Lecture 15: Hearing and Speech

Outline

I. Introduction (computational theory)

What information do we extract from sound?

What are the physical properties of sound?

Why is extracting information from sound computationally challenging?

Invariance problems: same source produces diff sounds

Ill-posed problem: cocktail party problem, reverb

II. Speech Perception

What is the structure of speech sounds?

Phonemes, formants, consonants & vowels

Why speech perception is computationally challenging

III. The auditory processing pathway

Peripheral transduction of sound & the cochlea (bare basics)

Primary auditory cortex

tonotopic organization

linear spectrotemporal filters

Speech-selective cortex

Just by listening, we can….

- *Identify the scene we are in and what is going on in it*
- *Localize* those events/people/objects
- *Recognize* them, e.g.:

  environmental sounds

  speech (what is being said)

  voices (who is saying it)
- *Selectively attend* to 1 source among many (“cocktail party effect”)
- Enjoy music
- Determine what things are made of…..

all this, just from variations in pressure arriving at the ear!

What is Hearing Good For?

- Detecting sounds, e.g. a but hidden dangerous animal, a friend calling out to you, a car horn
- Recognizing sounds
  environmental sounds, voices, speech
- Localizing sounds
- Selectively attending to one sound source
- Enjoying music
- Determine what things are made of

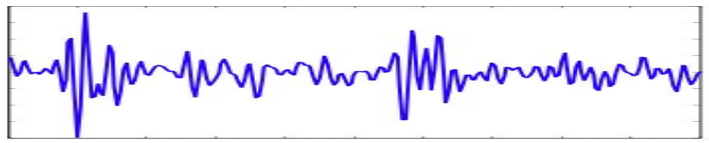
How do we do all this?

How do we find out how hearing works?

- Characterize is the input, & the physics of sound
  what would be involved in getting a machine to perform the above
  what cues are in the stimulus,
  what are key computational challenges &which problems are ill-posed?
- Characterize auditory abilities behaviorally
- Measure neural responses

Hearing: Extracting Information from Sound

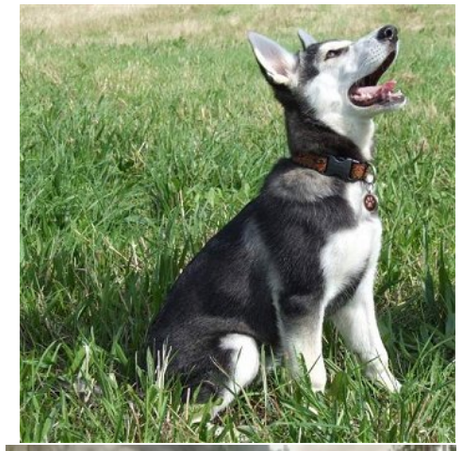
Sound Input



Now that we have some idea what sound is, let's think about how we extract information from it.

?

Why is this computationally challenging?



Several challenges in extracting information from sound:

1. **Invariance** problems:

A given sound source sounds different in different situations

e.g. diff people saying the same word

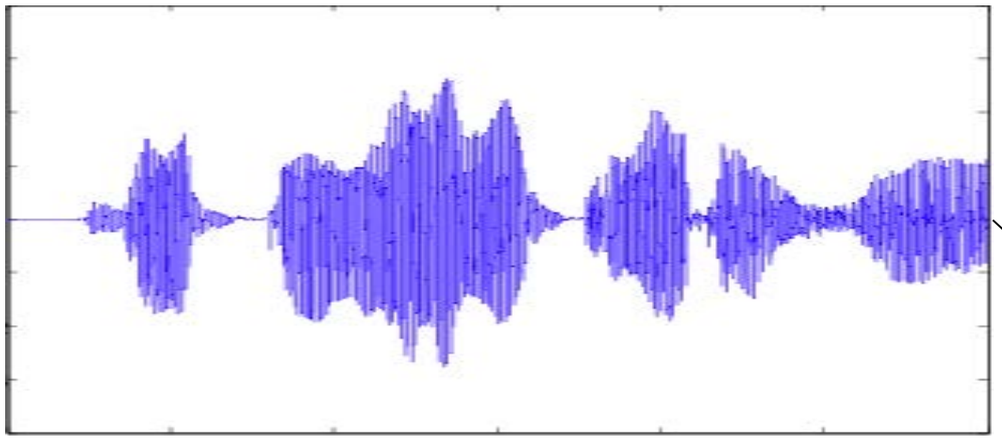
the same voice saying different things.

We need to appreciate the sameness across these diffs.

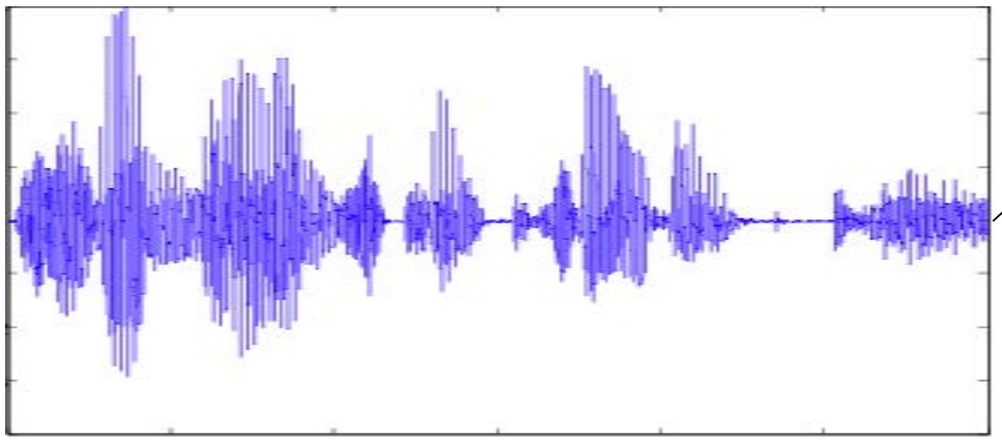
What else?

Cocktail Party Problem:
ear receives mixture of sources, which add linearly:

Source 1

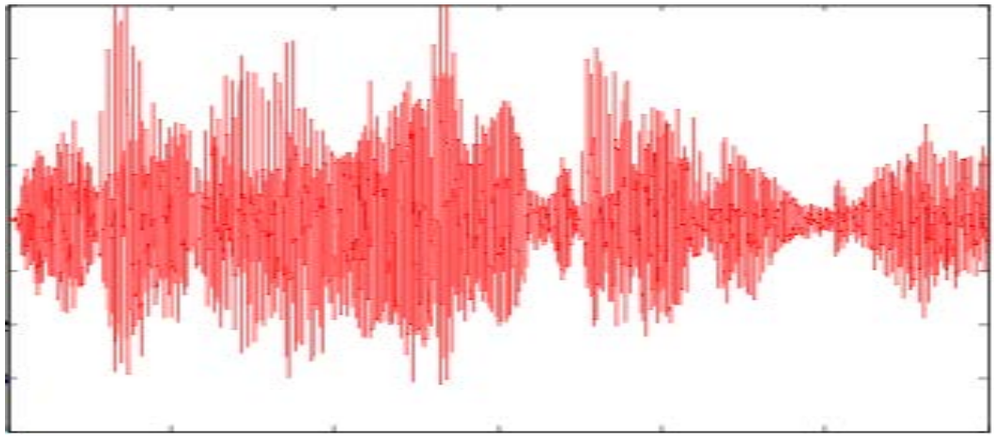


Source 2



\+ =

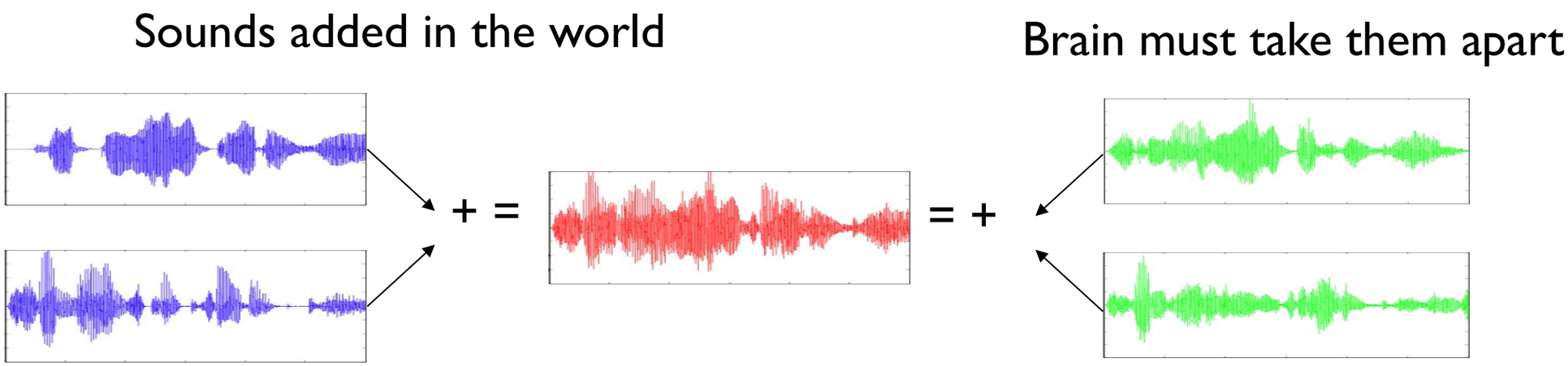
Mixture



BUT: The listener is usually interested in individual sources, which must be inferred from the mixture.

slide adapted from Josh McDermott

Cocktail Party Problem:
ear receives mixture of sources, which add linearly:



The problem is *ill-posed*.
Like: X + Y = 9, now solve for X and Y
many solutions, how do we know which is right?

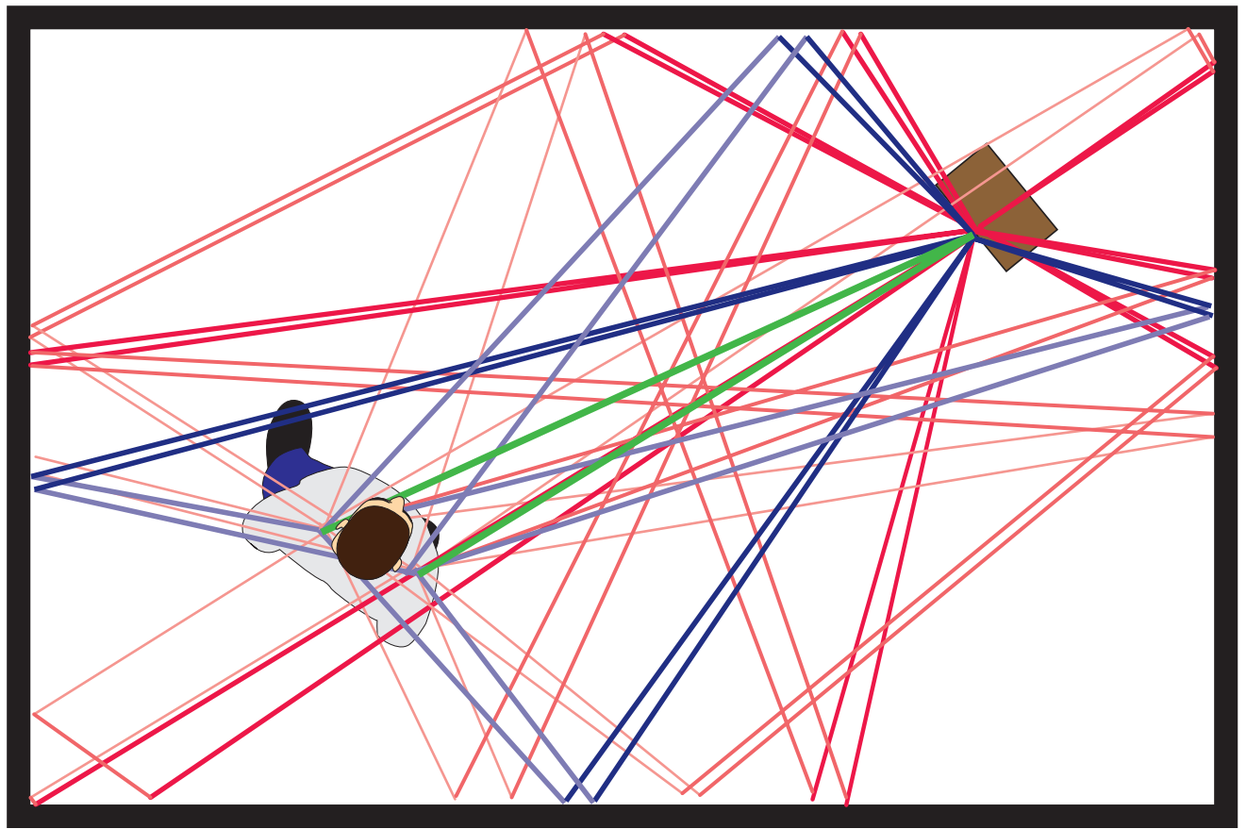
A classic problem in audition (studied by McDermott and many others).
Can be solved only by using knowledge of the properties of natural sound.s

Another challenge in inferring sources from auditory information…..

slide adapted from Josh McDermott

Reverberation

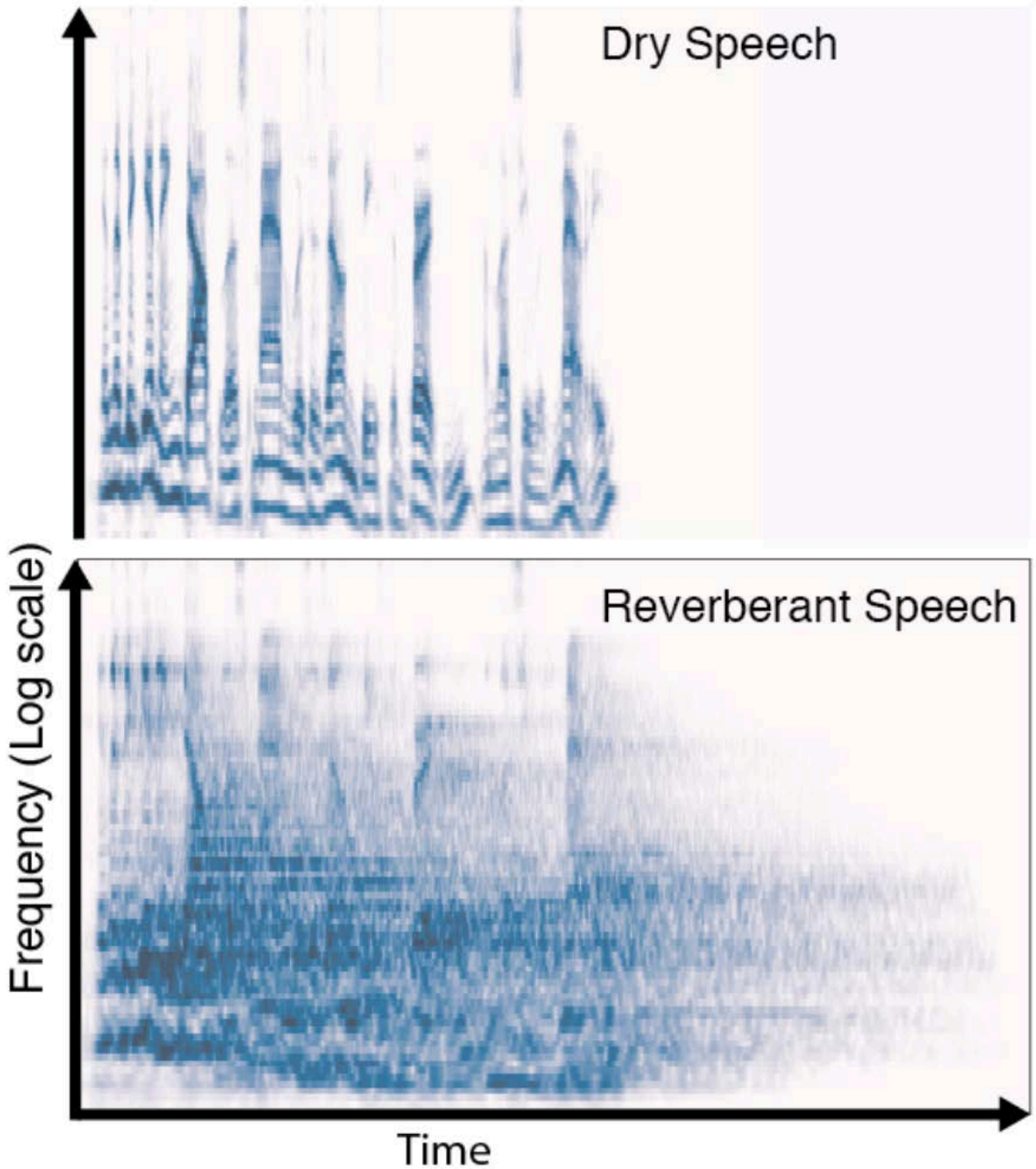
Sound sources interact with environment on way to ear:



Culling & Akeroyd, 2010

slide adapted from Josh McDermott

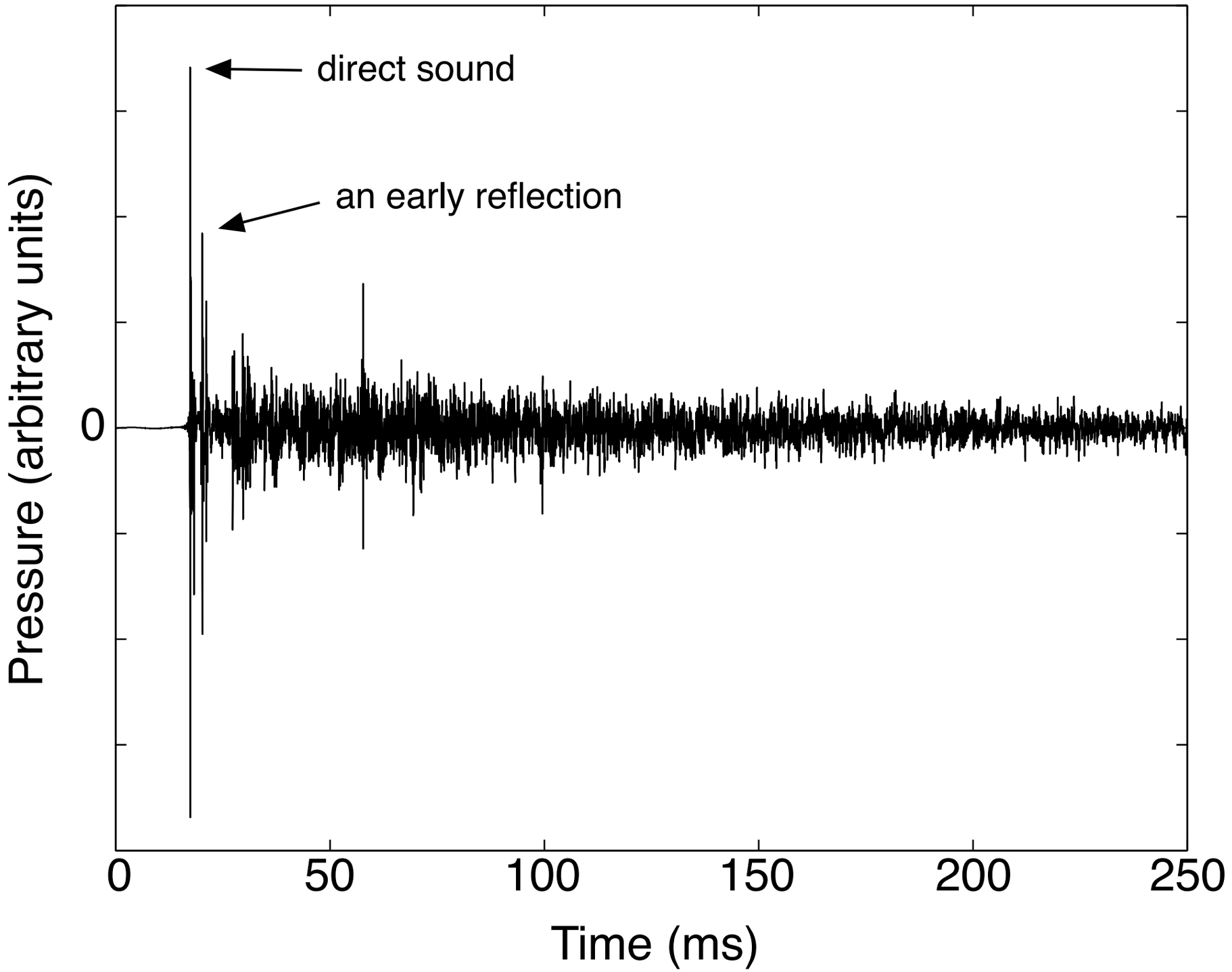
Reverberation profoundly distorts sound signals.



slide adapted from Josh McDermott

To measure reverb for a given location in a standard way:

Record Impulse Response.



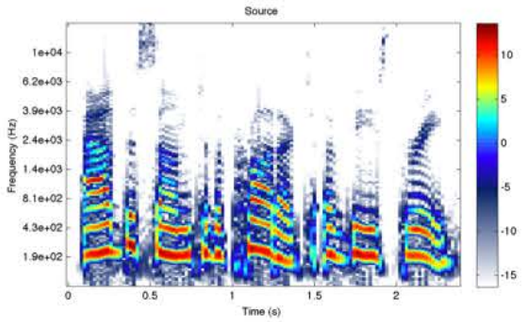
slide adapted from Josh McDermott

sound from source

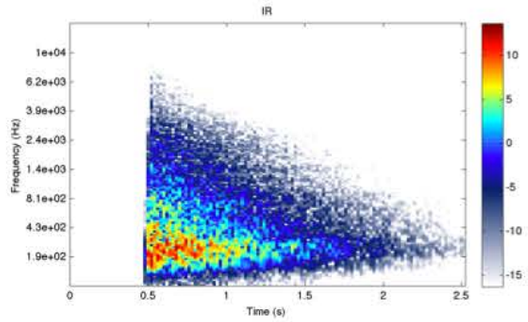
environmental impulse response

sound entering ear

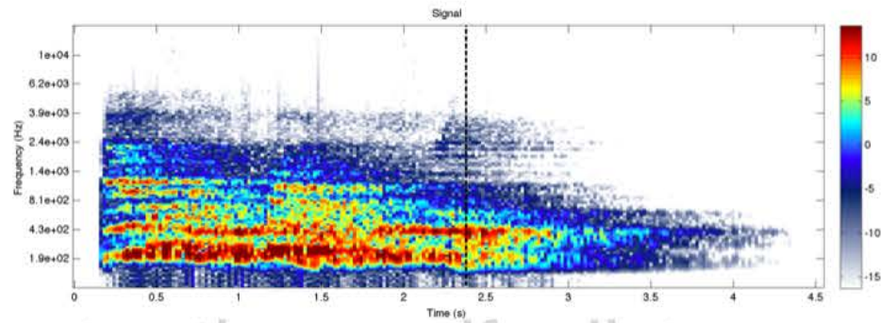
$$s(t) * f(t) = r(t)$$



*



=



Sound received by ear thus combines effects of reverberation and a sound source.

Problem: Listener is interested in source, and/or environment, rather than their combination.

Analogous to trying to infer the color of an object itself (reflectance) given just the light coming from it (luminance), given that this depends on the color of the light shining on the object (illuminant):

R * I = L given only L

slide adapted from Josh McDermott

Perception of reverberation as a process of separating the sound source from the reverb

unknown

unknown

sound entering ear

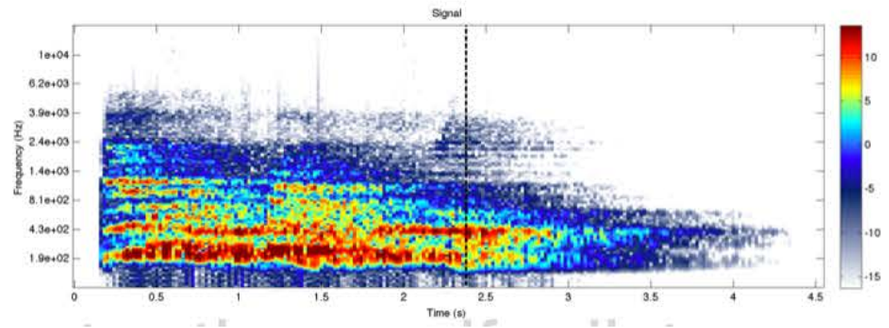
$$s(t) * f(t) = r(t)$$

?

*

?

=



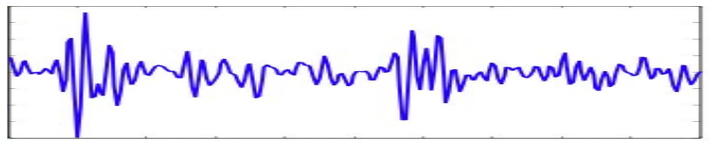
- Not solvable without assumptions about source and filter.
- Traer & McDermott (2016):
  Measured reverb in natural environments.
  Found that reverb has specific physical properties
  e.g. frequency-dependent exponential decay
  Listeners' auditory systems "know" this :
  they can decompose sounds into sources & spaces
  but only for reverb that has realistic filter properties
  not for reverb that violates these properties

So the ill-posed problem is solved in part by building in knowledge of the physics of reverb which constrains the solution.

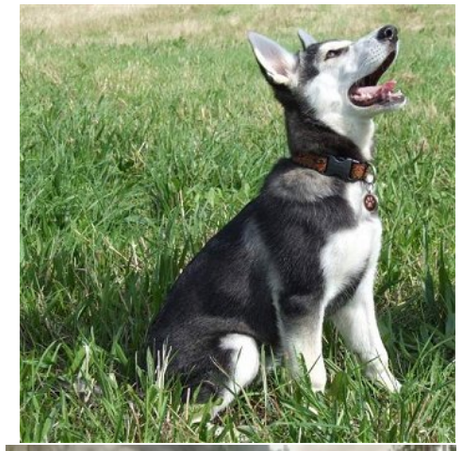
Hearing: Extracting Information from Sound

Why is this a computationally challenging problem?

Sound Input



?



Several challenges in extracting information from sound:

1. Invariance problems
   - same source sounds different in different contexts
2. Cocktail party problem: sources are superimposed in input,
   - need to separate them
   - ill-posed
   - need to constrain the problem with information about sounds
3. Reverb: sources interact with environment
   - use knowledge of properties of reverb to constrain the problem
   - thus enabling perception of both the sound source
   - and the space the sound happened in

Lecture 15: Hearing and Speech

Outline

I. Introduction (computational theory)

What information do we extract from sound?

What are the physical properties of sound?

Why is extracting information from sound computationally challenging?

Invariance problems: same source produces diff sounds

Ill-posed problem: cocktail party problem, reverb

II. Speech Perception

What is the structure of speech sounds?

Phonemes, formants, consonants & vowels

Why speech perception is computationally challenging

III. The auditory processing pathway

Peripheral transduction of sound & the cochlea (bare basics)

Primary auditory cortex

tonotopic organization

linear spectrotemporal filters

Speech-selective cortex

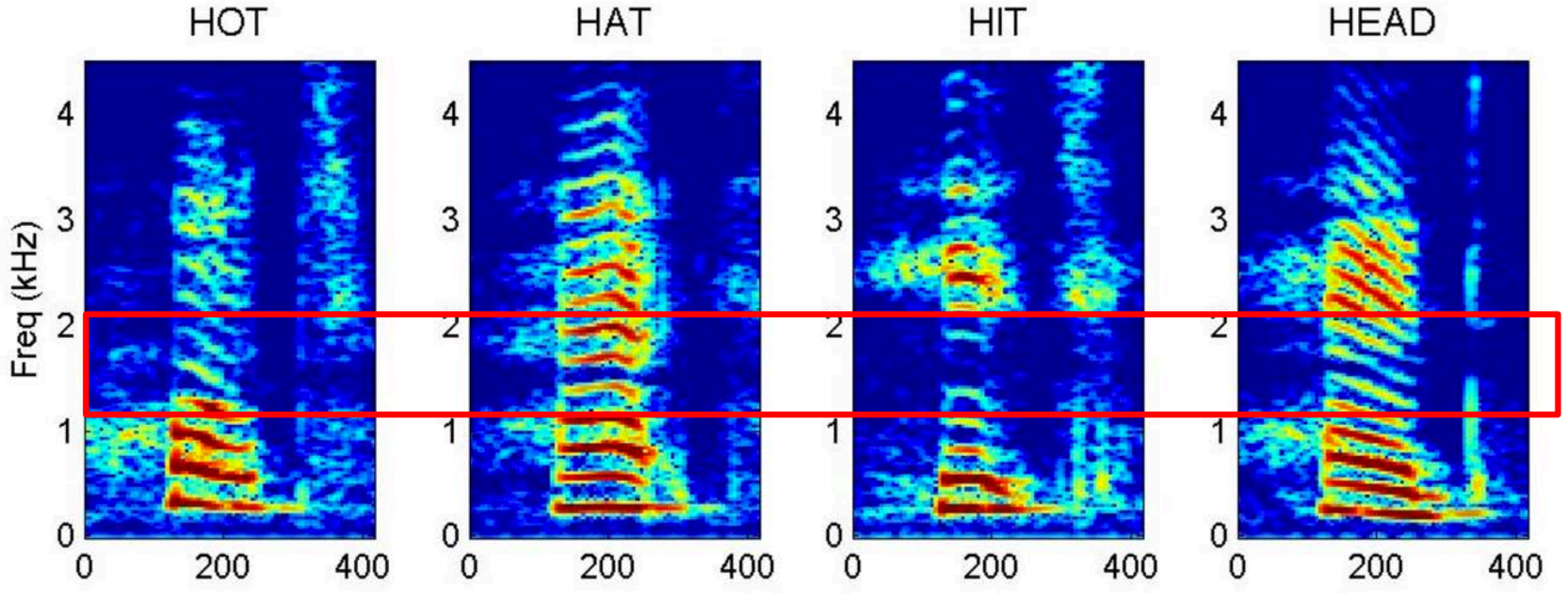
What is the Structure of Speech Sounds?

- Vowels have regularly spaced harmonics red stripes);
these determine pitch

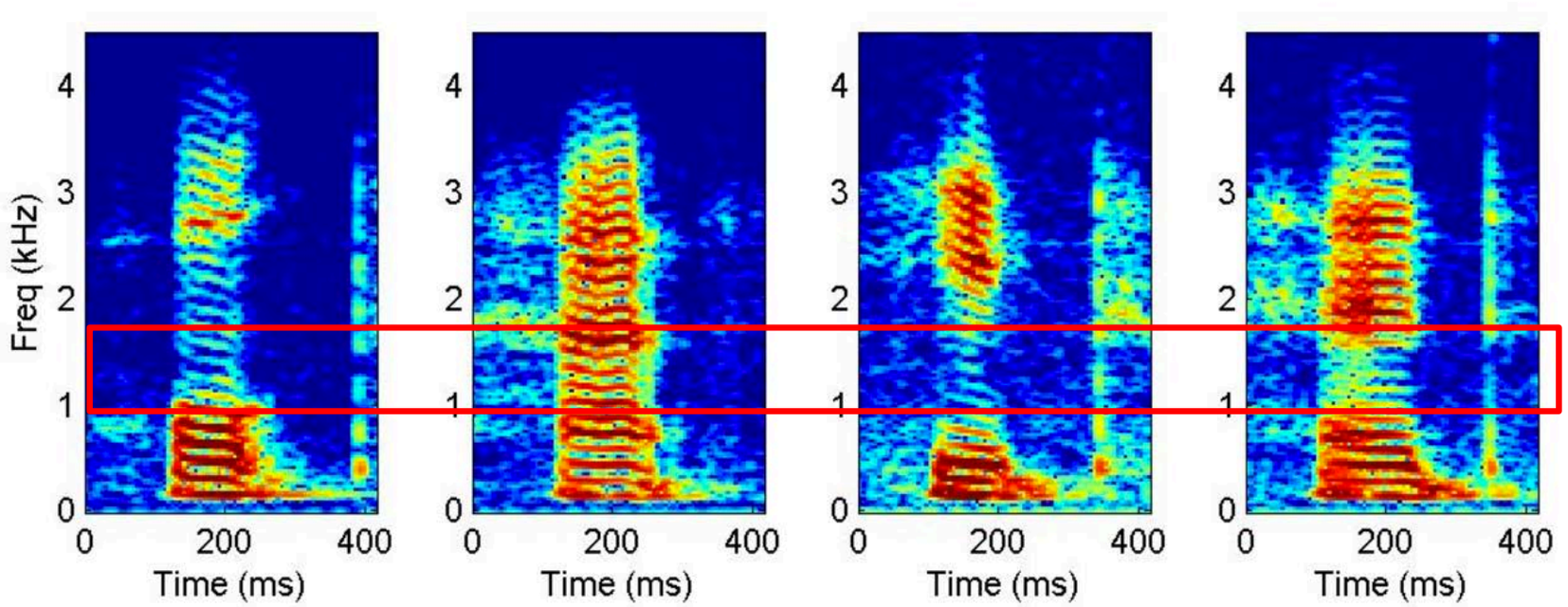
- Certain frequency bands where there is a lot of energy = “formants”, arise in vocal tract.

- Some frequency bands show differences for different vowels.

Spoken with a **high-pitched voice**



Spoken with a **low-pitched voice**



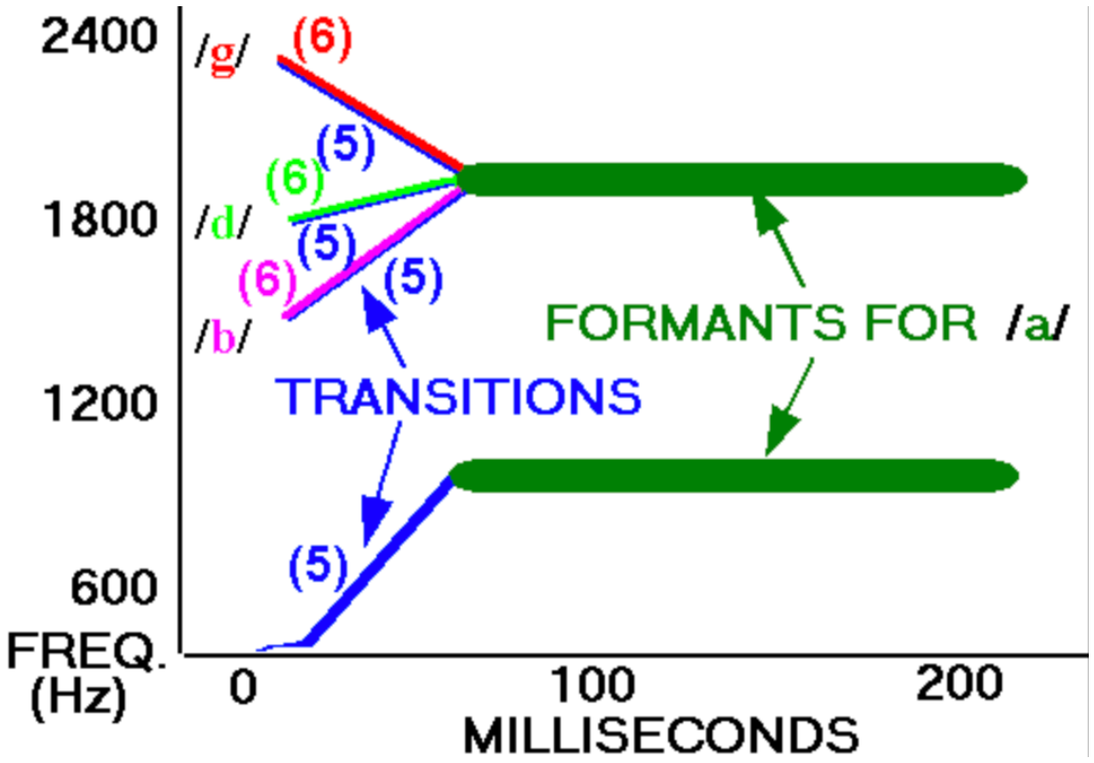
Check this out here:
https://musiclab.chromeexperiments.com/Spectrogram/

from: https://auditoryneuroscience.com/vocalizations-speech/formants-harmonics

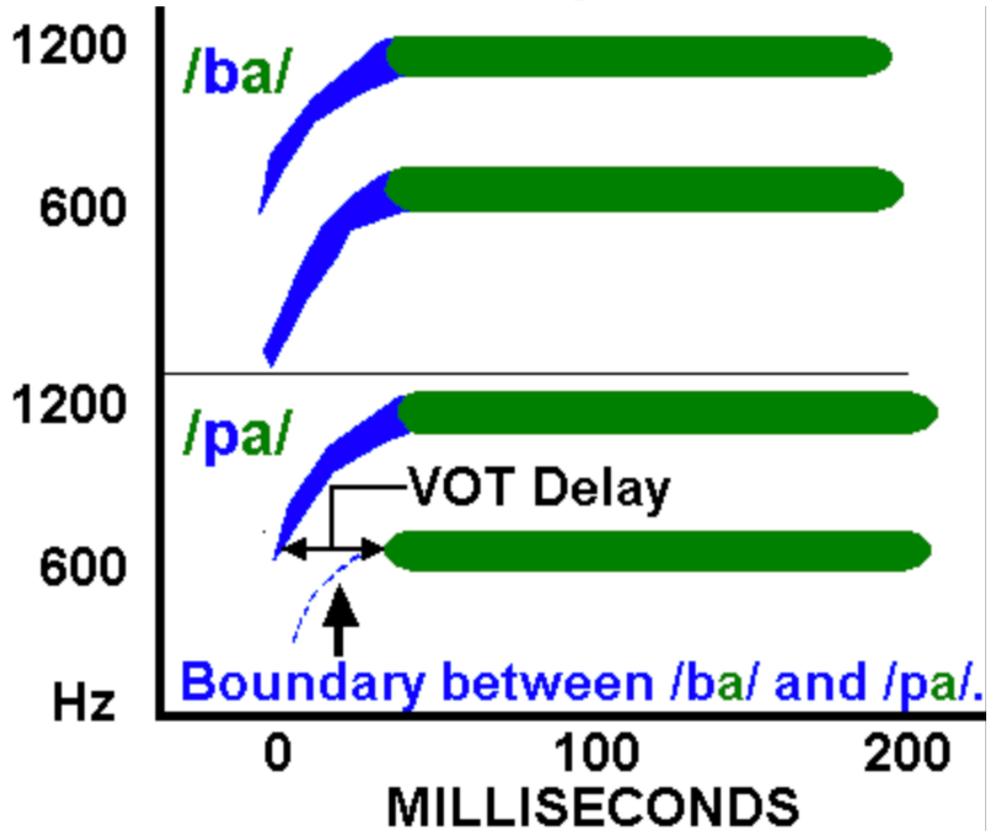
What is the Structure of Speech Sounds?

- vowels: see harmonics (parallel horizontal bars) in human voice.
  note sustained harmonics for voxels, and the diff between “EE” “OO” and “Ah”
- natural speech: note that voxels are in there but short, punctuated by lots of vertical nonharmonic sounds. those are the consonants.
- note the very subtle diff between “ba” and “pa in the first formant

Formants (voxels)
Transitions (consonants)



Voice onset time
“ba” vs “pa” 65 ms delay



Check this out here:
https://musiclab.chromeexperiments.com/Spectrogram/

Varieties of Phonemes

Each vowel or consonant sound is a "phoneme", i.e. a unit of sound that distinguishes one word from another.

Many phonemes are shared across languages, but many are not.

(r/l in US but not Japan;

d/d in Hindi)

(screenshot removed)

Some particularly awesome phonemes are the click consonants in some southern African languages.

See lecture video for Xhosa language tounge twister clip

Here for example is Test from Zimbabwe saying this sentence in Xhosa:

"The skunk was rolling and accidentally got cut by the throat."

Native English speakers use a few click consonants, though not as phonemes.

What are they?

Speech Perception

Is computationally challenging because the same phoneme does not always sound the same. Three sources of variation are:

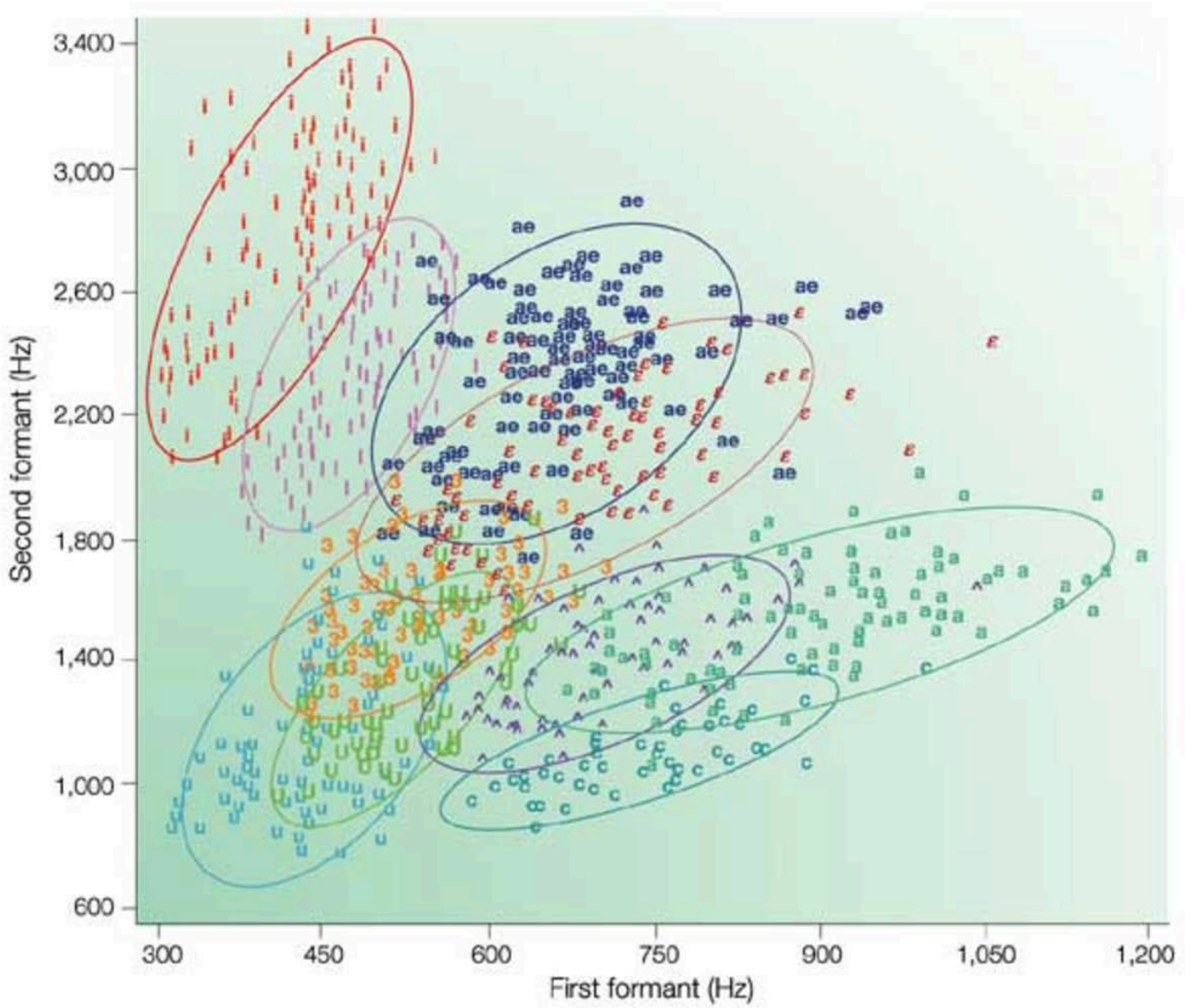
Rate variability
Slow speech results in different acoustic properties from faster speech, making physical descriptions of phonetic units difficult.

Context variability
The acoustic values of a phonetic unit change depending on the preceding and following phonemes.

Talker variability
When different talkers produce the same phonetic unit, such as a simple vowel, the acoustic results vary widely. This is because of the variability in vocal tract size and shape, and is especially different when men, women and children produce the same phonetic unit. For example….

Talker Variability in Speech

Each ellipse represents an English vowel, and each symbol within the ellipse represents one person's production.



When different talkers produce the same vowel, the acoustic results vary widely.
Worse: a given acoustic sound can correspond to multiple different vowels (overlap of ellipses).
A classic ill-posed problem in perception.
We have to learn about each speaker's voice to solve this problem, since voice and phonemes depend on each other.

Interdependence of voice & speech perception

1. Talker variability: Speech perception is affected by voice
Listeners perceive speech faster and more accurately from a familiar or consistent voice compared to unfamiliar and inconsistent voices. (Mullennix JW, Pisoni DB, 1990)

2. The language-familiarity effect (LFE): Voice perception is affected by speech
Listeners identify talkers more accurately in their native language than an unknown, foreign language. (Thompson, C.,1987).

a corollary (skip if running out of time)…….
Dyslexia is widely though to result from a deficit in speech perception
If this is true then the LFE might be reduced in dyslexics….

Lecture 15: Hearing and its Brain Basis

<u>Outline</u>

I. Introduction (computational theory)

What information do we extract from sound?

What are the physical properties of sound?

Why is extracting information from sound computationally challenging?

Invariance problems: same source produces diff sounds

Ill-posed problem: cocktail party problem, reverb

II. Speech Perception

What is the structure of speech sounds?

Phonemes, formants, consonants & vowels

Why speech perception is computationally challenging

III. The auditory processing pathway

Peripheral transduction of sound & the cochlea (bare basics)

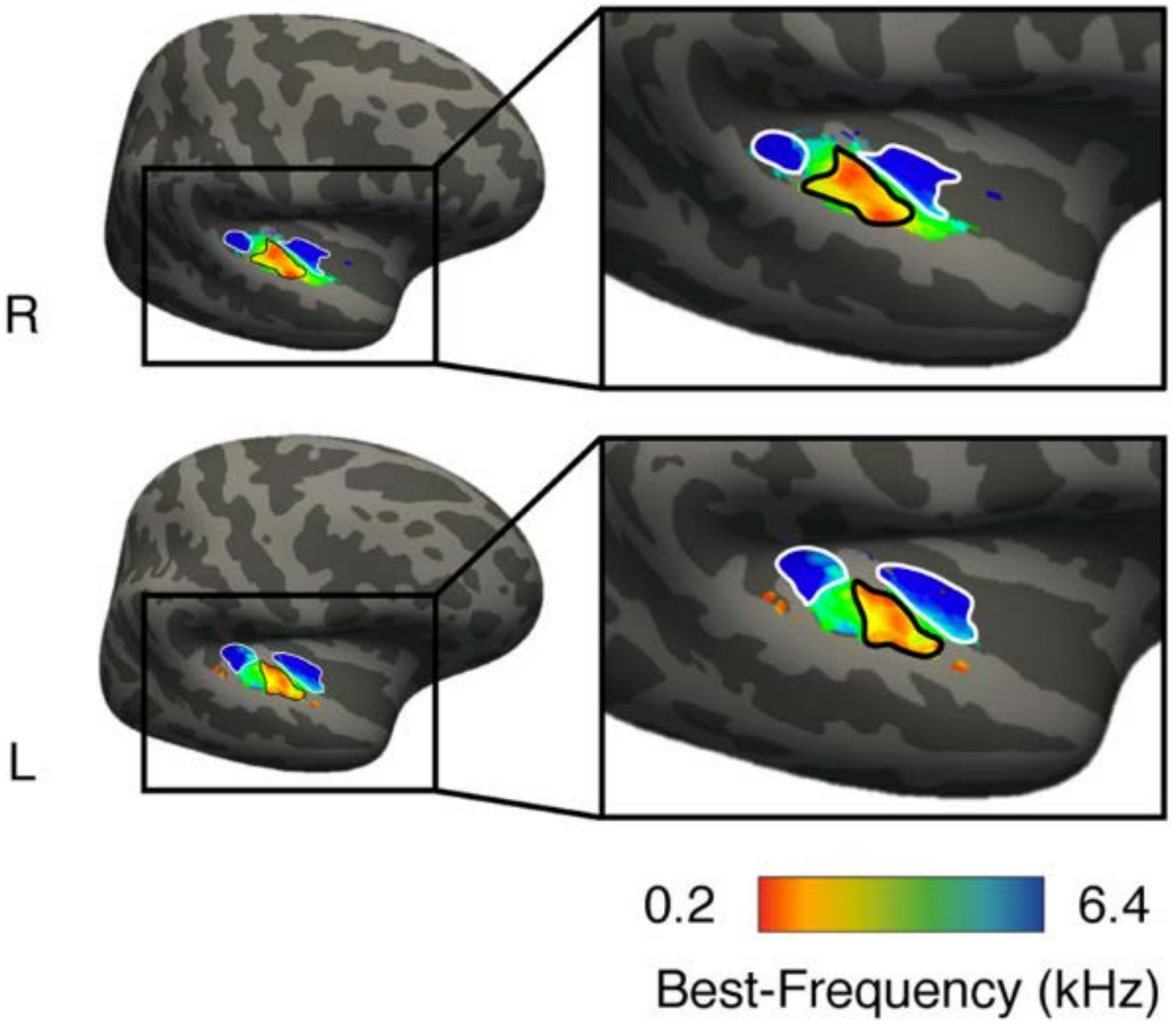
Primary auditory cortex

tonotopic organization

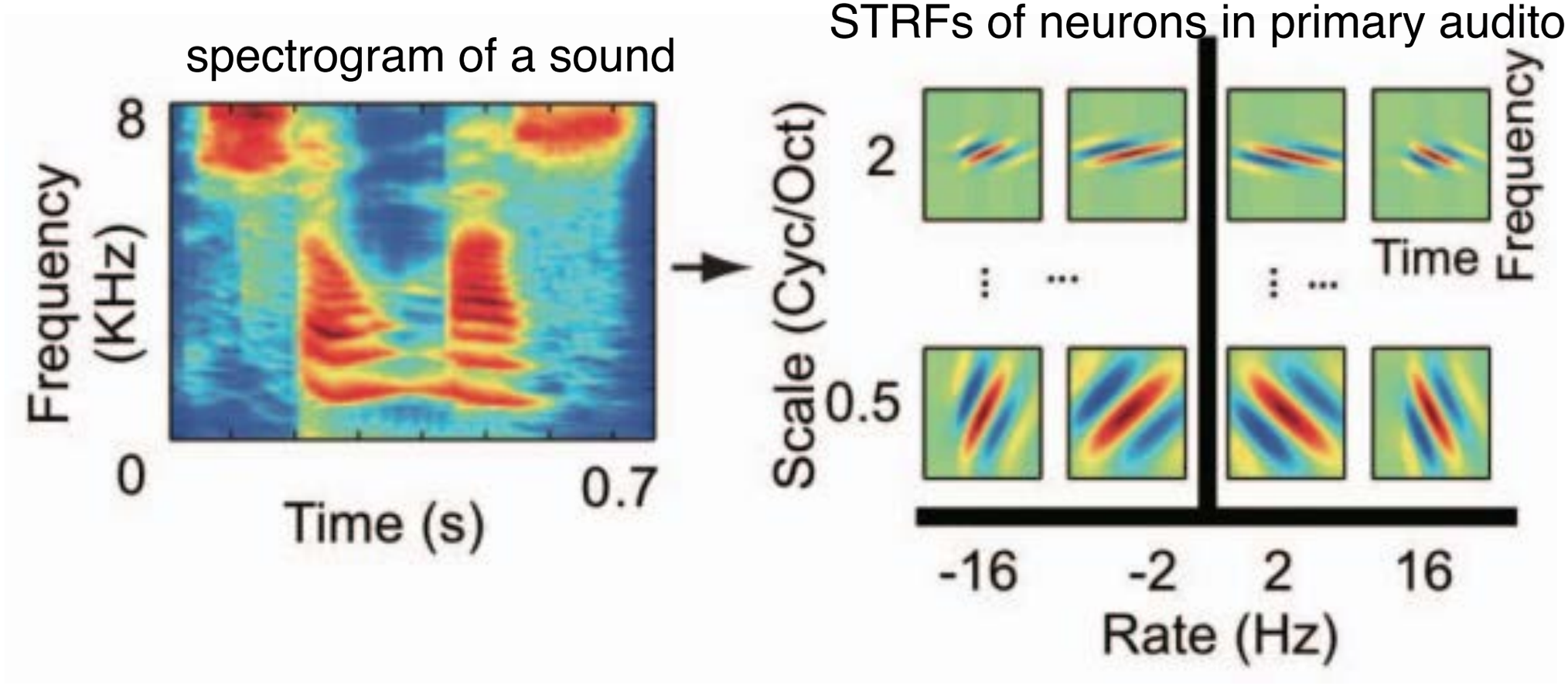
linear spectrotemporal filters

Speech-selective cortex

Primary auditory cortex has tonotopic maps



Standard model of neural responses in auditory cortex: linear spectrotemporal receptive fields (“STRFs”)



Neurons in primary auditory cortex can be thought of as a bank of linear filters selective for specific frequency changes over time.

How good a model is this of human primary auditory cortex?

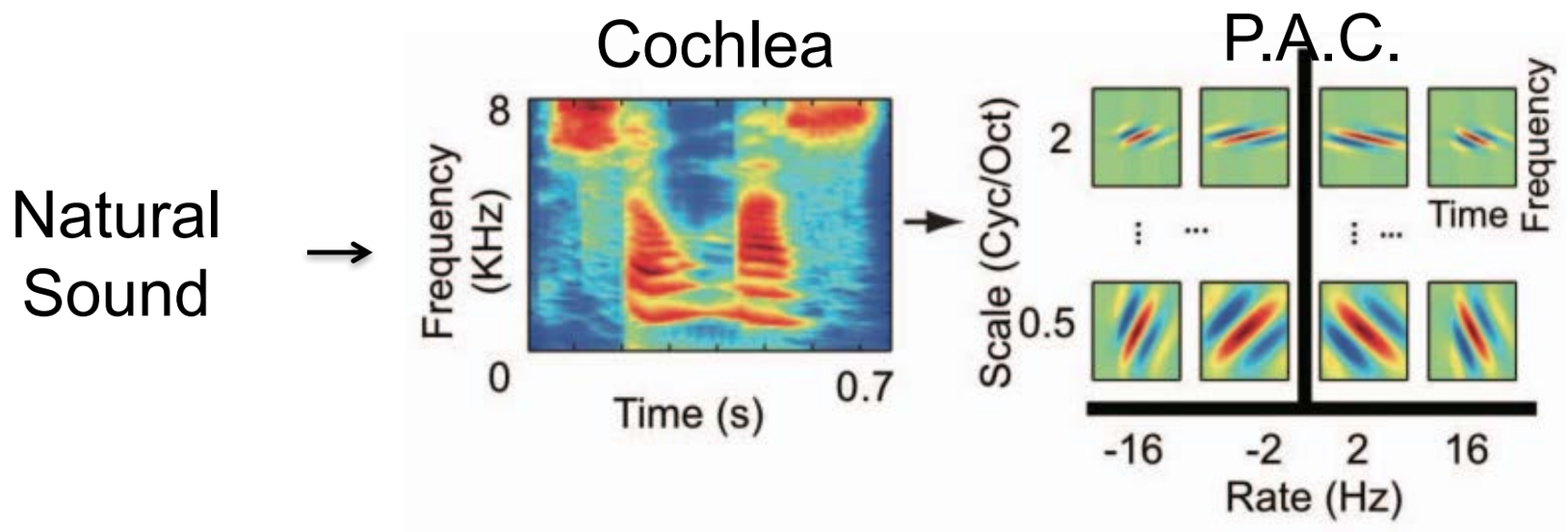
slide adapted from McDermott

Shamma and colleagues

Common model of auditory cortex: linear spectrotemporal filtering ("STRFs")

Testing theories with "model-matched stimuli":

- Idea: present a natural sound, and a synthetic signal that produces same response in a model
- If model is good description of neural response, responses to natural and model-matched sounds should be similar.
- Try this on a STRF-like model of primary auditory cortex



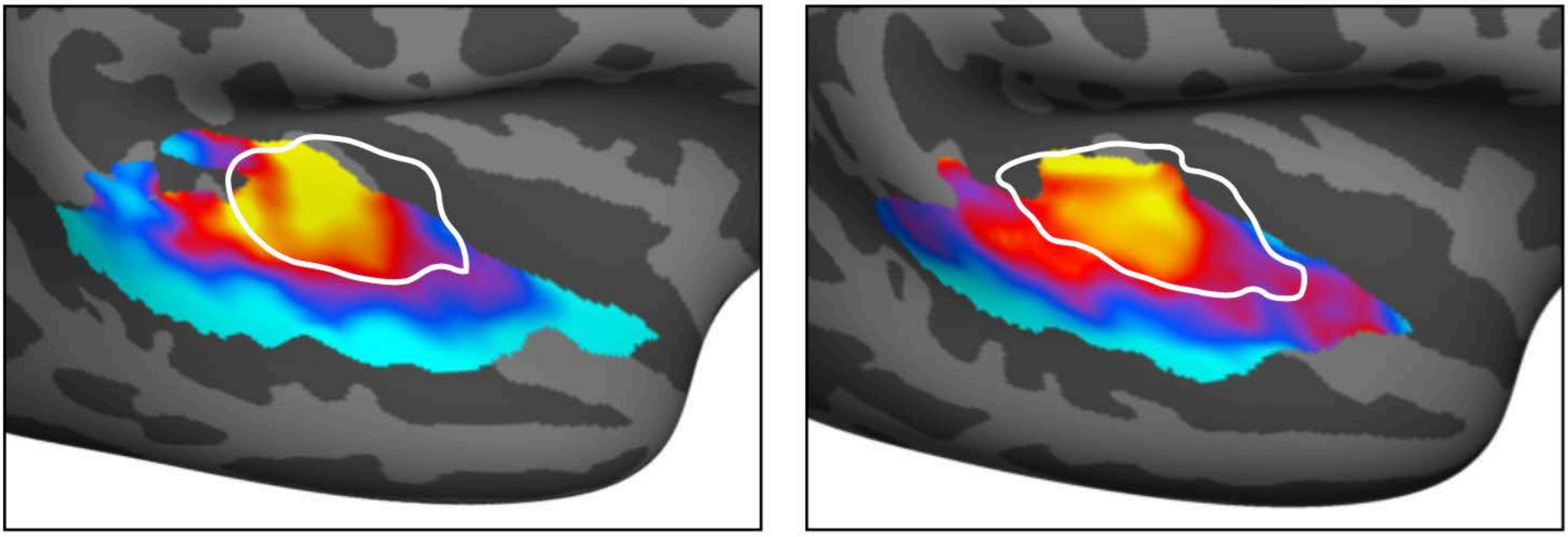
slide adapted from McDermott

Norman-Haignere & McDermott (2018)

Correlation between voxel response to real and synth sounds:

Left Hemisphere

Right Hemisphere

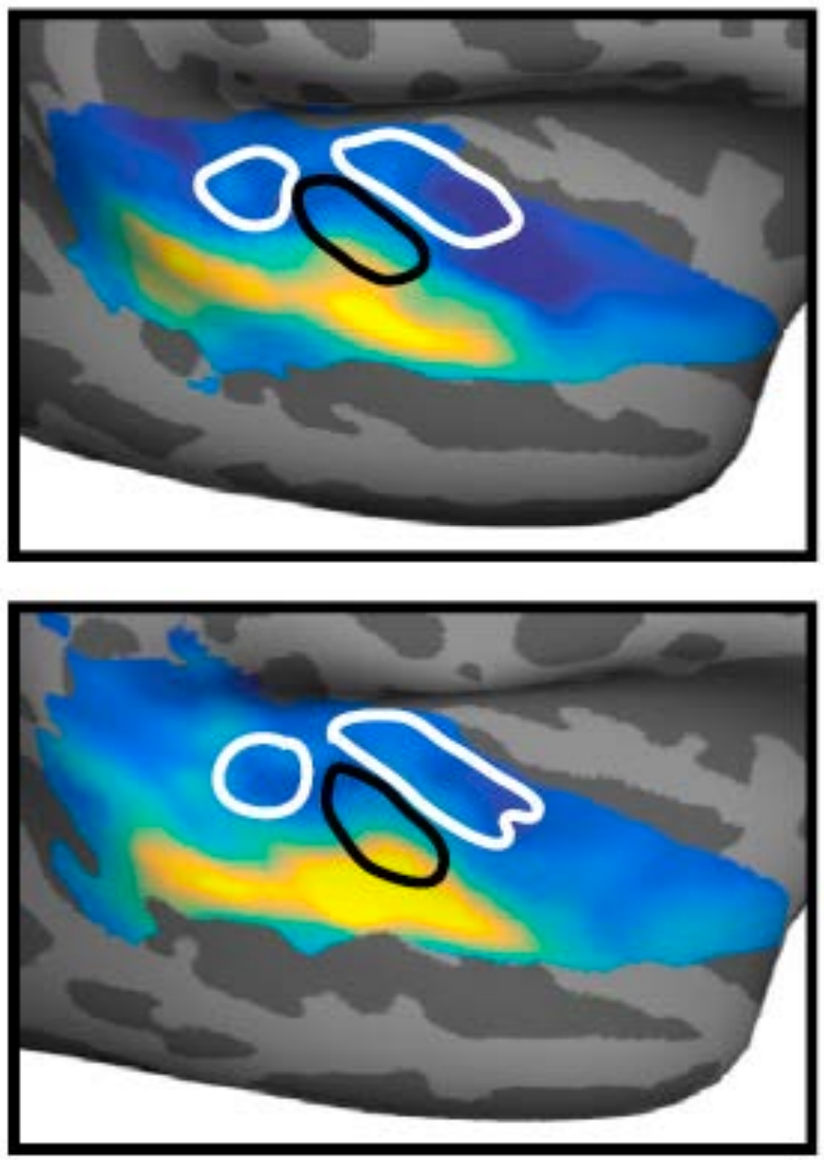


- Standard linear “STRF” model accounts for much of the voxel tuning near primary auditory cortex, very little in non-primary regions.
- What goes on after primary auditory cortex?
  One thing….

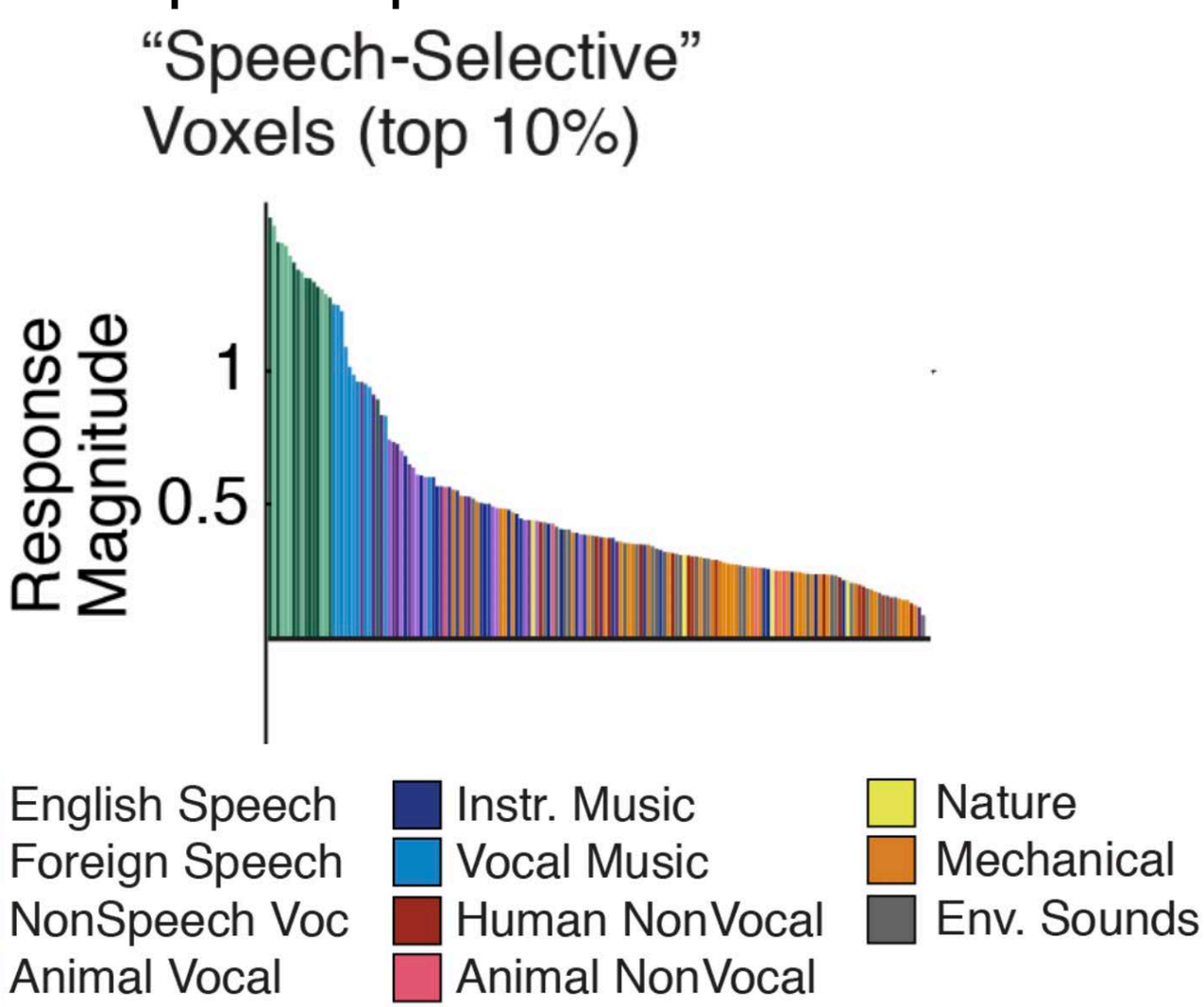
slide adapted from McDermott

Speech-Selective Cortex!

Speech> nonspeech sounds



Response profile



Note: this is speech-selective cortex, not language-selective cortex. What aspects of speech are coded here?

Norman-Haignere et al (2015). Figure S6

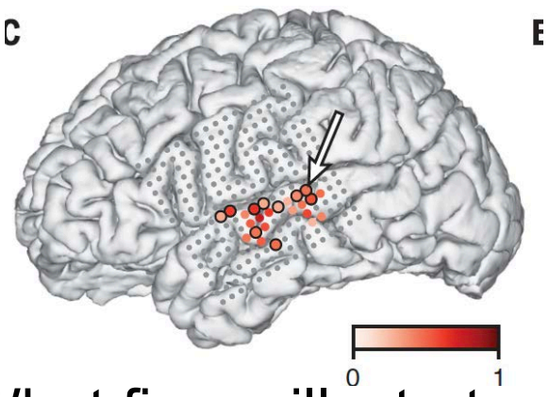
Tang, Hamilton, & Chang (2017)?

0. What question is this paper asking?

How is speech coded by the brain? Are speaker identity and intonation coded separately, even though both depend on pitch??

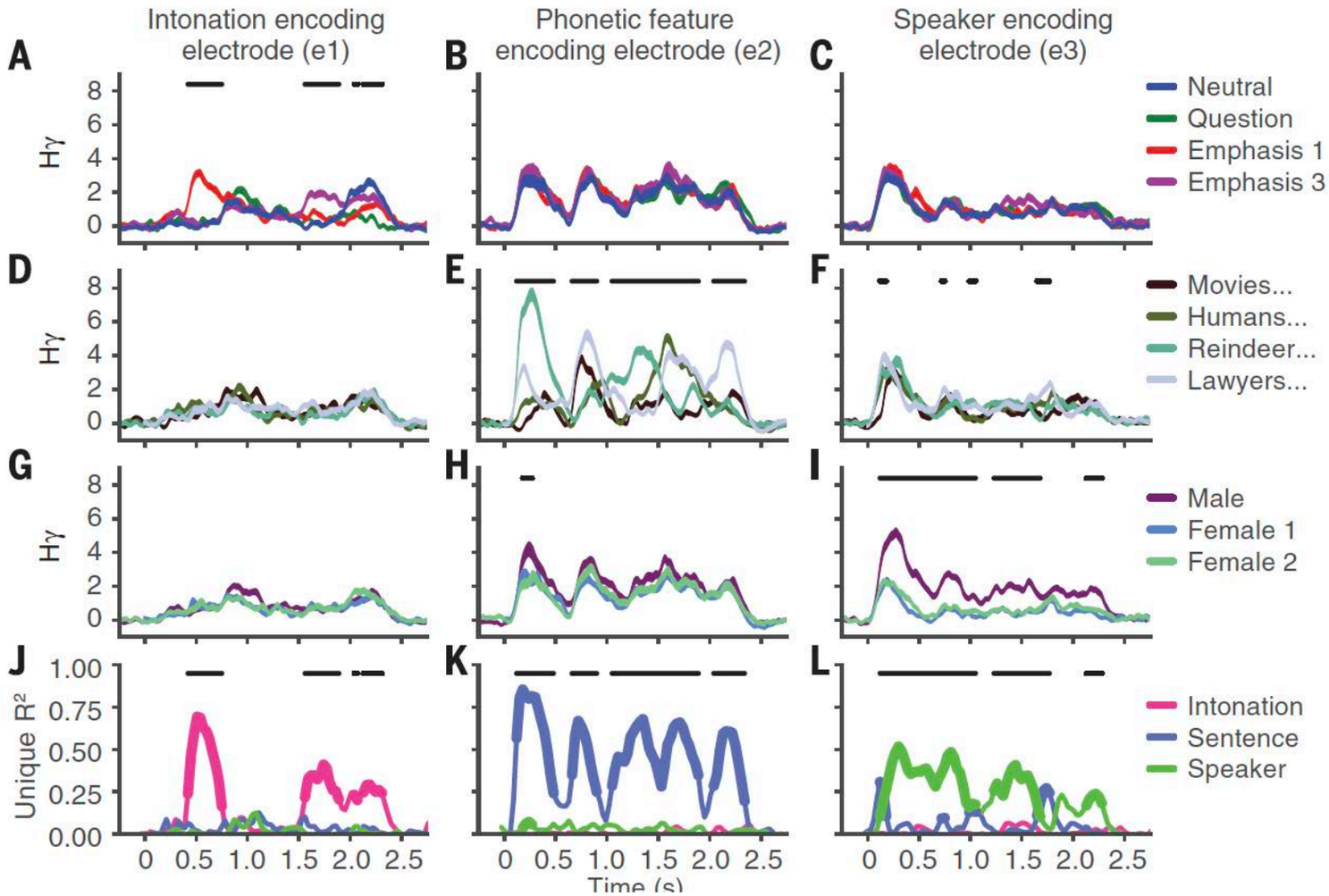
1. What method does this paper use?

Direct intracranial recording from the surface of the human brain. (or ECoG)



2. What is the main (most important) empirical finding in the paper? What figure illustrates this finding? *Figure 2 A-L*

That intonation and phonemes and speaker identity are coded independently, such that: Individual ecog sites code for just one of these dimensions and no site shows an interaction across these dimensions.This also means that each of these dimensions is coded invariant to the other dimensions.



3. What is the design of this experiment? That is, what factors were manipulated (list each), and how many different conditions were used?

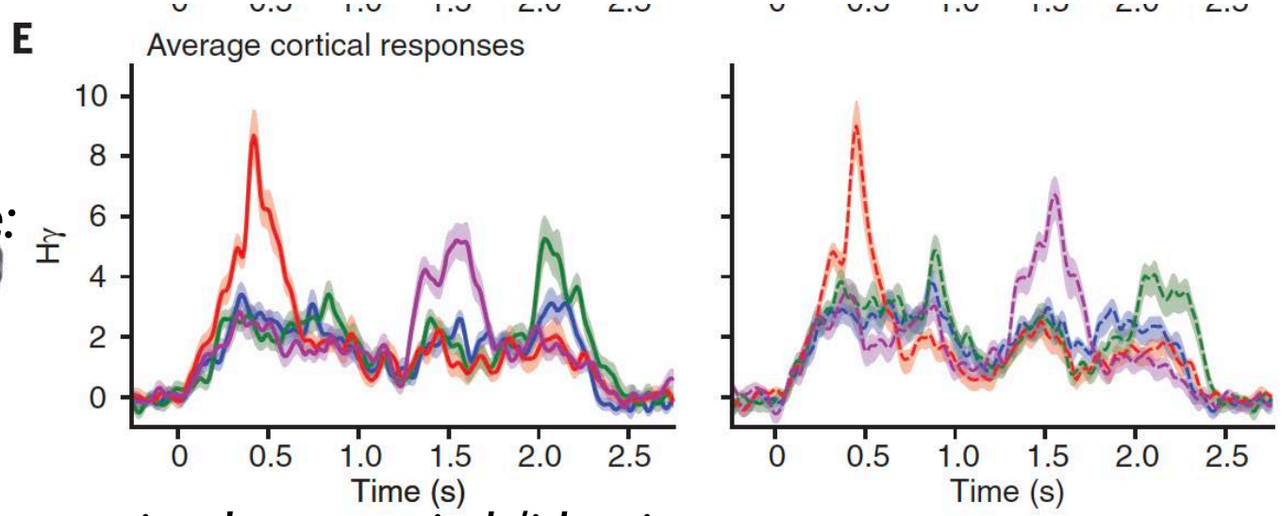
identity (3 levels) x intonation (4 levels) x phonemes (4 levels)

Tang, Hamilton, & Chang (2017)?

4. What does Figure 1E show?

Response of one electrode:

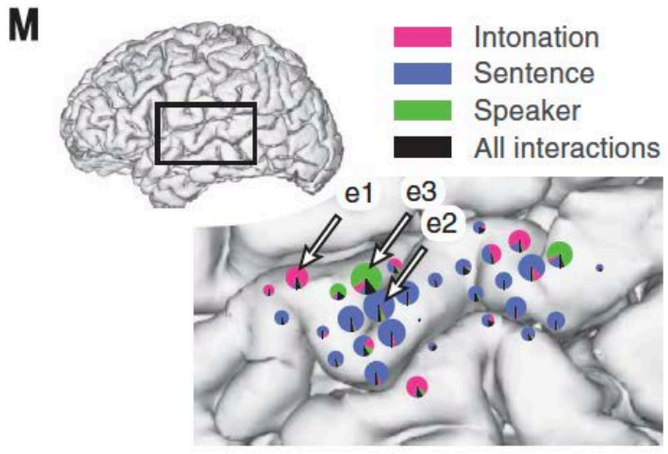
- Neutral
- Emphasis 1 (E1)
- Emphasis 3 (E3)
- Question (Q)



An electrode sensitive to intonation but not pitch/identity

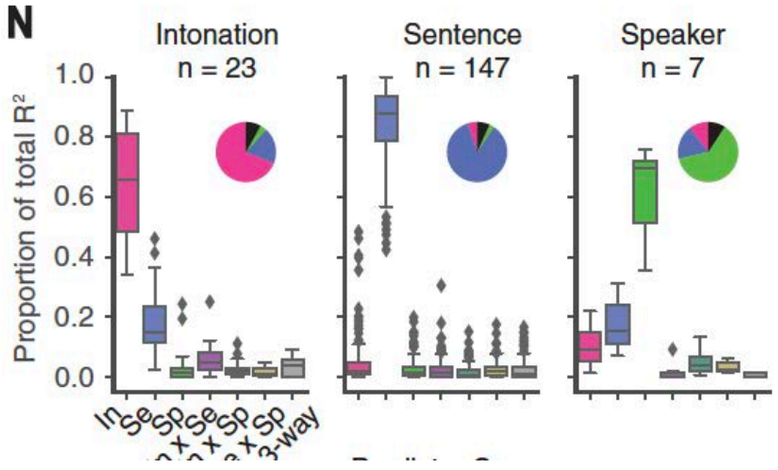
5. The middle graph in figure 2N shows 147 electrodes that had information about which sentence was spoken. Could you decode from these electrodes whether the speaker saying those sentences was male or female?

Those electrodes only show effects of sentence, not of identity/ gender.



6. Would you expect the electrodes that can discriminate intonation to be able to discriminate musical melodies (in which each note has a different pitch)? If so, how would you characterize the function of those sites?

Maybe, as both are pitch contours! That would suggest these electrodes are not selective for speech processing alone, but more generally for pitch.



7. If the same experiment were repeated using fMRI, what results would you expect?

First, what fMRI method would you use?

Which would be most likely to work?

9.13 The Human Brain
Spring 2019