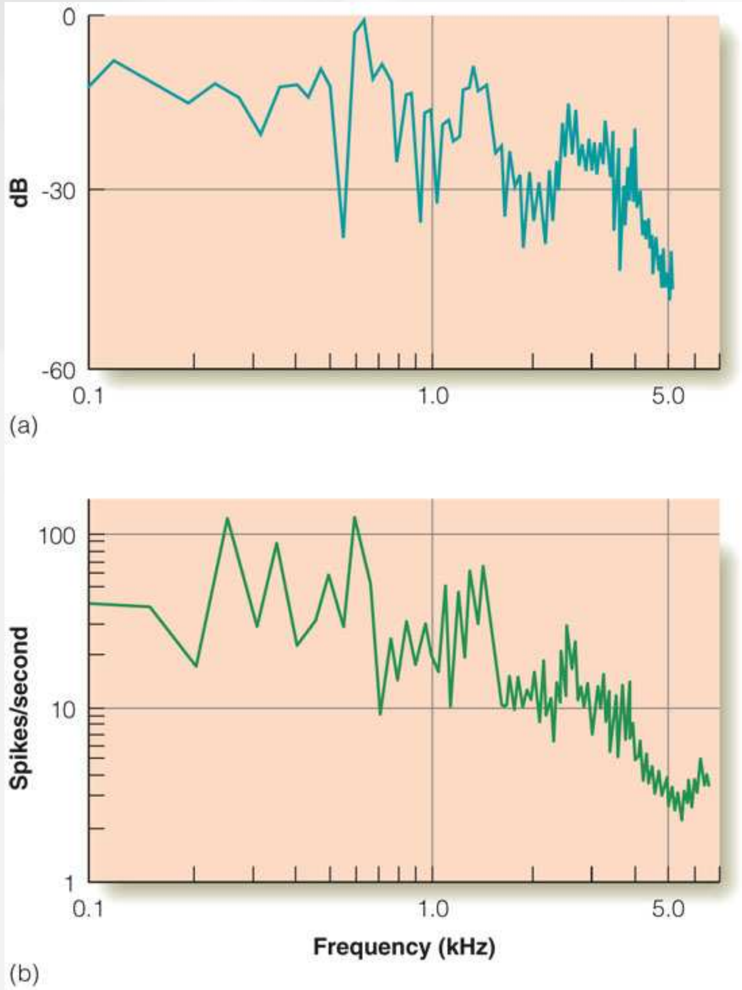
When trying to say: "The dog needs to go out so I will take him for a walk."

"You know that smoodle pinkered and that I want to get him round and take care of him like you want before,"

Speech Perception and the Brain

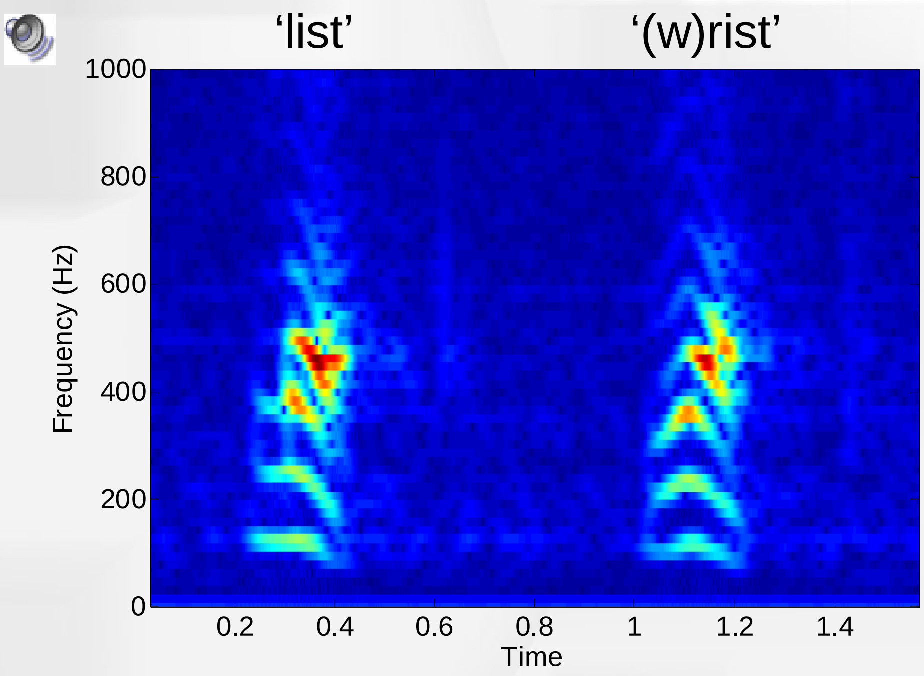
- Measurements from cats' auditory fibers show that the pattern of firing mirrors the energy distribution in the auditory signal
- Brain scans of humans show that there are areas of the human *what* stream that are selectively activated by the human voice

/da/



Experience Dependent Plasticity

- Before age 1, human infants can tell difference between sounds that create all languages
- The brain becomes “tuned” to respond best to speech sounds that are in the environment
- Other sound differentiation disappears when there is no reinforcement from the environment



Experience Dependent Plasticity

By adulthood, we are ‘tuned’ to recognize and produce only a subset of possible sounds.

Demonstration:

1) Record your voice
2) Play it backwards
3) Imitate and record the backward sounds
4) Play *that* backwards.

Why? Backward sounds contain sounds that aren’t normal (English) phonemes.

We can’t hear or produce these sounds properly.

Speech Perception is Multimodal

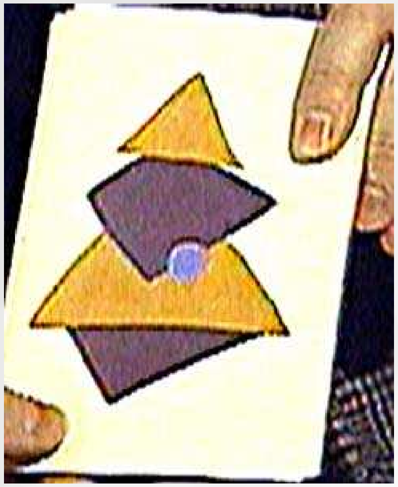
- Auditory-visual speech perception
  - The McGurk effect
    - Visual stimulus shows a speaker saying “ga-ga”
    - Auditory stimulus has a speaker saying “ba-ba”
    - Observer watching and listening hears “da-da”, which is the midpoint between “ga” and “ba”
    - Observer with eyes closed will hear “ba”

Speech Perception is Multimodal

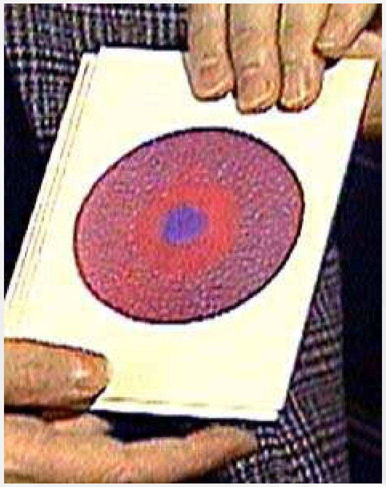
Demonstration from YouTube

Other sensory interactions: Synesthesia

music - color synesthesia, individuals experience colors in response to tones or other aspects of musical stimuli (e.g., timbre or key). Tone-color synesthetes often have perfect pitch.



Doorbell ringing



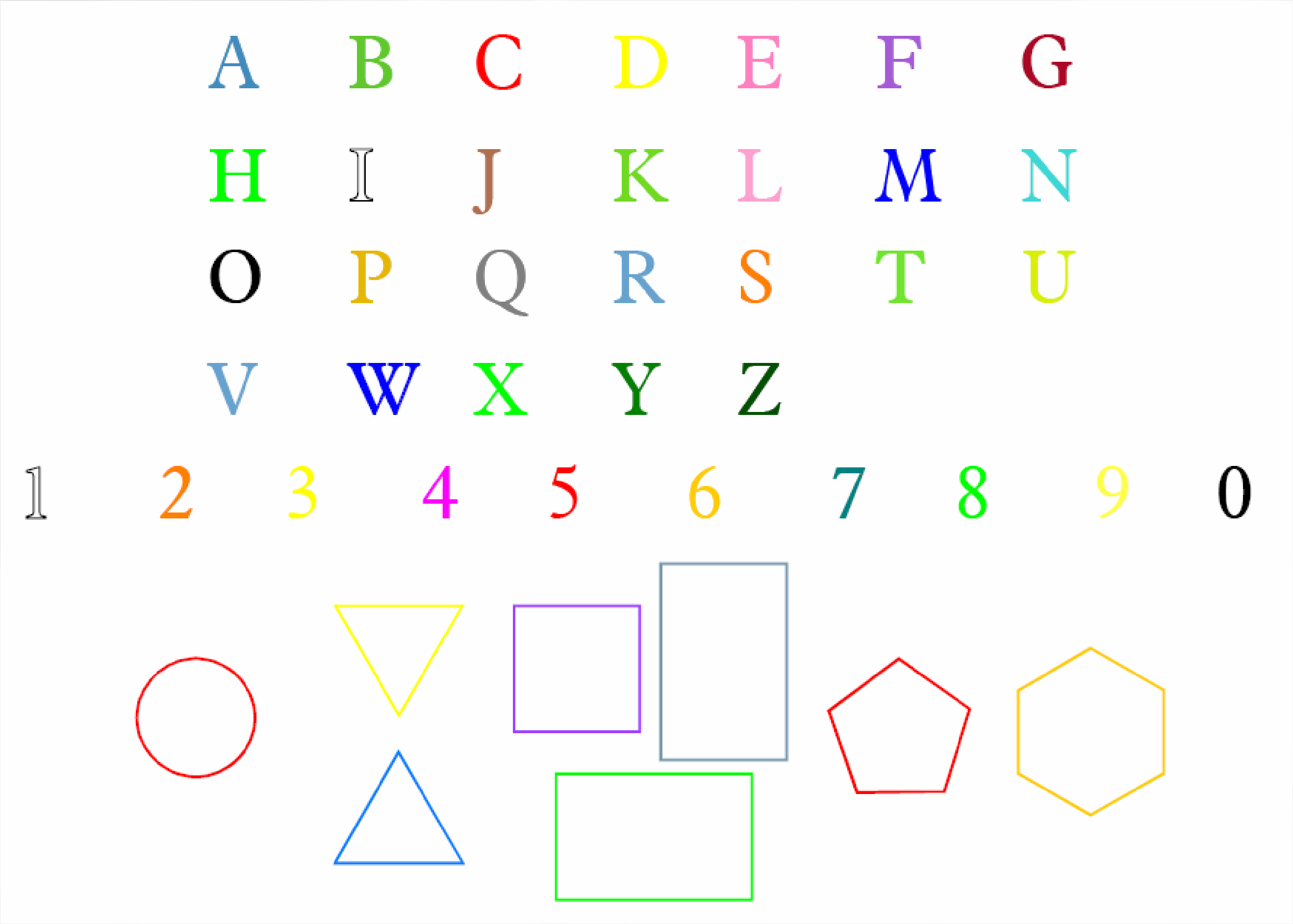
Dog barking

Artist Carol Steen's drawings of common sounds.

One individual's color and pitch perceptions :

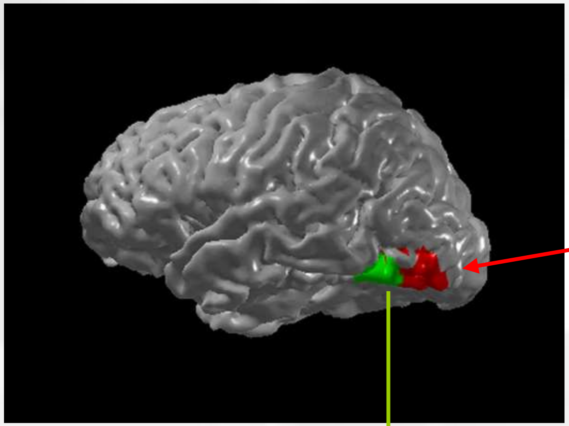
C- white
C# navy blue, somewhat metallic
D- gray-green
D# yellow-green; Eb gold, metallic
E- bright yellow
F- crimson red, tending toward magenta. Very vivid and rich.
F# maroon, a bit redder; Gb maroon, slightly darker with a metallic tone
G   brown-orange, browner the lower the note is.
G# orange-copper, not shiny, but bright. Ab metallic copper/brass.
A   orange
A# magenta; Bb a beautiful royal purple--more violet, reddish-purple hue
B   a very crisp black.

grapheme- color synesthesia: letters or numbers are perceived as inherently colored



Other sensory interactions: Synesthesia

grapheme- color synesthesia: letters or numbers are perceived as inherently colored



Area V4
(color processing)

Visual word-form area

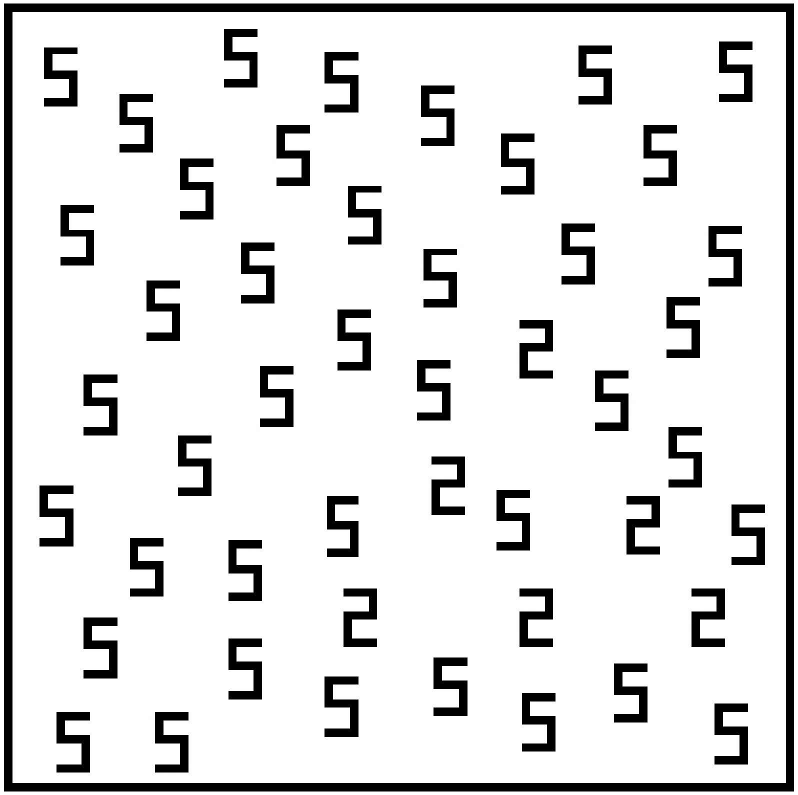
fMRI responses to letters invoke responses in V4 for synesthetes

The Stroop effect: it is difficult to override the written meaning of the word when naming the color of the text.

BLUE	GREEN	YELLOW
PINK	RED	ORANGE
GREY	BLACK	PURPLE
TAN	WHITE	BROWN

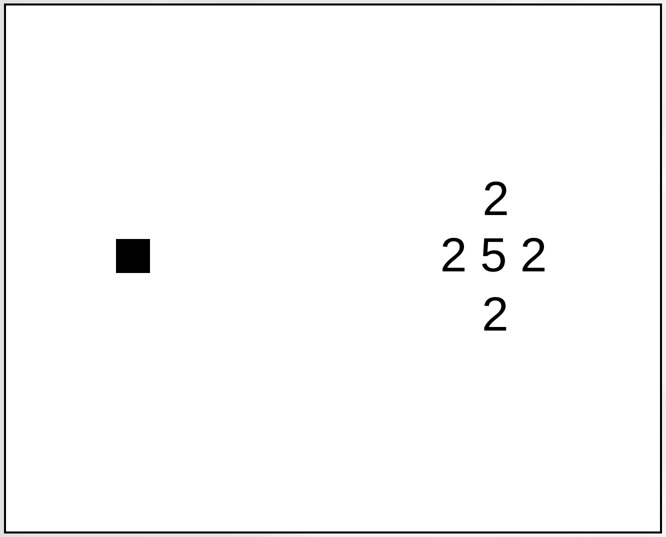
Grapheme-color synesthetes suffer from the Stroop effect with black letters on a white background.

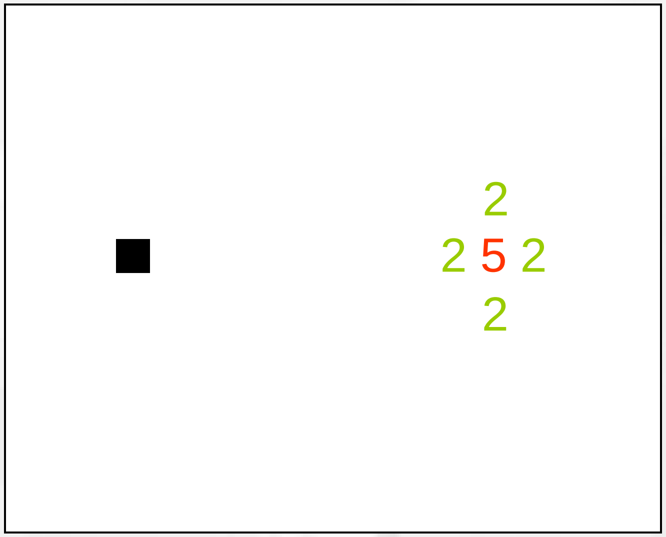
Ramachandran and Hubbard showed that grapheme-color synesthetes are faster at finding the triangle of ‘2’s imbedded in the background of ‘5’s



Crowding task: when placed in the periphery, it is difficult to identify the center number when surrounded by other numbers.

But if the center number is a different color, it is easier to identify.





Given black letters on a white background, grapheme-color synesthetes identify the center number faster and more accurately than control subjects.

Number - form synesthesia: numbers, months of the year, and/or days of the week elicit precise locations in space (for example, 1980 may be "farther away" than 1990), or may have colors, or have a three-dimensional view of a year as a map (clockwise or counterclockwise).

January, February, March, April, May, June, July, August, September, October, November, December.

Lexical - gustatory synesthesia In a rare form in which words and phonemes of spoken language evoke the sensations of taste in the mouth.

Table 1

Examples from JIW's inducer words (to the left of the arrow) and concurrent tastes (to the right of the arrow) that overlap in semantics (*Lexical-semantic*) or phonology (*Lexical-phonological*), or that are mediated by another word or concept (*Indirect lexical links*)

Lexical-semantic	Lexical-phonological	Indirect lexical links
Blue → "inky"	*Virginia* → "vinegar"	*Crease* → "lard" (via grease?)
Brown → "marmite"	*Barbara* → "rhubarb"	*Shop* → "lamb fatty" (via chop?)
Bar → "milk chocolate"	*Sydney* → "kidney"	*Six* → "vomit" (via sick?)
Can → "bitter flat beer"	*Auction* → "Yorkshire pudding"	*Human* → "baked beans" (via being?)
Newspaper → "chips"[a]	*April* → "apricots"	*Trust* → "smooth crusted bread" (via crust?)
Baby → "jelly babies"	*Made* → "marmalade"	*Speak* → "bacon" (via streaky?)

[a] In the UK, chips (fries) are traditionally eaten out of newspaper.

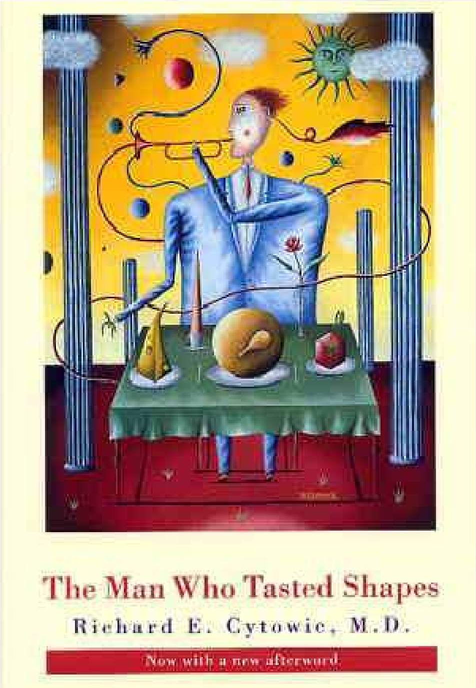
When coloured sounds taste sweet

Table 1 **Tastes triggered by tone intervals**

Tone interval	*Taste experienced*
Minor second	Sour
Major second	Bitter
Minor third	Salty
Major third	Sweet
Fourth	(Mown grass)
Tritone	(Disgust)
Fifth	Pure water
Minor sixth	Cream
Major sixth	Low-fat cream
Minor seventh	Bitter
Major seventh	Sour
Octave	No taste

seventh are both rated as bitter.

Taste – shape synesthesia: flavors invoke the perception of 3-dimensional shapes.



Includes the chapter: *"not enough points on the chicken"*

Face-color synesthesia: colors associated with individual faces. Could be the basis of why some people perceive 'auras'.

Subjective reports of synesthesia

For Patricia Duffy, a 46-year-old instructor in the United Nations' language and communication training program, the cause of her perceptions is less important than the richness they have brought to her life. She sees the words she speaks fly by in a rainbow of colors. She sees a year as an oblong circle, a week as a sidewalk with seven colored squares of pavement. The month of January is garnet red; December is dark brown. "I don't really know where it comes from," she said. "I just know it's always been that way."