Speech Perception

LIN 350

Scott Nelson

Stony Brook University

August 8, 2023

Topics for today

- Categorical Perception
- Integrating Visual Information
- Lexical Influence
- Normalization

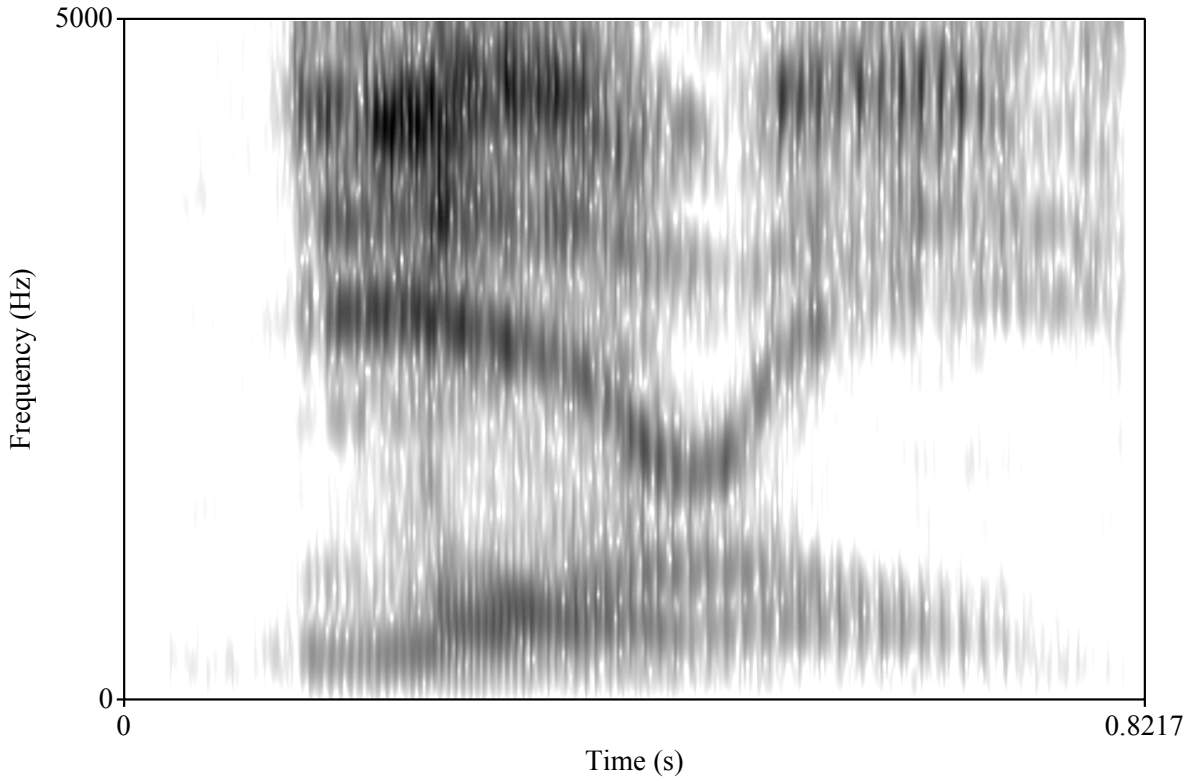
Speech Perception

What do you hear?

- Let's listen to a short audio clip
- What name do you hear?

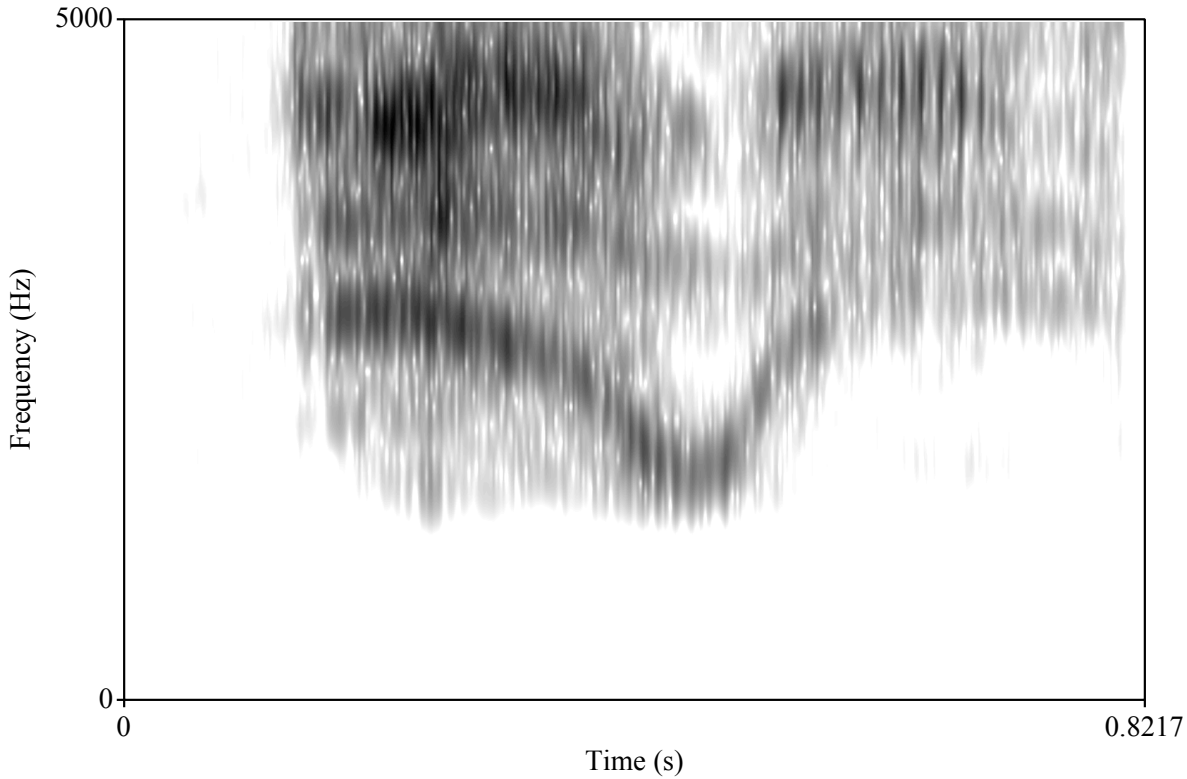
What do you hear?

- Let's listen to a short audio clip
- What name do you hear?
- Let's look at the acoustics



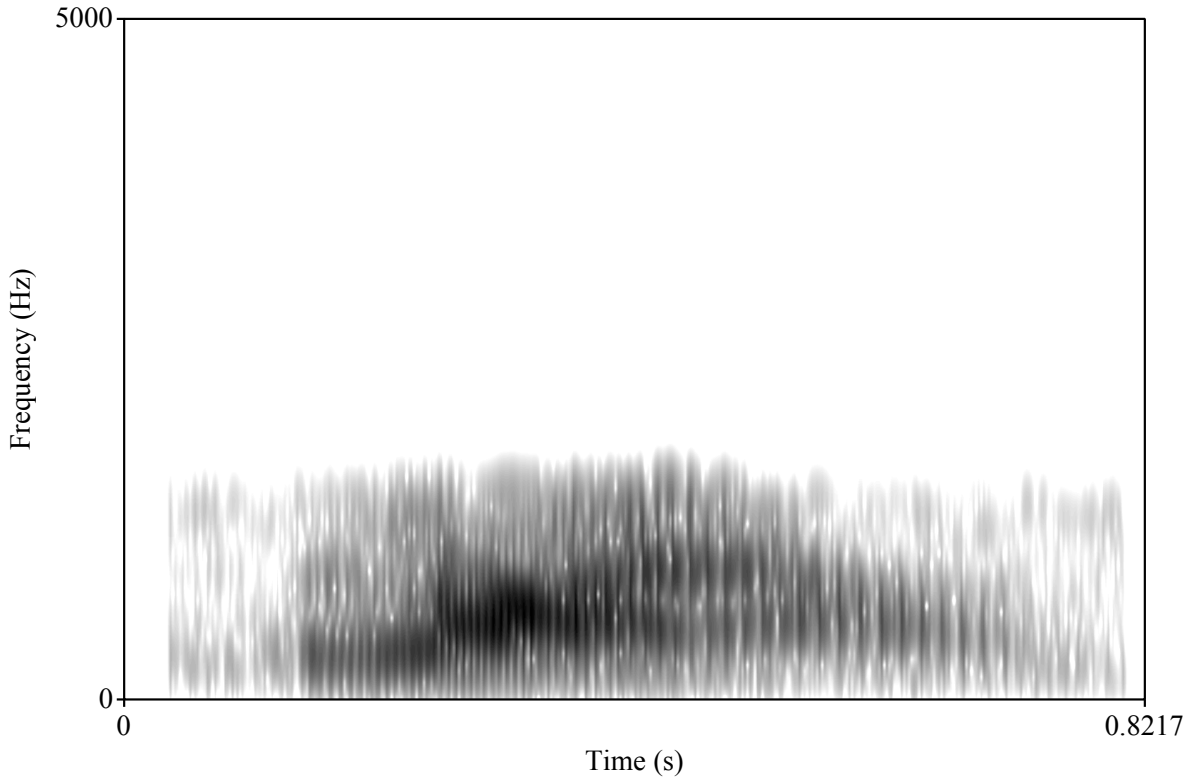
What do you hear?

- Let's listen to a short audio clip
- What name do you hear?
- Upper band contains "Yanny" cues



What do you hear?

- Let's listen to a short audio clip
- What name do you hear?
- Lower band contain "Laurel" cues



Some questions to get us thinking

- What type of information do we use when trying to figure out what a person has said?
- Why do we sometimes mishear what has been said?

Categorical Perception

- The biggest issue for models of speech perception is the *lack of invariance problem*
- This means there is no single acoustic feature that signifies one sound over another
- The actual acoustic properties of a phoneme can vary based on speaker, context, speech rate, and many other variables
- One way that perceivers seem to overcome the lack of invariance problem is by integrating both top-down and bottom-up information

Categorical Perception

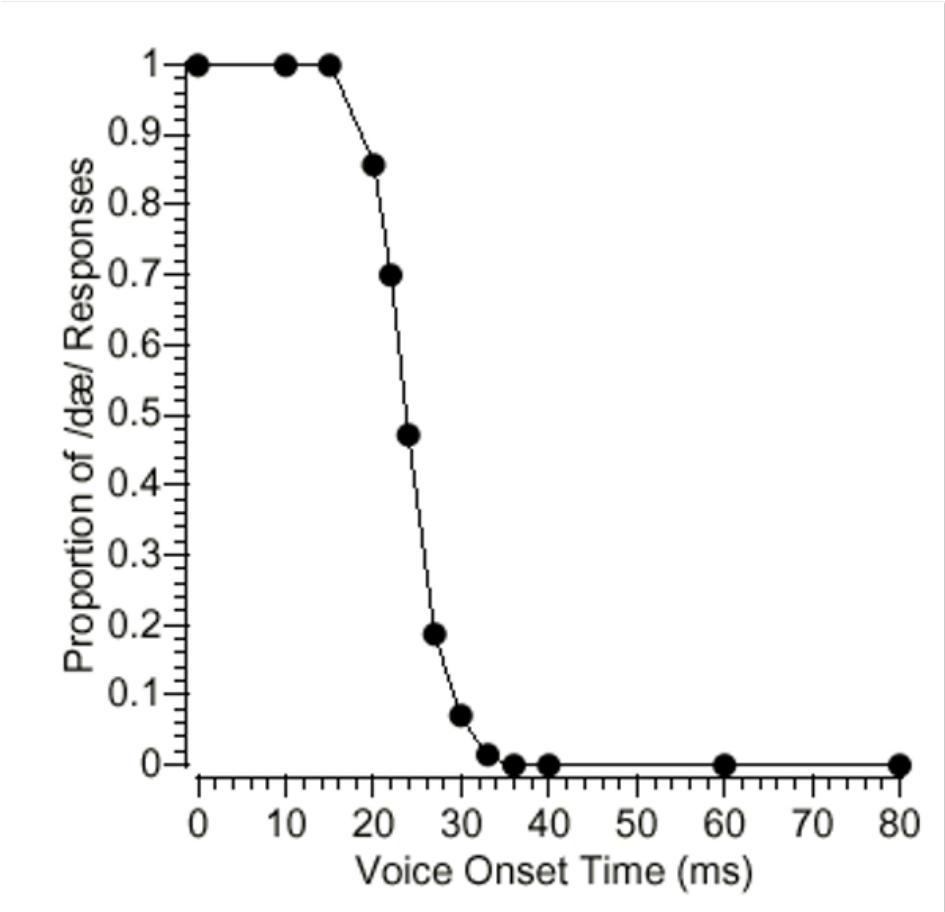
- Top-down information
  - Knowledge of the grammar
  - Syntax, Semantics, Phonology, etc...
- Bottom-up information
  - The acoustic signal
  - Other perceptual inputs (visual)

Categorical Perception

- One type of top-down effect in speech perception is categorical perception
- Even though properties of speech are continuous, we still seem to perceive them as categories (i.e., phonemes)
- Sounds at the boundaries of two categories may be harder to categorize, but we still always choose one category or the other

Categorical Perception

- One example of categorical perception is VOT
- For English speakers, the boundary is around 25 ms
- Any stop with >25 ms VOT will be perceived as voiceless
- Any stop with <25 ms VOT will be perceived as voiced



Categorical Perception

- Another example is the r/l distinction in English
- Japanese speakers don't have two separate categories
- English speakers are good at telling them apart
- Japanese speakers are always around 50% even when English speakers can tell them apart

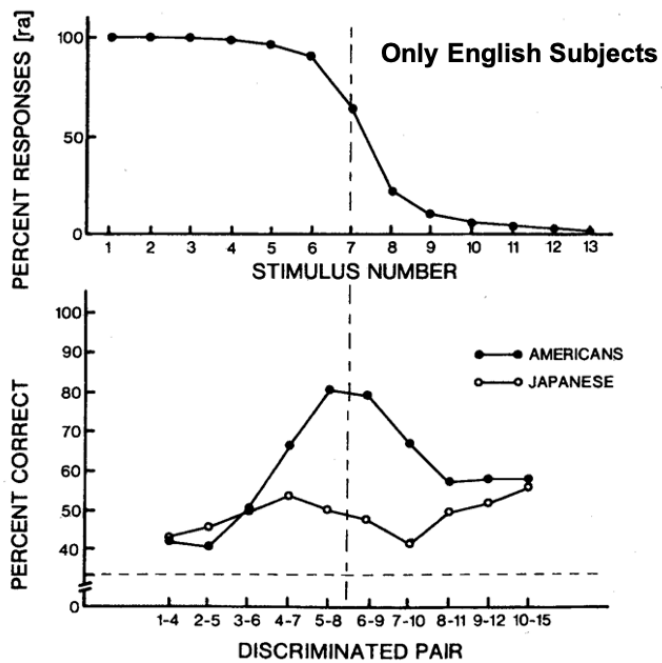
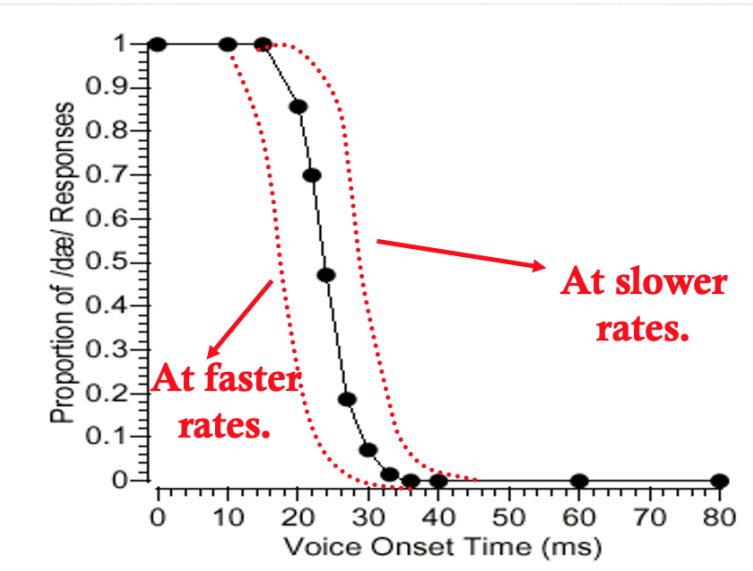


Figure 12.2. Test of the categorical perception of /ra/ and /la/ by American and Japanese adults. American listeners show the characteristic peak in discrimination at the phonetic boundary; Japanese listeners do not. (From Miyawaki et al., 1975.)

Categorical Perception

- It also seems to be the case that we adjust our perception based on prosody and speech rate
- At faster speech rates all our gestures are shorter so the VOT boundary moves to a shorter location
- At slower speech rates all our gestures are longer so the VOT boundary moves to a longer location
- Generalized over all sounds and not limited to ones you have recently heard at fast/slow speech rates



Activity

- Let's see how categorical perception works

Integrating Visual Information

- Which vowels do you think the speaker is making in the images below?

Integrating Visual Information

- Which vowels do you think the speaker is making in the images below?
- The left image shows the speaker making an [u] vowel
- The right image shows the speaker making an [o] vowel

Activity

- Let's watch this video to see how visual information can affect our perception of sounds
- https://www.youtube.com/watch?v=aFPtc8BVdJk

Integrating Visual Information

- When we perceive speech we use both audio and visual information!
- In the youtube video the audio information corresponds to [bɑ] but the visual information corresponds to [gɑ]
- Our brain "blends" both together into a percept that sounds like [dɑ]
- An example of (possibly automatic) bottom-up information

Lexical Influence

- Our knowledge of the lexicon also affects the way we perceive speech
- Ambiguity disappears if there is a close enough lexical item
- We are constantly predicting what sounds come next based on the words we know

Activity

- Let's do some listening experiments!

Lexical Influence

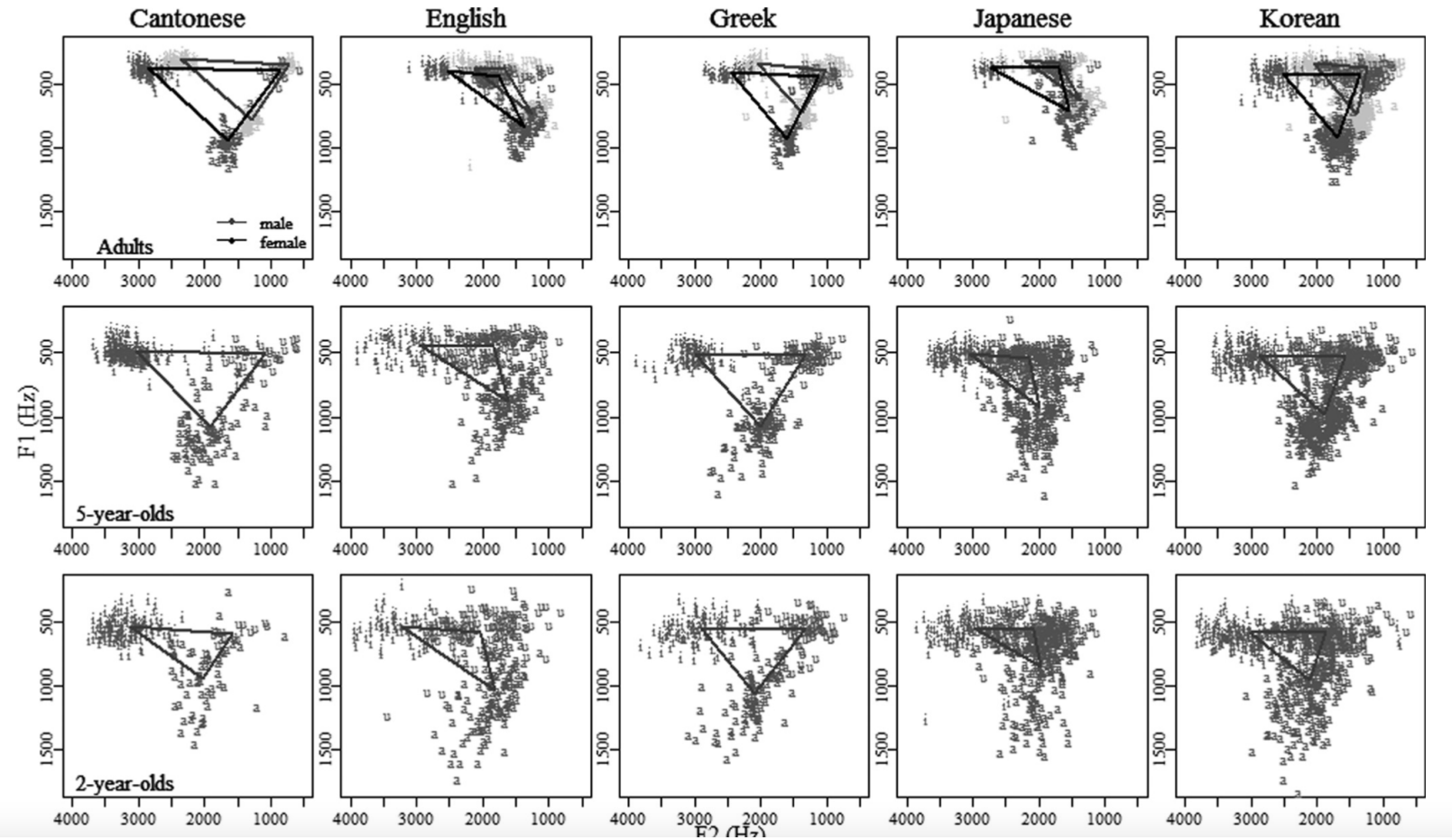
- A sound halfway between /d/ and /t/ is usually perceived as...
  - /d/ before “ash” because *dash* is an English word but *tash* is not
  - /t/ before “ack” because *tack* is an English word by *dack* is not
  - 50/50 between /d/ and /t/ before “ath” because neither *dath* nor *tath* are English words

Vowel Normalization

- Last week we discussed how we can predict vocal tract length based on formant values
- This highlights a problem for speech perception research/the lack of invariance problem
- The formant space can vary drastically based solely on physiological makeup
- One way to see this clearly is child vs. adult vowel spaces

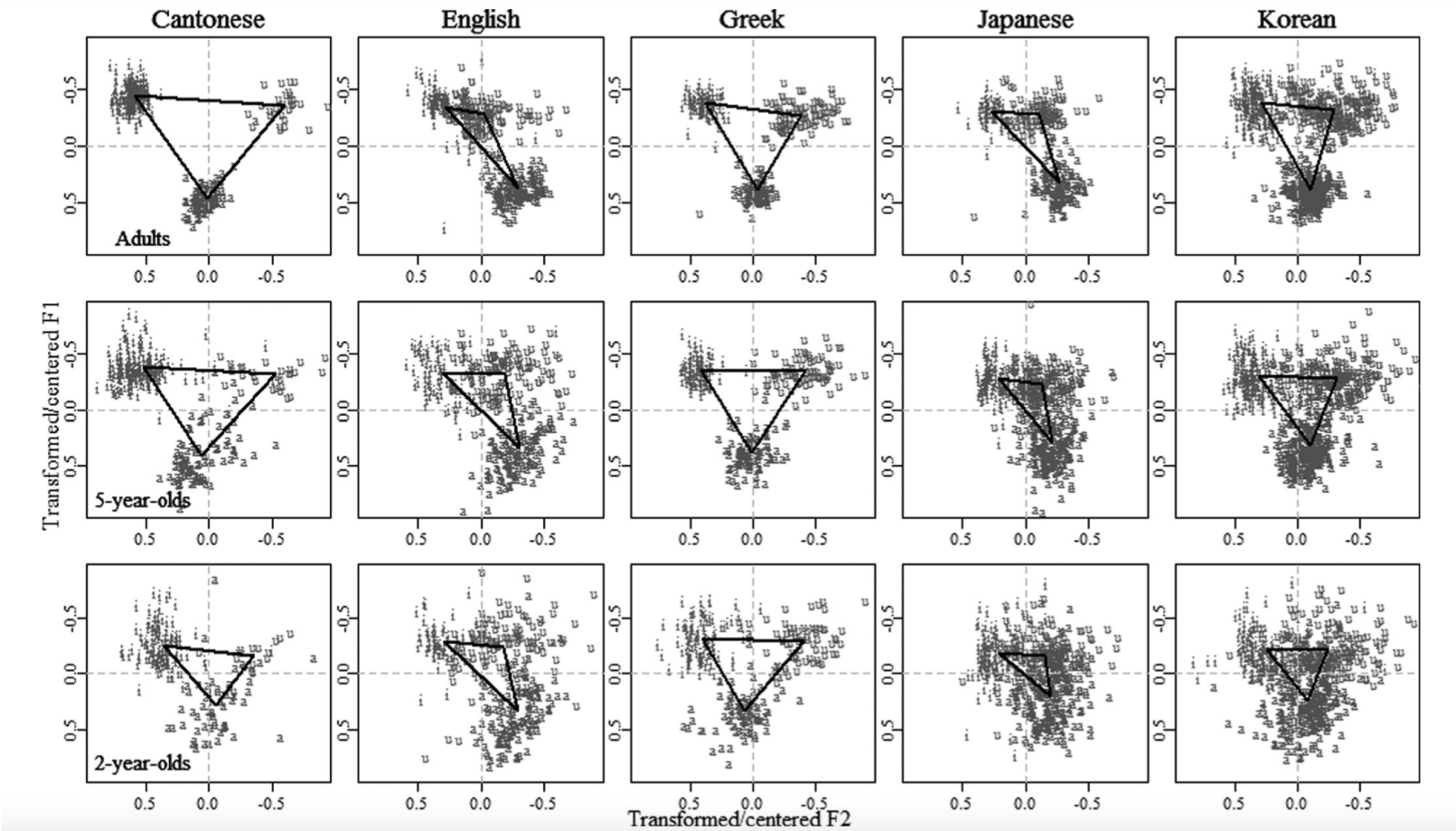
Vowel Normalization

- Notice how high the F1 and F2 values are for children



Vowel Normalization

- Once we normalize the data the difference is drastically reduced



Vowel Normalization

- Vowel systems convey at least three types of information:
  - phonemic information
  - anatomical/physiological information
  - sociolinguistic information
- When perceiving speech our goal is to obtain the first bit of information but we can't do that without using the second and third pieces of information

Vowel Normalization

- Four goals for vowel normalization:
  - minimize inter-speaker variation due to physiological or anatomical differences
  - preserve inter-speaker variation in vowel quality due to dialectal or social factors, or to preserve features in a sound change
  - maintain phonological differences
  - model cognitive processes that allow listeners to understand different speakers.

Vowel Normalization

- Vowel normalization is based on three traits
  - speaker
  - vowel (token)
  - formant
- Each of these traits can be based on either intrinsic or extrinsic properties
- **Intrinsic:** based on only a single speaker, vowel, or formant
- **Extrinsic:** based on all speakers, vowels, formants (in the data set)

Conclusion

Conclusion

- Speech perception is hard to model because of the *lack of invariance problem*
- Speech perception involves integrating all sorts of information and even then humans often mishear/misperceive/misunderstand what is being said