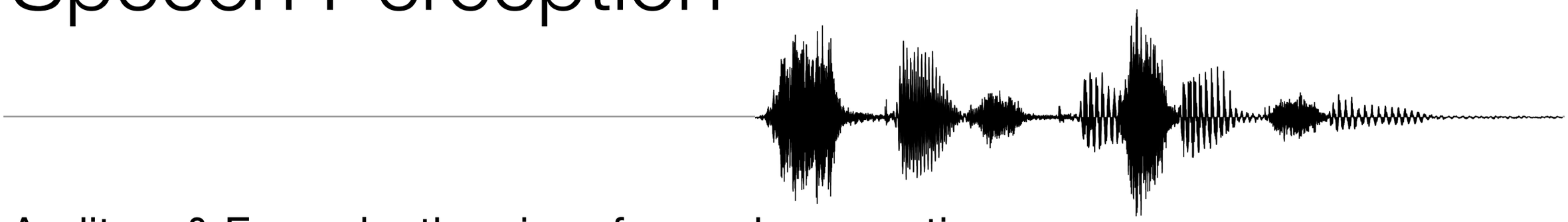
Speech Perception



Auditory & Exemplar theories of speech perception

College of Arts and Sciences
Department of Linguistics

2019 LSA Institute (Beddor & McGowan)

The main question of speech perception (traditionally):

How do listeners interpret the acoustic –or, rather, auditory– signal as linguistic forms?

how are we so consistent at it?

Motor Theory

- Listeners perceive (intended) gestures
- Speech is perceived in a specialized speech module
- Speech perception recruits the motor system

Direct Realism

- Listeners perceive (actual) gestures
- Speech perception is not special
- No proposed role of motor system in perception.

How do listeners interpret the input acoustic signal as linguistic forms?

General Auditory Theory

- Listeners perceive auditory signal
- Speech is not special
- No proposed role of motor system in perception.

Note: we are abusing the word *theory* here. Theory, in science, is generally reserved for models that have been rigorously tested and have withstood the tests of time and empirical study. Evolution is a ‘theory’ because we’re about as sure it’s true as it’s possible to be sure of anything.

But then we’re linguists, not prescriptivists. :)

But what were gesturalist approaches *for*?

i.e. what **observable problem(s)** do they solve?

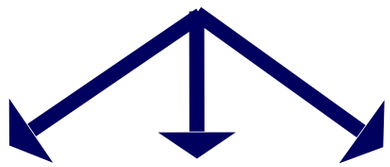
Lack of invariance!

Issues that arise in acoustics-to-phoneme mapping

Lack of Invariance

The problem: In some cases, there appear to be no acoustic properties that reliably correspond to the segments of linguistic analysis and perception.

Acoustic signal



Multiple percepts

One-to-many mapping

Acoustic signals

Percept

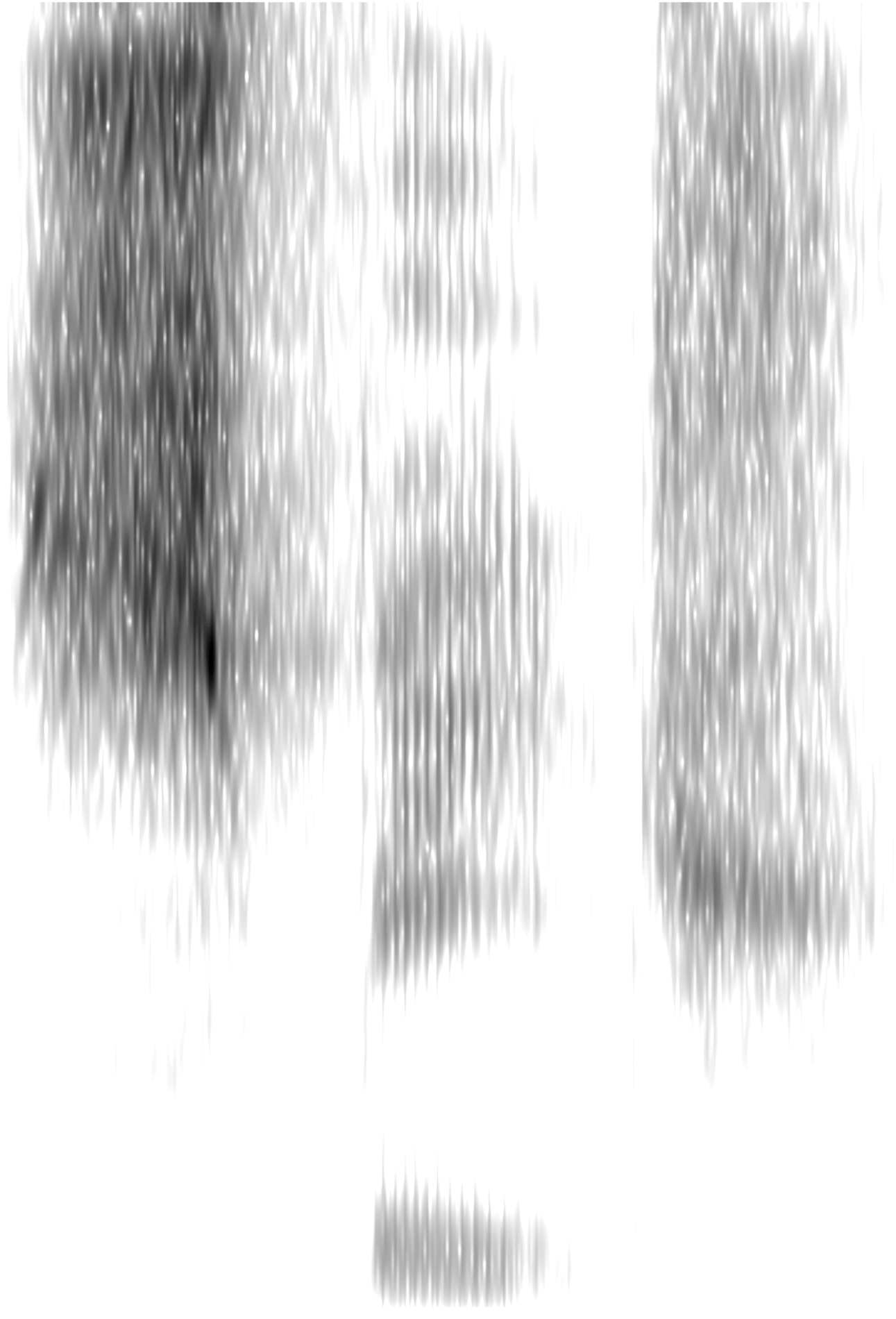
Many-to-one mapping

Issues that arise in acoustics-to-phoneme mapping

- Segmentation problem
- Lack of Invariance problem

Auditory theories and gestural theories both have to address these problems!

Praat “speech” demo



We *hear* sounds.

But what do we *perceive*?

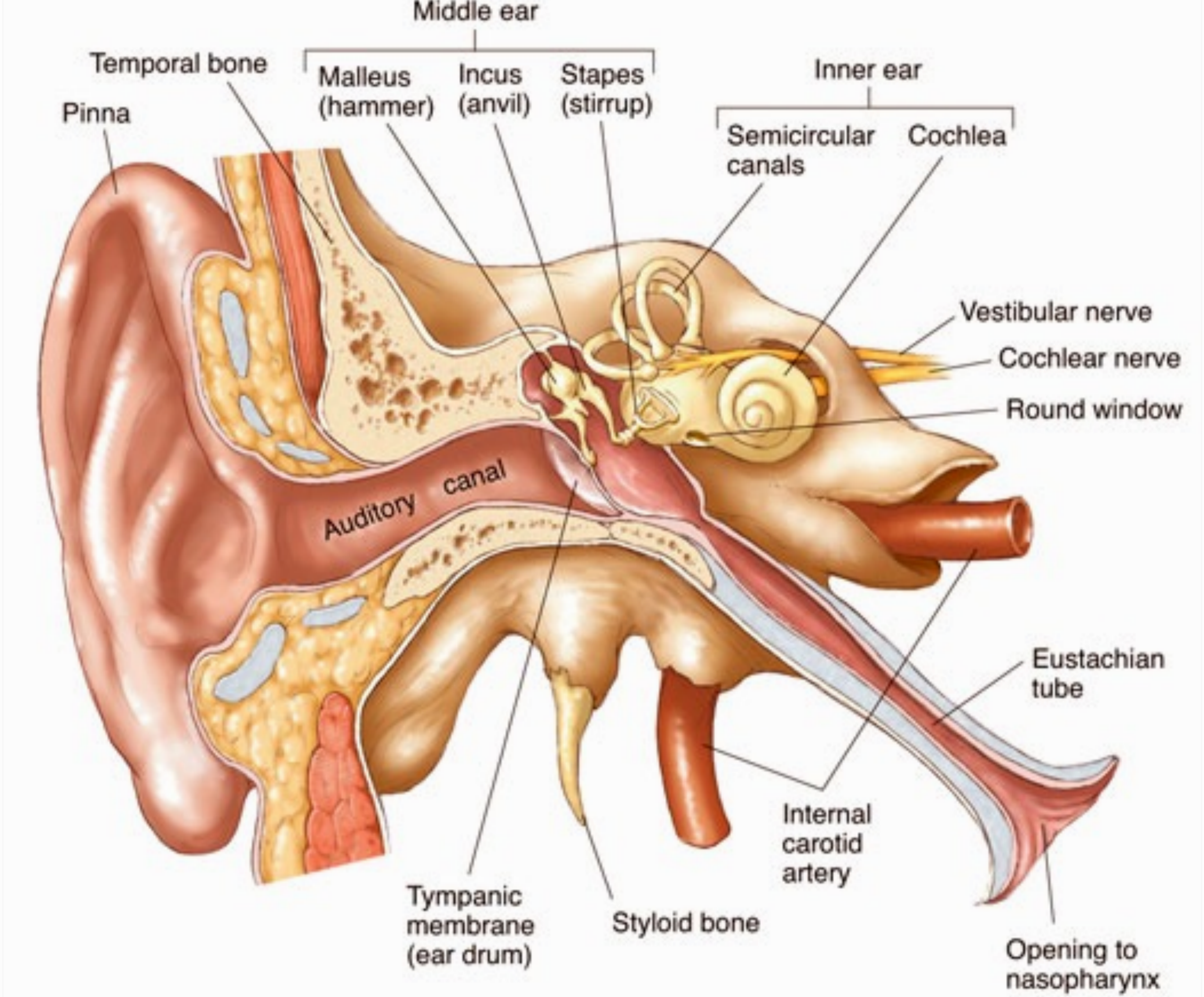
One nice feature of speech perception research is that everyone pretty much agrees about what it is we have to account for.

Disagreements arise from the fact that we are trying to study something that happens entirely within the privacy of our minds.

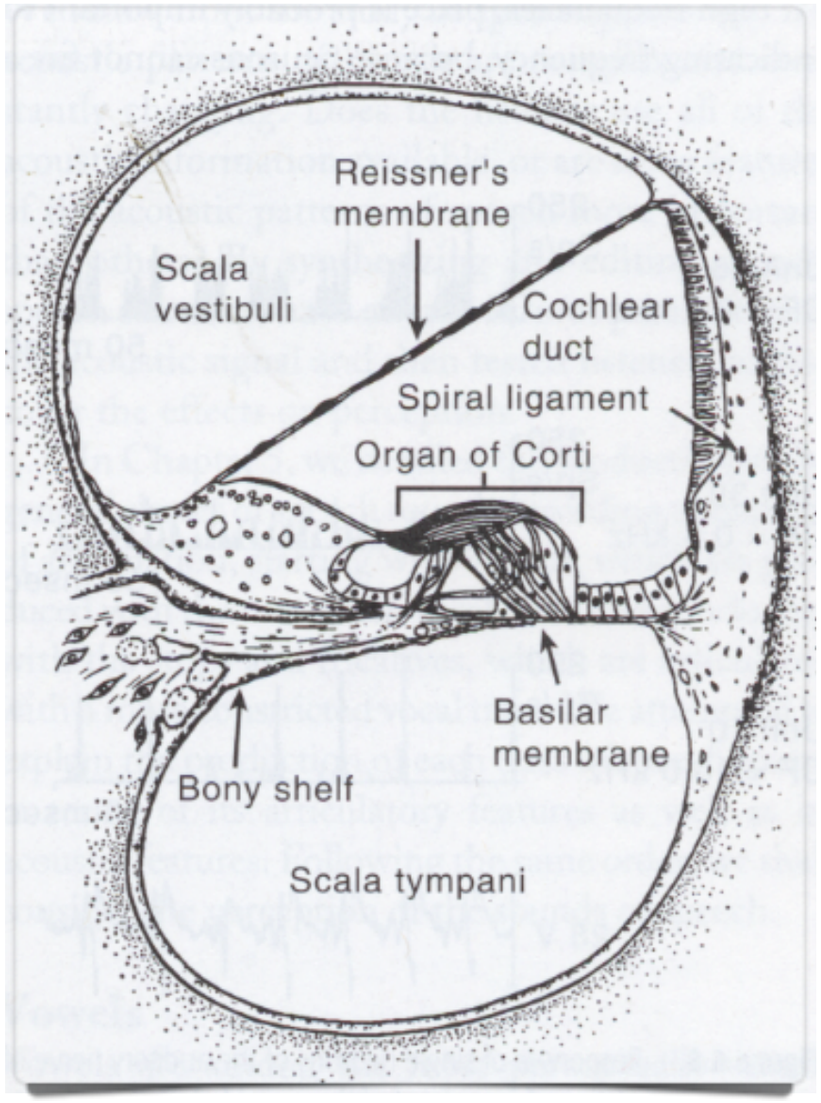
(Something even the experiencer only has occasional conscious awareness of.)

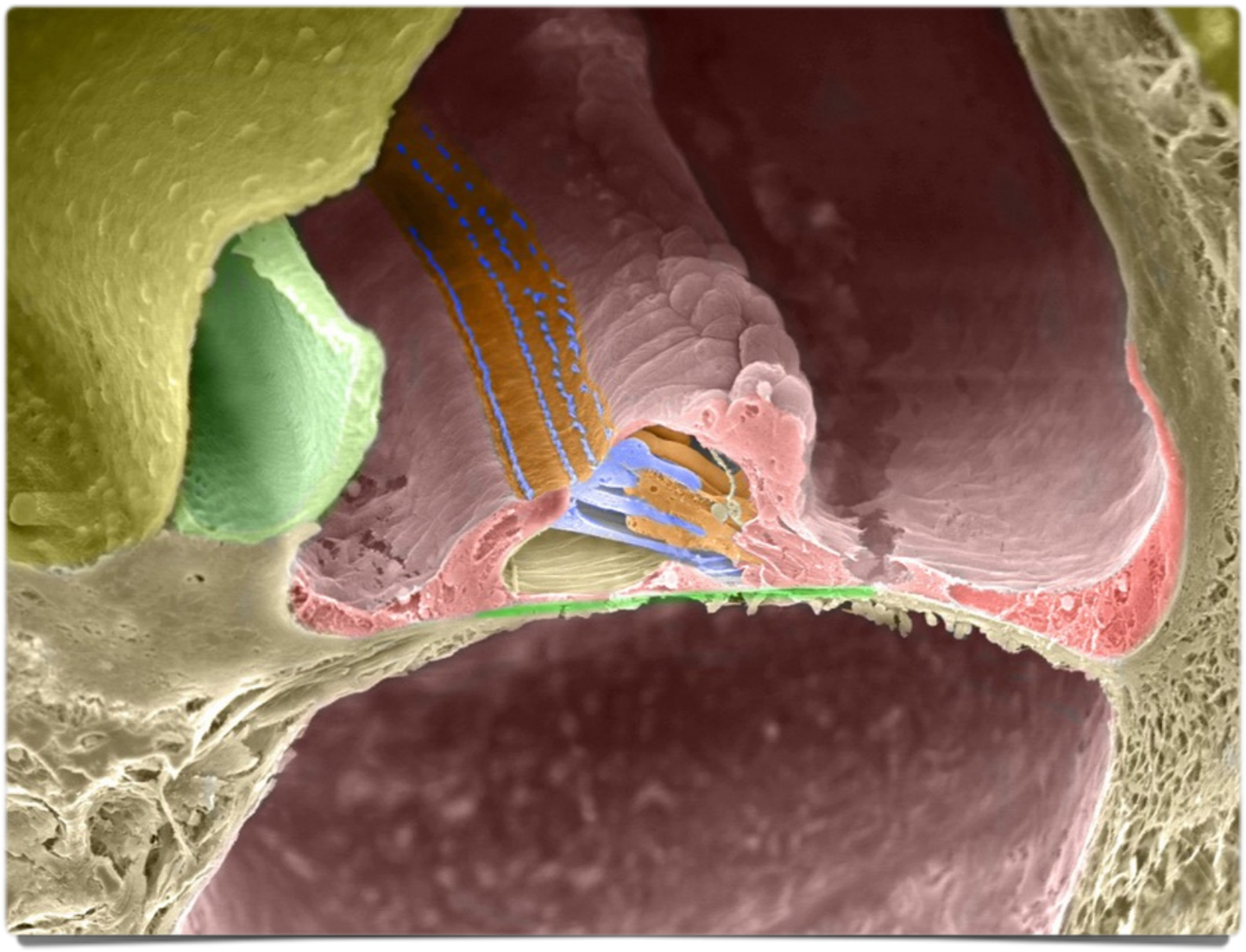
So what *can* we know about the auditory signal?

- Sound is movement.
- Hearing is a highly-specialized sense of touch.
- The ear, even a healthy ear, introduces a number of non-linearities and distortions to the acoustic signal


Middle ear
Temporal bone
Malleus (hammer)
Incus (anvil)
Stapes (stirrup)
Inner ear
Pinna
Semicircular canals
Cochlea
Vestibular nerve
Cochlear nerve
Round window
Auditory canal
Eustachian tube
Internal carotid artery
Tympanic membrane (ear drum)
Styloid bone
Opening to nasopharynx

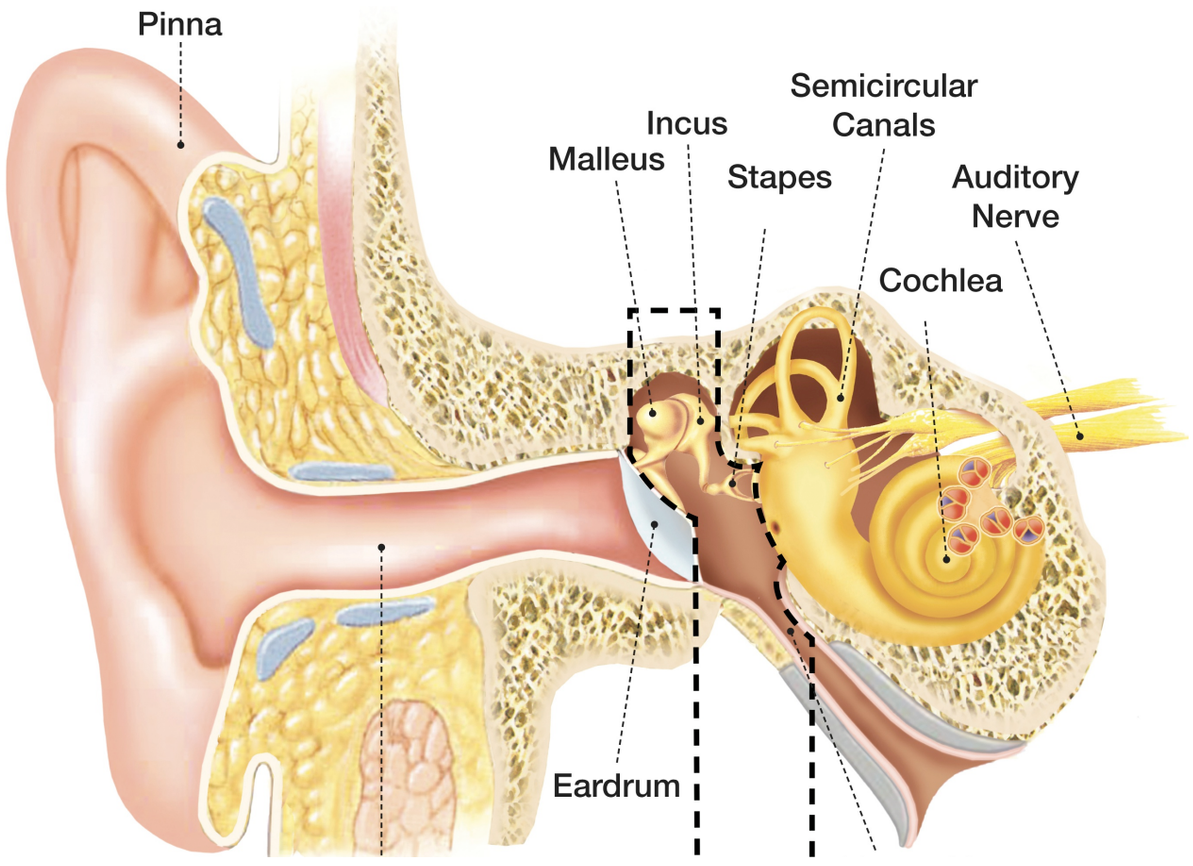
Organ of Corti





https://www.youtube.com/watch?v=dyenMluFaUw

BTW **that** was not the auditory system…



peripheral

central

Hearing ability

- **Intensity:** A healthy ear can transduce a vibration which causes the eardrum to move about the diameter of a hydrogen atom

- **Frequency:** A healthy ear can transduce frequencies as low as 20 Hz and as high as 20,000 Hz

- The transmission of information to the auditory cortex of the brain involves the highest information transmission rate in the human nervous system.

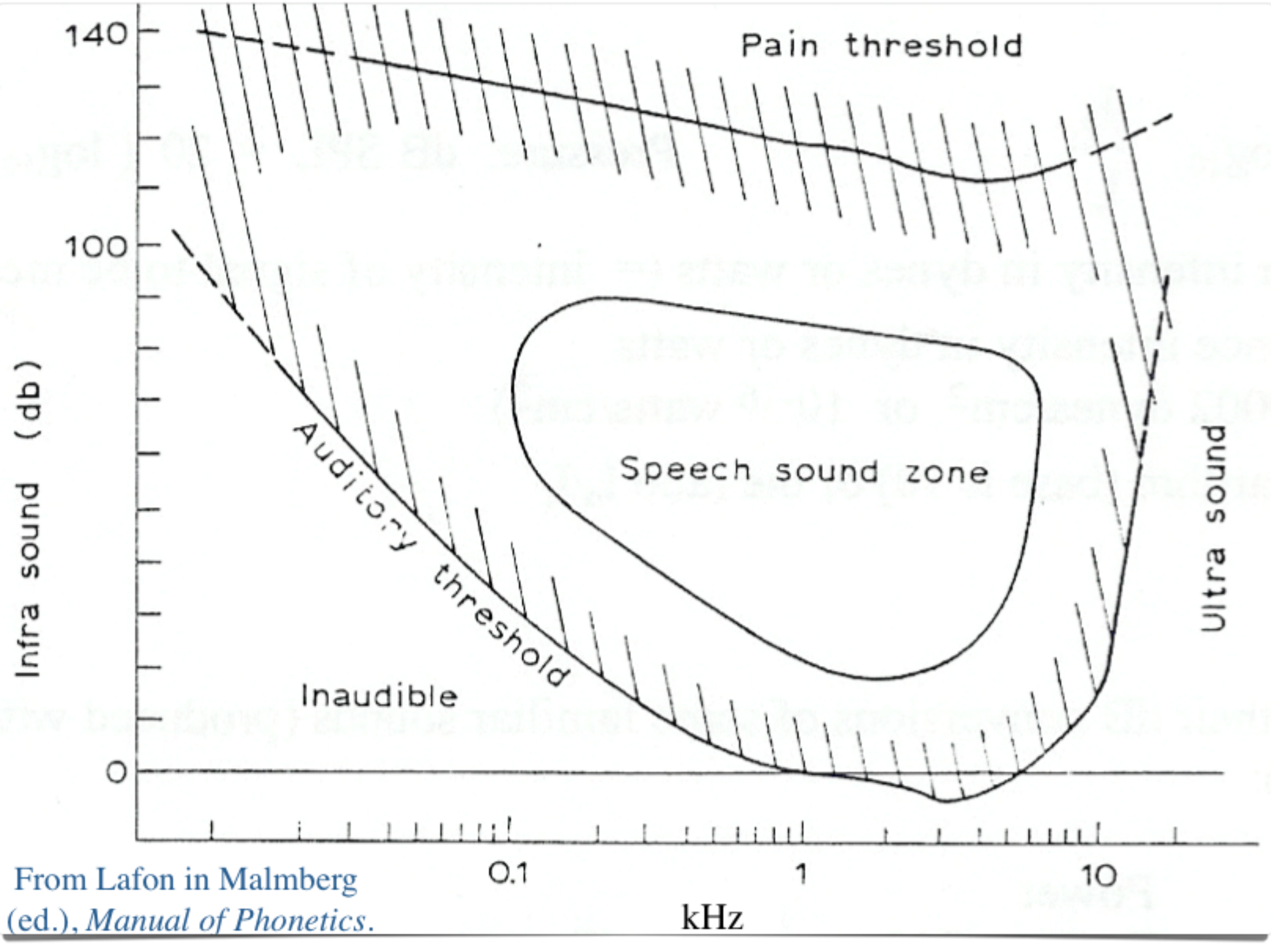
Loudness & Intensity

- Our sensation of loudness correlates closely with the measurable intensity of a sound.

- From the most to the least intense sound that the human ear can transduce (without damage), the ratio of intensities is 1,000,000,000,000:1

- But the **subjective loudness** differences that sounds evoke in listeners is nowhere near that great.

- Therefore, a commonly used scale for measuring intensity, the decibel scale (dB), reflects this nonlinearity of loudness perception.



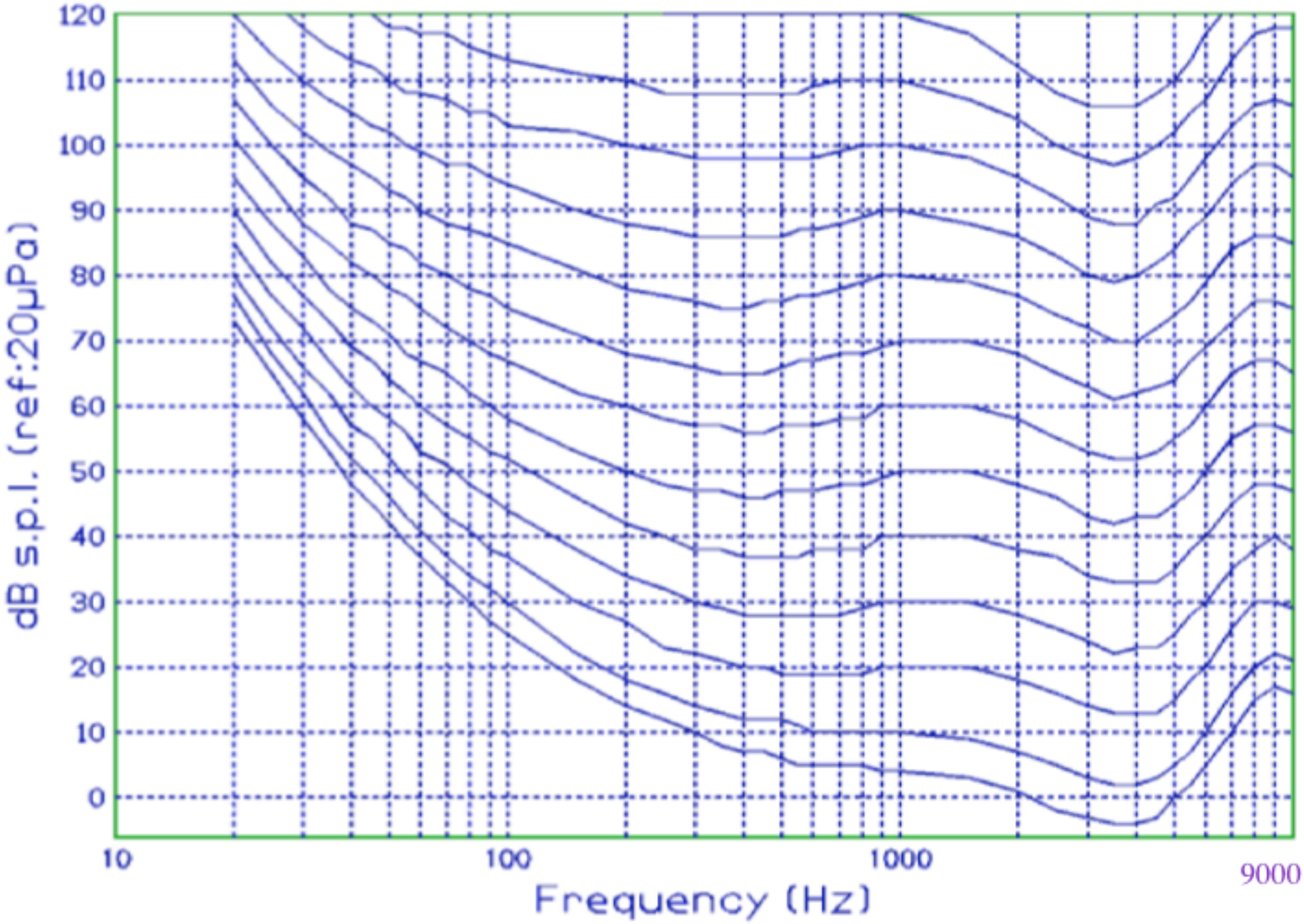
From Lafon in Malmberg (ed.), *Manual of Phonetics*.

Relative Loudness Scales

- **Phon**: The phon scale is determined by having listeners adjust the intensity of a **1000 Hz** tone until it has the **same loudness** as a comparison tone of a different frequency. Sounds judged to have equal loudness in this way are assigned the same "phon" value (e.g., all tones judged as having the same loudness as a 20 dB 1000 Hz tone have a loudness of 20 phons).

- **Sone**: The sone scale is determined by having listeners adjust the loudness of a tone until it is **twice as loud**, or half as loud, as another tone. 1 sone = loudness of a 40 dB 1000 Hz tone. 2 sones = sound judged to be 2x as loud as this.

Phon: equal loudness scale



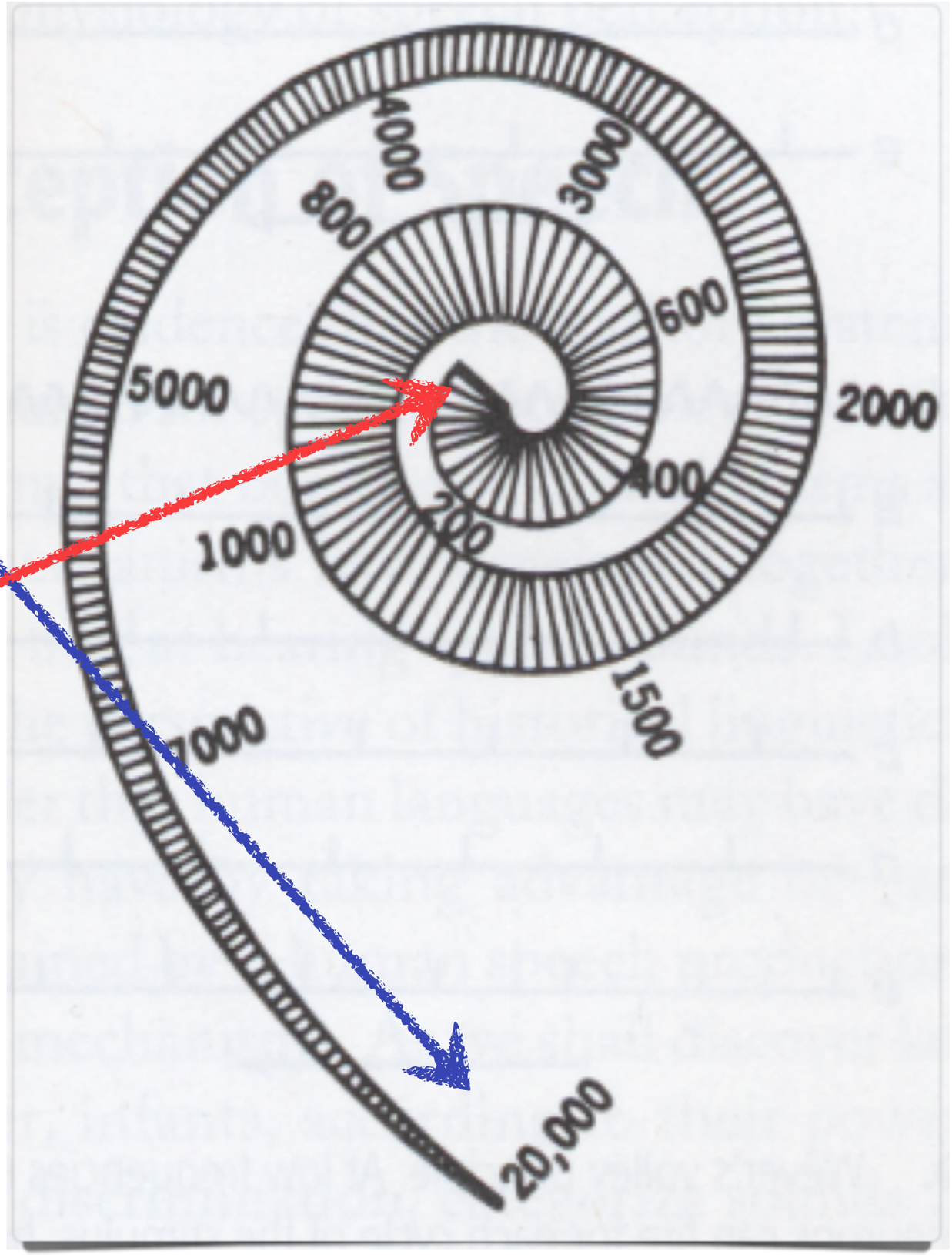
From http://www.ling.mq.edu.au/units/sph301/auditory_representations/intensity.html

Pitch & Frequency

- Just as the relation between loudness and intensity is non-linear, so is the relation between pitch and frequency.

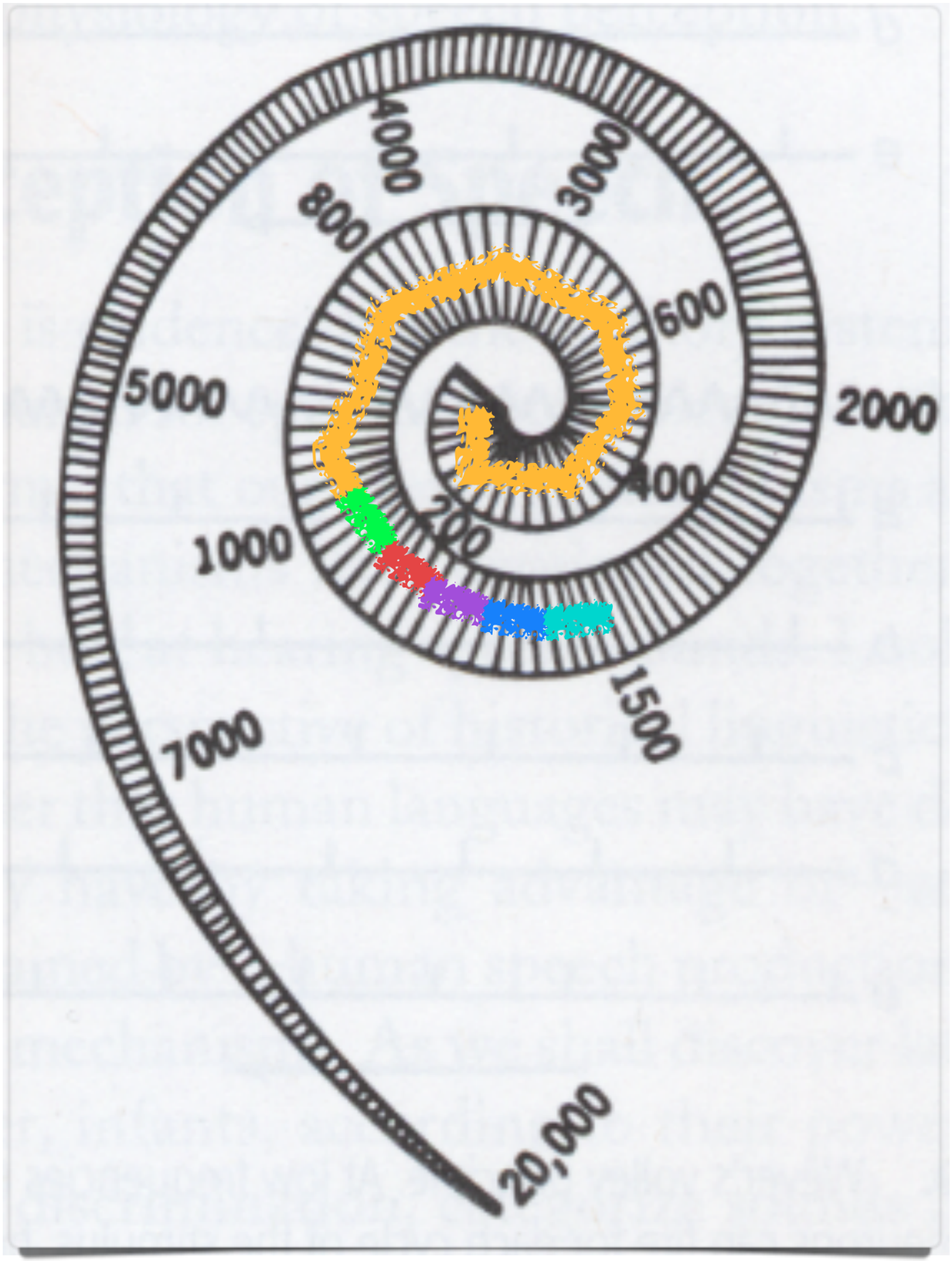
Basilar Membrane

- The basilar membrane is narrow and stiff at the basal (oval window) end, where it responds with greatest amplitude to high frequencies.
- At the **apical** end, where it is thicker and less stiff, the greatest amplitude of response is to low frequencies.



Basilar Membrane

- Thus the basilar membrane is a spectrum analyzer, performing a kind of Fourier analysis on complex waves.
- The largest portion of the basilar membrane responds to sounds in the 20-1000 Hz range.



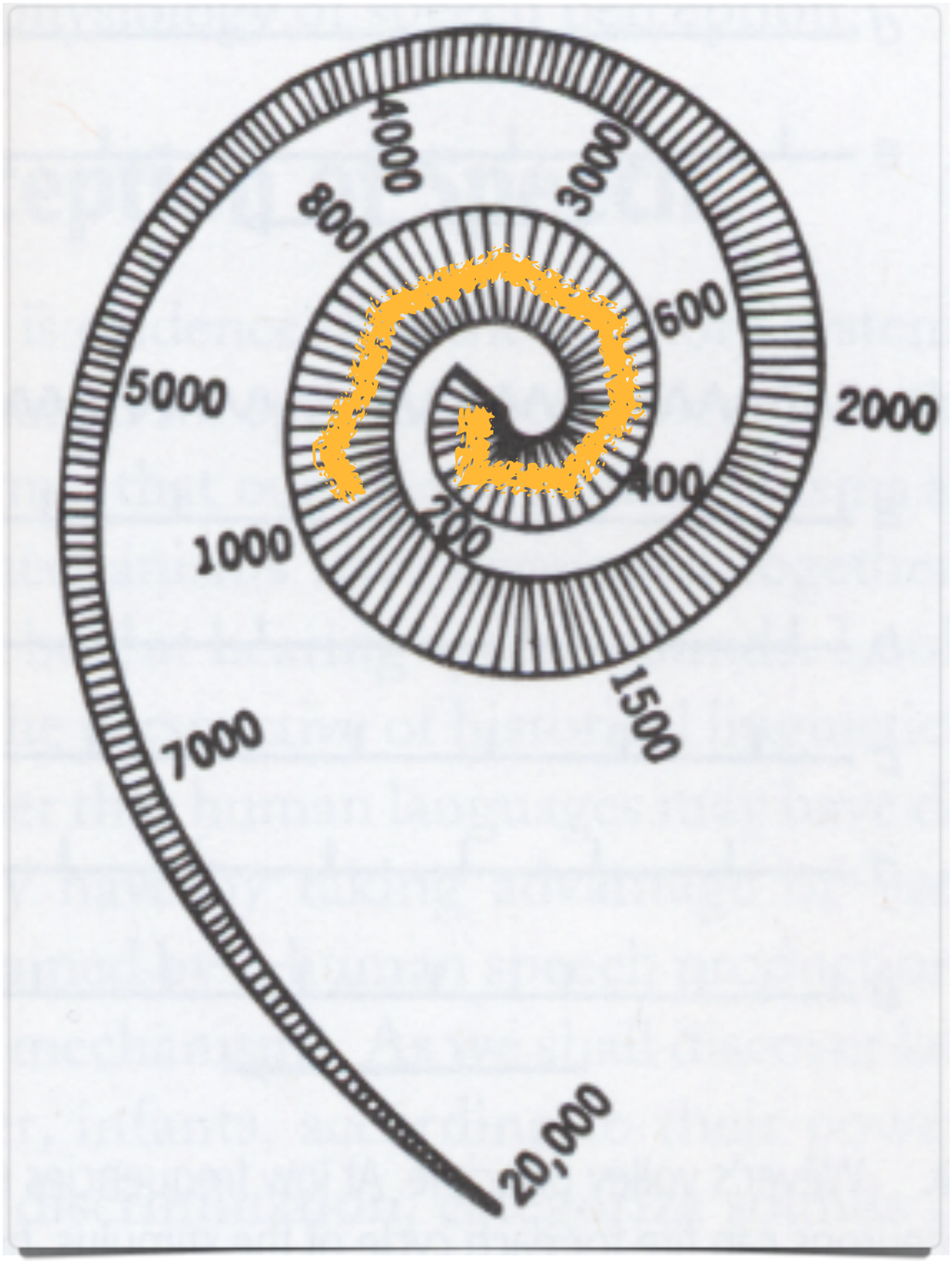
Basilar Membrane

- This non-linearity has psychoacoustic consequences:

- **the mammalian ear is more sensitive to fine frequency differences in the lower than in the higher frequencies.**

- We can hear a difference of about 1 Hz at 1,000 Hz

- We need more



Gabor Uncertainty Principle

- As in all signal processing, the cochlea's response has a time-frequency trade-off:

- Basal end (high frequencies) provides poorer frequency but better temporal resolution;

- Apical end (low frequencies) provides better frequency but poorer temporal resolution.

How does this time-frequency trade-off relate to what we know about speech sounds? Which speech sounds have primarily high frequency energy? Which have fine frequency differences? What are their time courses like?

Pitch & Frequency

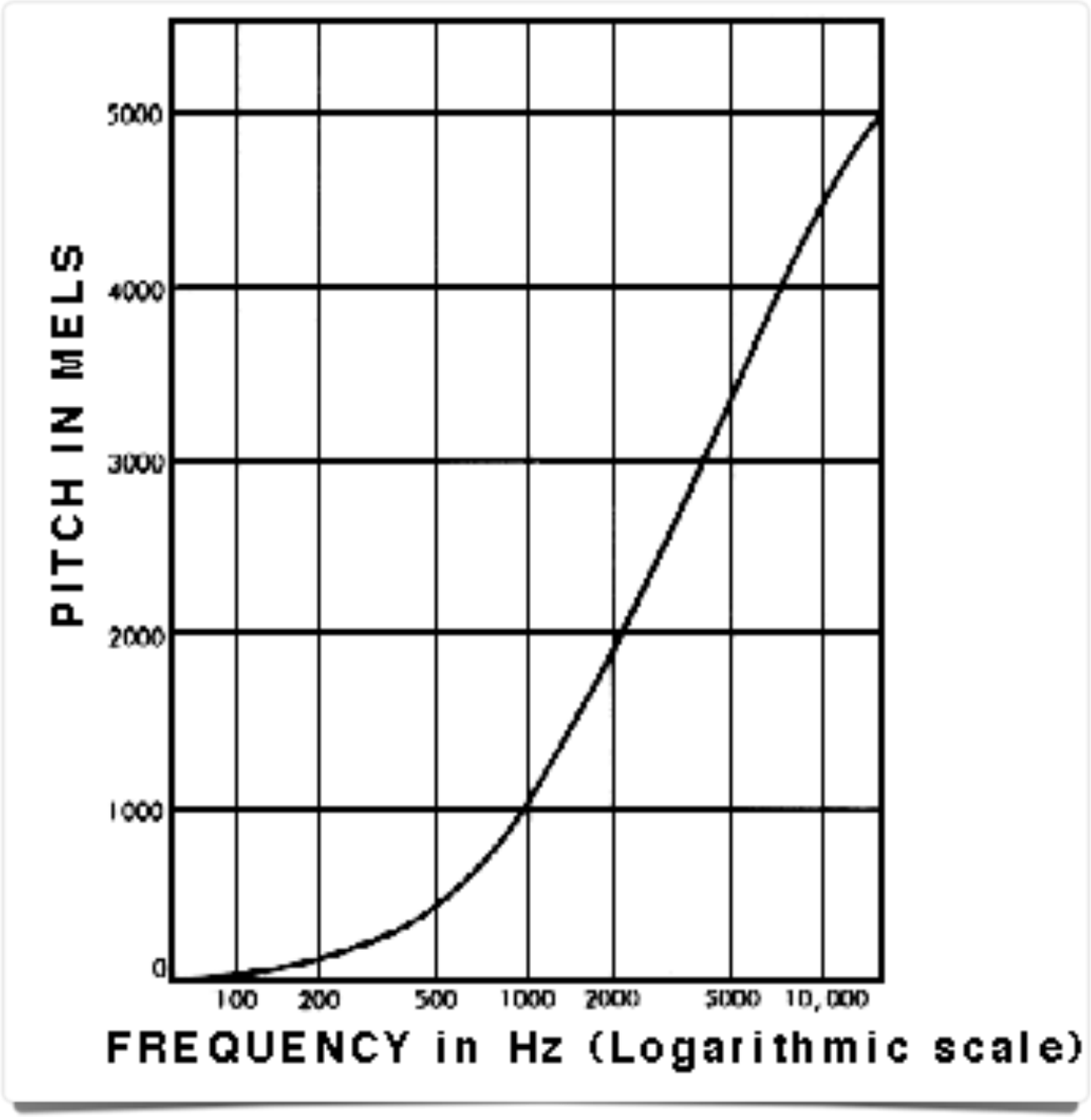
- Just as the relation between loudness and intensity is non-linear, so is the relation between pitch and frequency.

- **Mel**: Scale is based on experiments with pure tones in which listeners adjust the frequency of a test tone to be half as high (or twice as high) as that of a comparison tone.

- **Bark**: Scale is intended to be a frequency scale in which equal distances between frequencies are perceived by listeners as equally distant

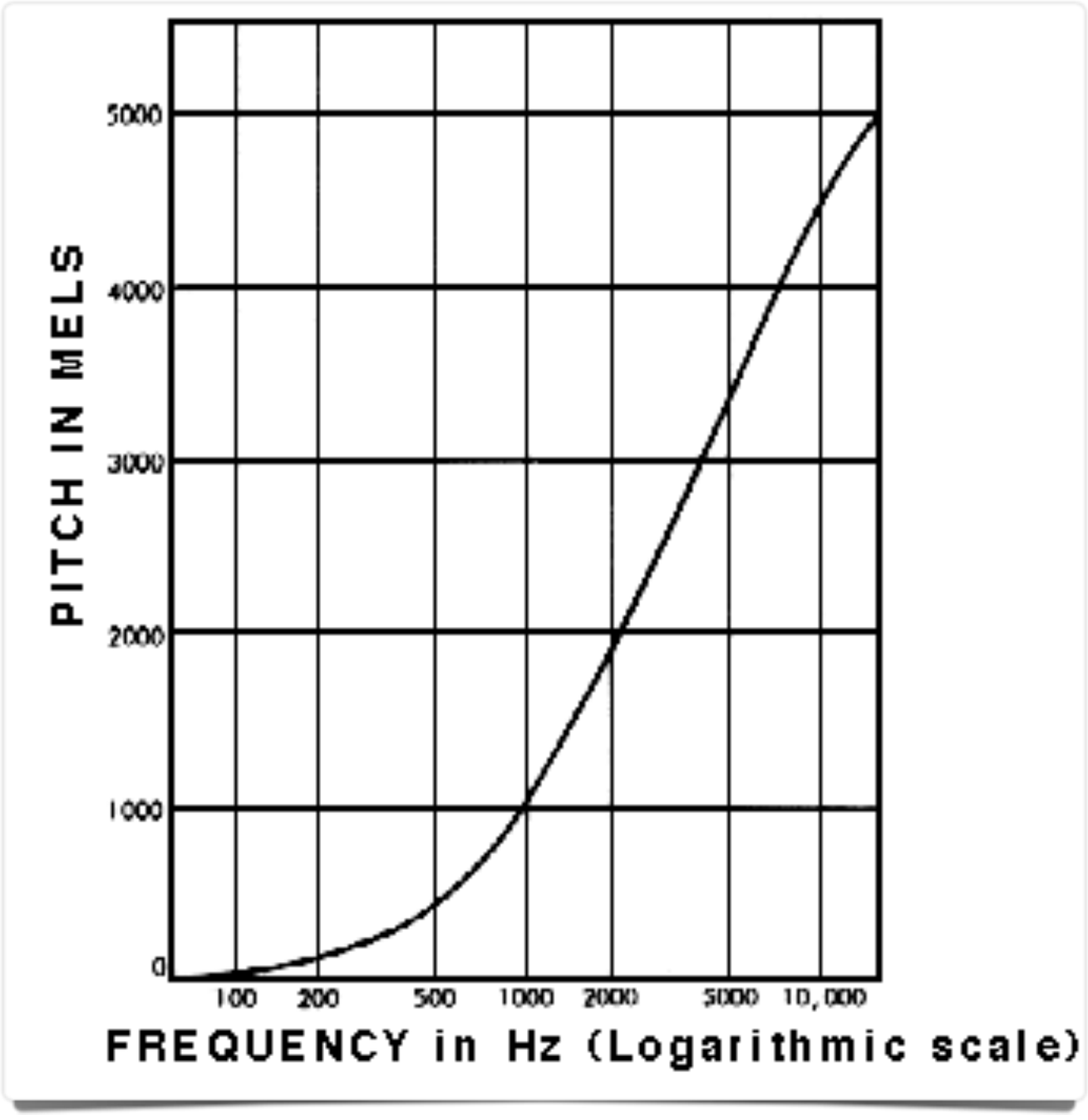
Mel Frequency

- **Mel**: The mel scale is based on experiments with pure tones in which listeners adjust the frequency of a test tone to be half as high (or twice as high) as a comparison tone.



Mel Frequency

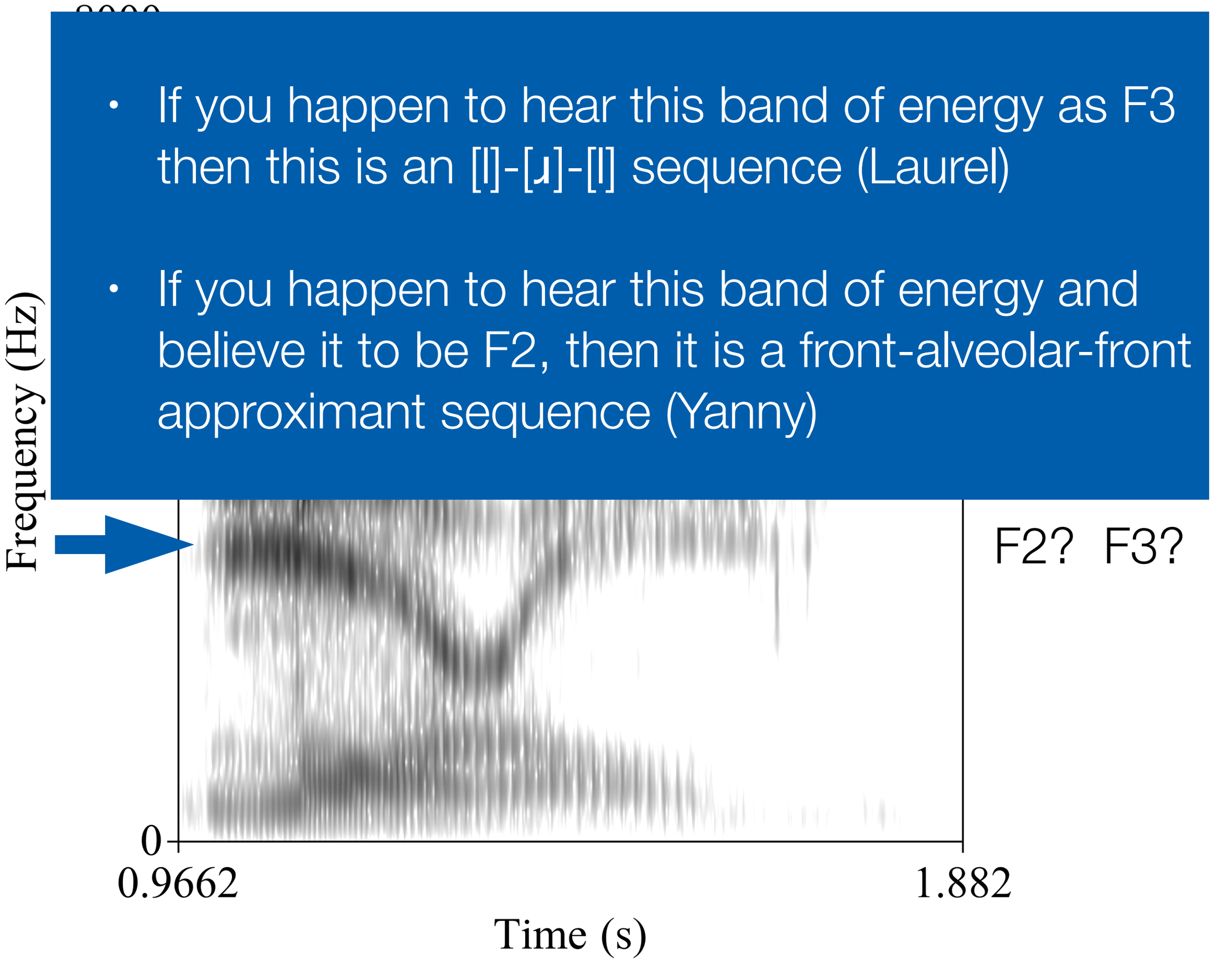
- 1,000 mel = pitch of 1,000 Hz tone (by def)
- 500 mel = pitch of tone that sounds half as high.
- The mel scale corresponds to Hz up to ~500 Hz. At higher frequencies the mel scale is (nearly) logarithmic.

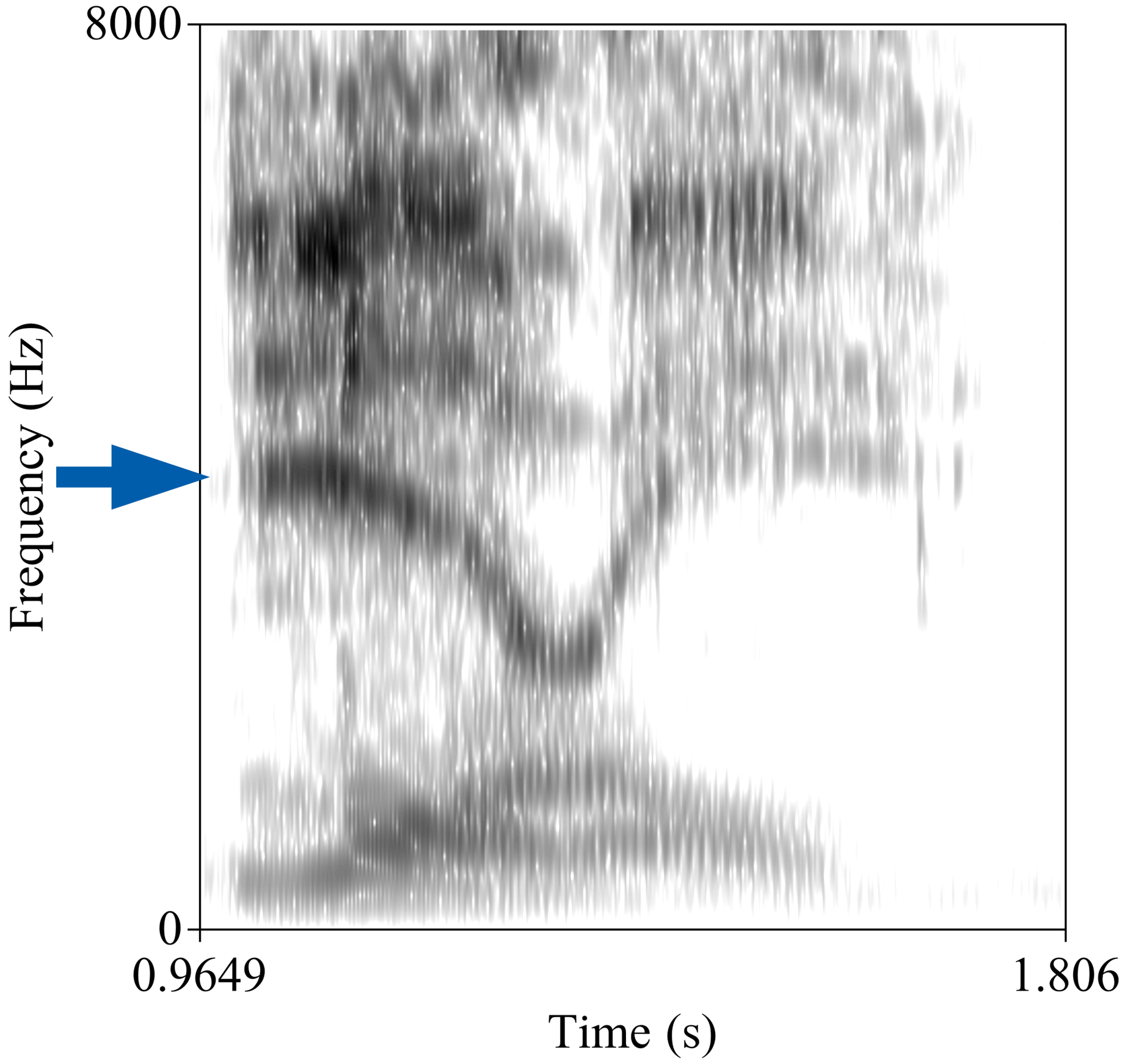


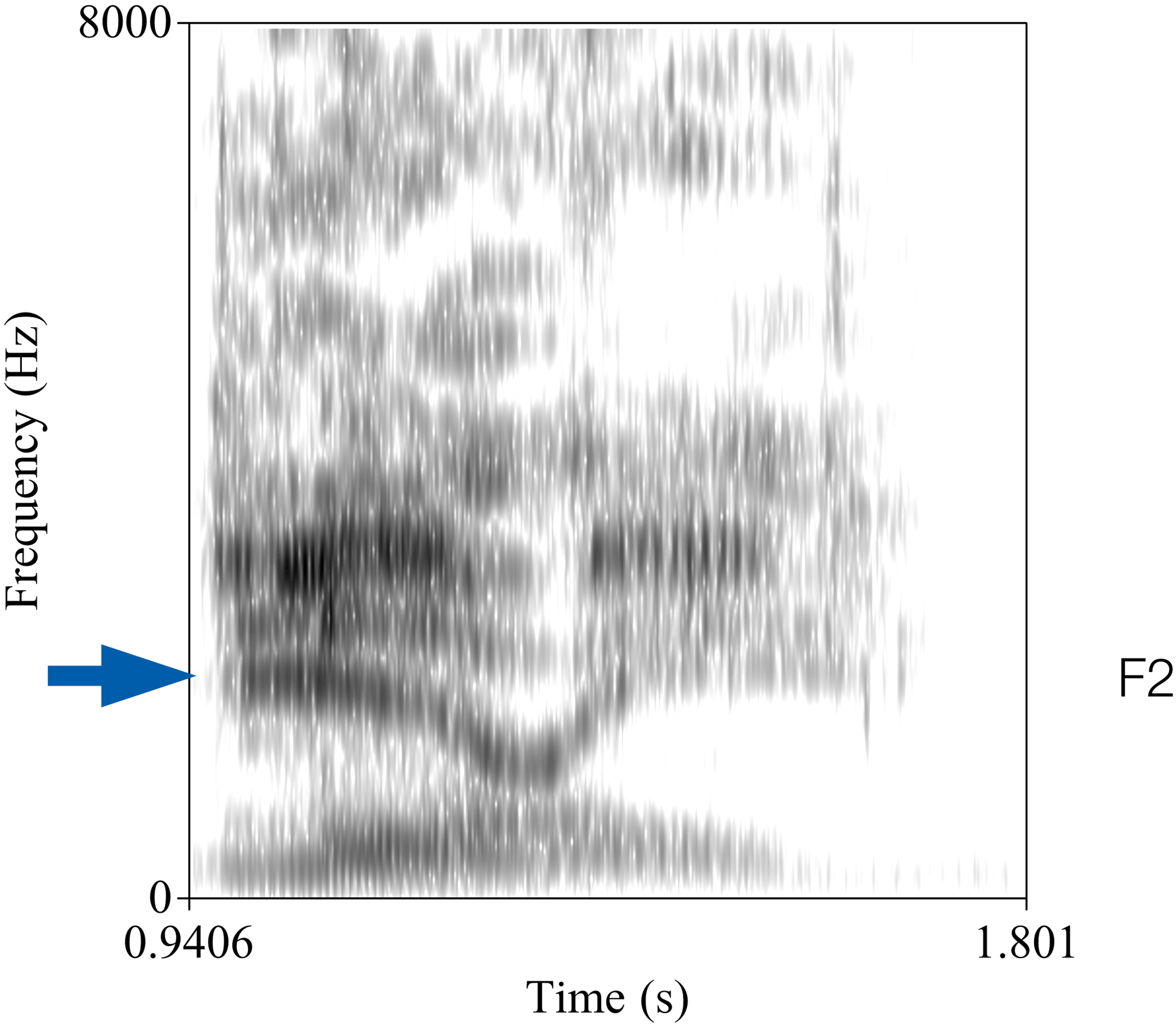
Individual Differences

Yanny vs Laurel was a hugely popular demonstration of the fact that different people listen differently.

- If you happen to hear this band of energy as F3 then this is an [l]-[ɹ]-[l] sequence (Laurel)
- If you happen to hear this band of energy and believe it to be F2, then it is a front-alveolar-front approximant sequence (Yanny)




8000
0
0.9649
1.806
Time (s)
Frequency (Hz)
F3


8000
Frequency (Hz)
0
0.9406
1.801
Time (s)
F2

General Auditory Theories

- Frankly, GA theories tend to be defined in terms of *not* being gestural theories.

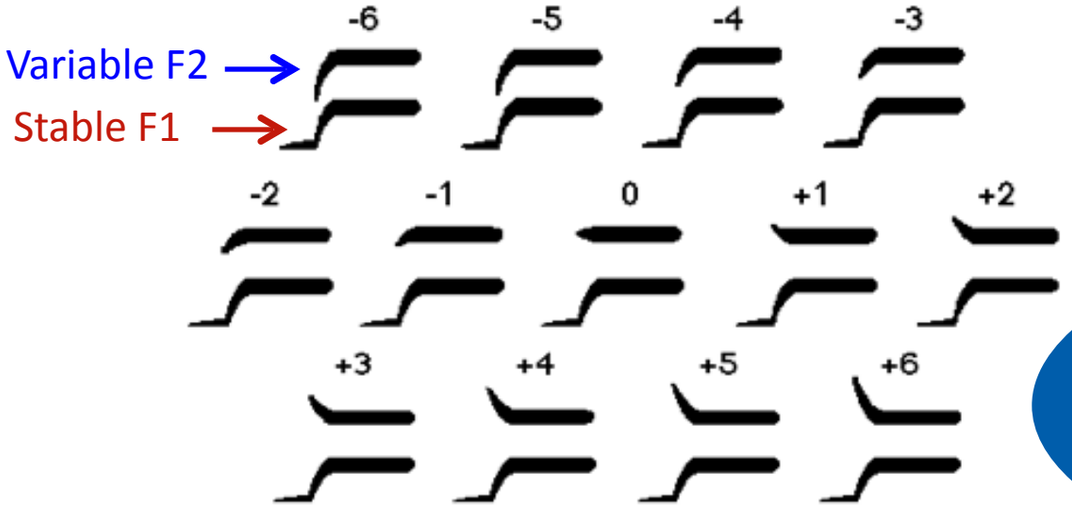
- "We hear sounds" therefore "sounds are the object of speech perception" is the "common sense," non-fanciful account of human speech perception and is even discussed this way in the literature.

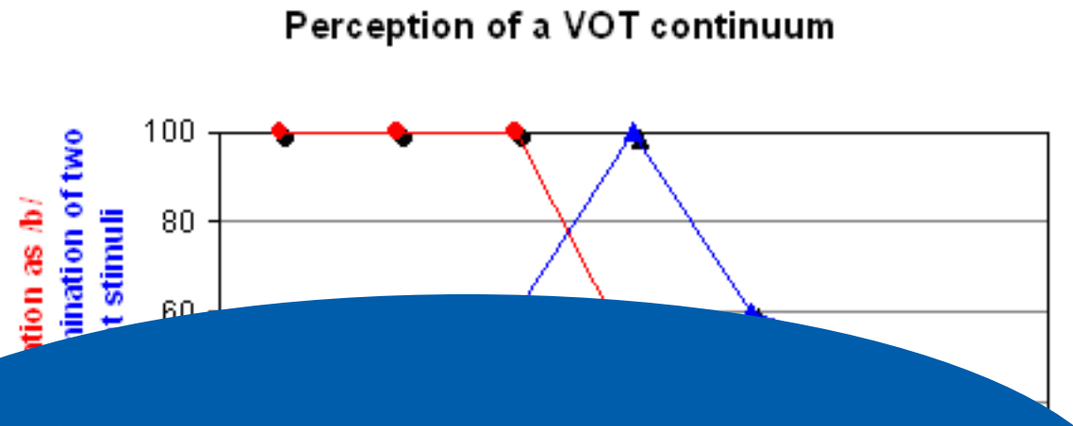
- The general in "General Auditory" means that listeners draw on general (not language-specific) cognitive resources to hear speech

- But being "common sense" doesn't make perceptual phenomena any easier to explain

Categorical Perception

- For GA theories, Categorical Perception is evidence of learning in humans.





Can you think of any ways to test this?

- Humans learn the relationships between auditory differences and category boundaries.

Compensation for coarticulation

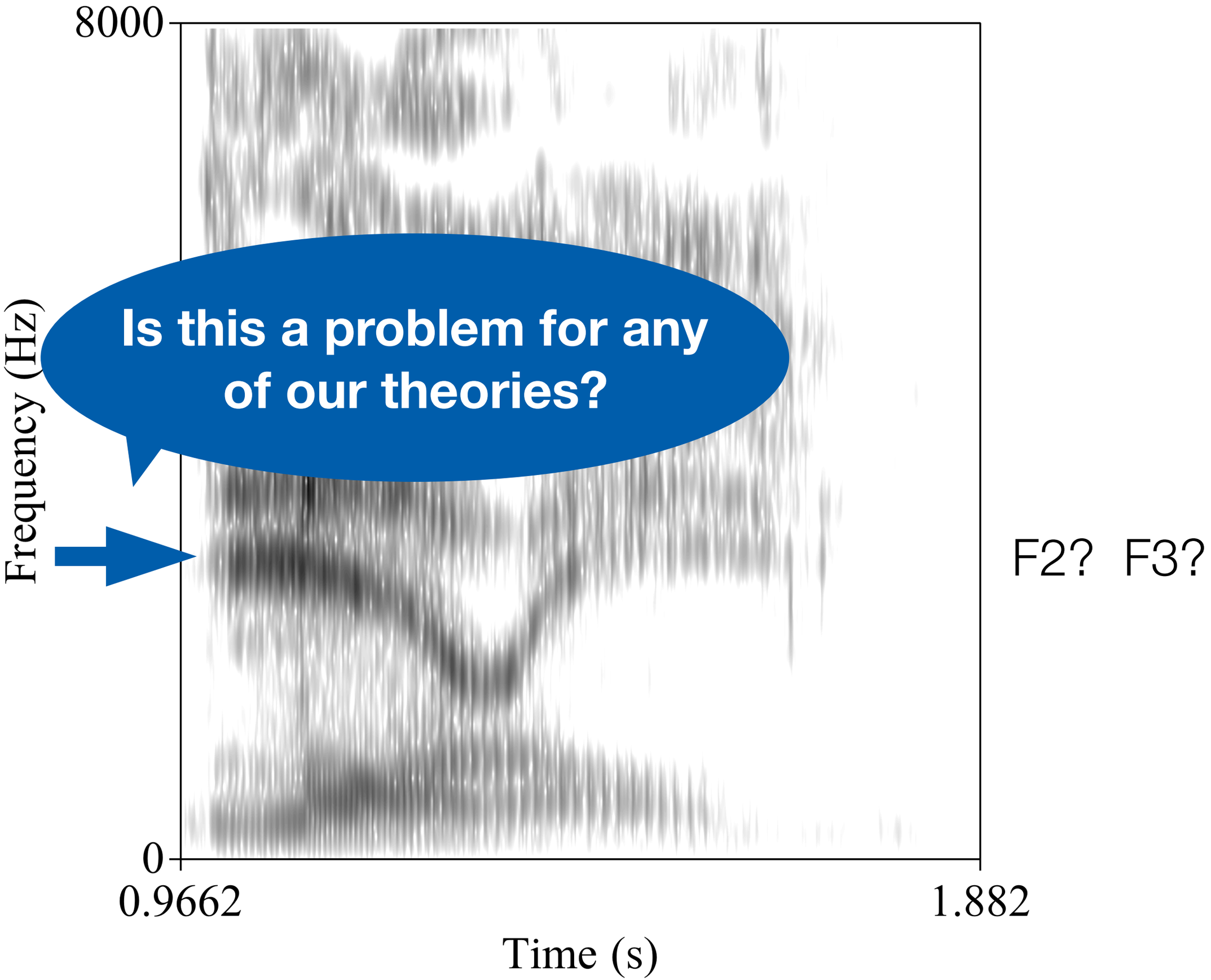
- As you may recall from day 2, General Auditory theories explain compensation for coarticulation as spectral contrast

- Lotto & Kluender (1998) used pure tones to mimic low F3 (/aɹ/) and high F3 (/al/) before an ambiguous [da]/[ga] syllable

- Because the tone is non speech, L&K argue that compensation is spectral contrast so the object of perception is **frequency of the sound itself** not the information this frequency might give the listener about gestures.

McGurk (and MacDonald) Effect

- What does Motor Theory say about McGurk?
- What does Direct Realism say about McGurk?
- What might General Auditory theories say about McGurk?


8000
Frequency (Hz)
Is this a problem for any of our theories?
F2? F3?
0
0.9662
1.882
Time (s)

Bill Bill Bill Bill Bill…

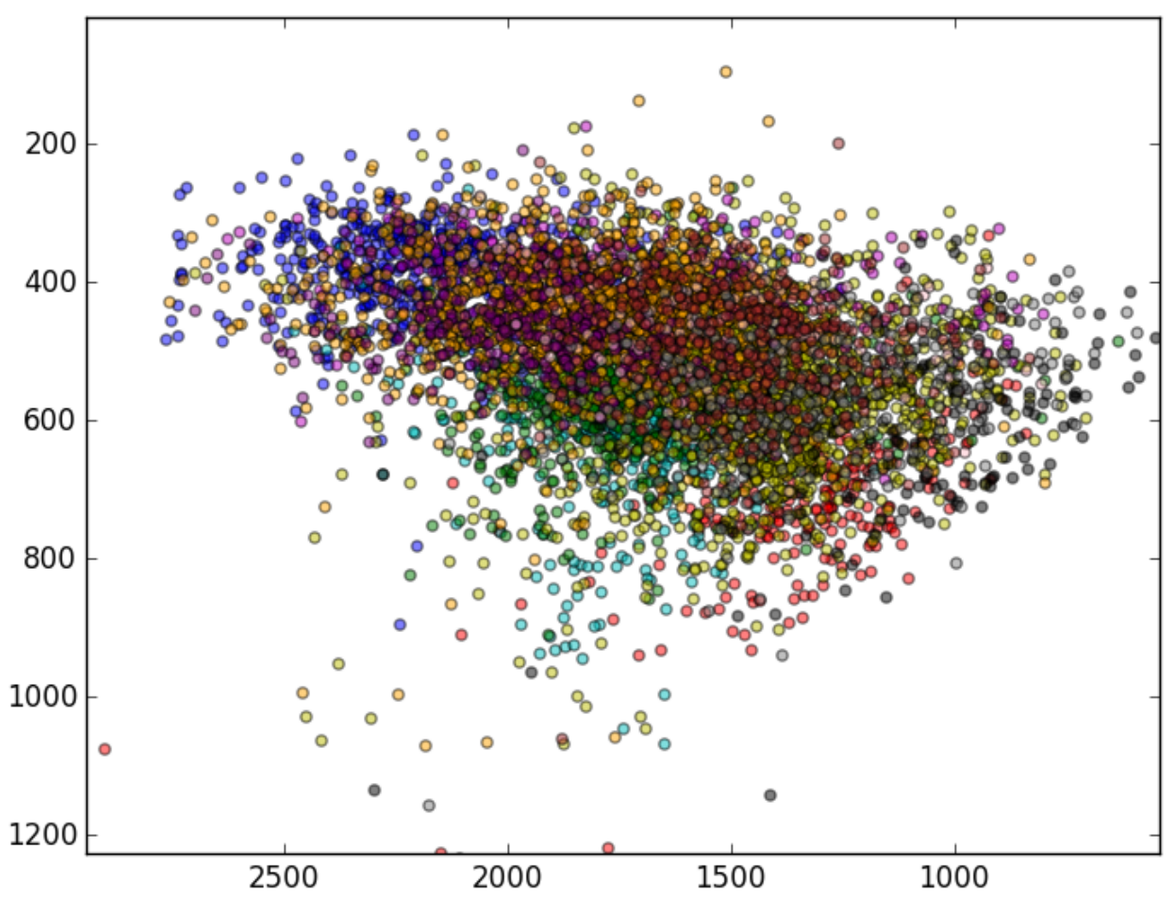
- How is this phenomenon different from/similar to the McGurk Effect?

- Is this a problem for any of the three major theories so far?

Vowel Overlap

- General auditory theories, in particular, face the problem of vowel overlap through a process of **normalization** through which 'irrelevant' information about the linguistic unit (coarticulation, gender, age, height, emotion, etc.) is discarded.

- We will devote more time to this on Thursday



General Auditory Theories

- This isn’t necessarily a problem (Occam’s razor is no guarantee of accuracy), but GA theories start to feel more like a collection of solutions and less like a general overarching theory of perception.

- Every phenomenon requires a different kind of explanation (normalization, spectral contrast, learning, auditory enhancement, training Japanese quail, etc.) that may coalesce into a theory, but may also seem ad hoc to critics

Exemplar models

- A proposal that offers a single coherent model that can be used to predict or explain a wide range of speech perception phenomena

- Exemplar models build on the GA response to, e.g. Categorical Perception and the McGurk effect to say that humans rely on learning patterns from experiences with speech to perceive novel speech inputs

Declarative Memory

- There are two types of declarative memory:
  - Semantic memory
    - General facts like “The first moon landing was 50 years ago this month”
  - Episodic memory
    - Personal facts like the experience of watching or listening to the first moon landing 50 years ago

Episodic traces

- Exemplar models propose that we store our personal memories of experiences with speech.

- These memories are stored along with feedback about whether we seem to have correctly understood what was said to us.

- For example: [‘kæt]

Issues that arise in acoustics-to-phoneme mapping

- Segmentation problem
- Lack of Invariance problem

How we frame the question matters.

- What do we mean by “linguistic message?” What is the domain or level of the message?

Feature?

Gesture?

Syllable?

Phoneme?

Word?

Meaning?

Exemplar models: Spoken word perception

- Does this solve the Segmentation problem?
- Does this solve the Lack of Invariance problem?
- Does it settle the auditory/gestural debate?

Next time: Approaches to vowel perception!