Lecture 14 Speech Perception

Wu Xihong

Peking University
School of Artificial Intelligence

The Hypo

- Speech is not only the future of windows, but is also the future of computing

  --- Bill Gates

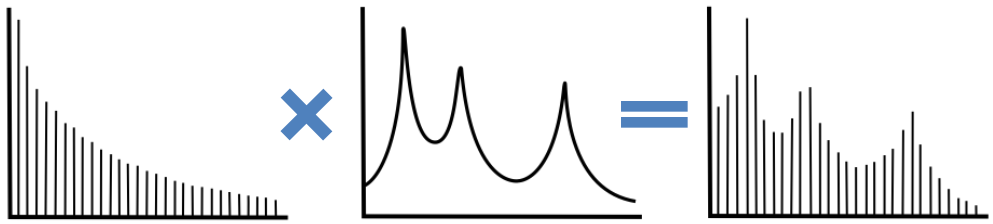
- The future of computing is speech

  --- Gordon Moore

Outline

- The Structure of Speech
  - ✓Articulatory & Acoustic Phonetics
- Why is speech recognition hard?
  - ✓lack of invariance & segmentation problems
- So how do we do it?
  - ✓ categorical perception
  - ✓ Top-down processing
- Acquisition of language

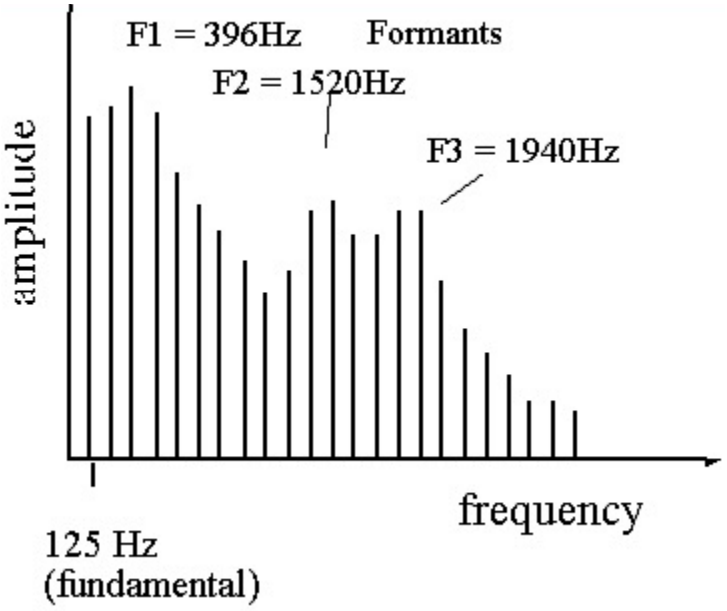
The Structure of Speech

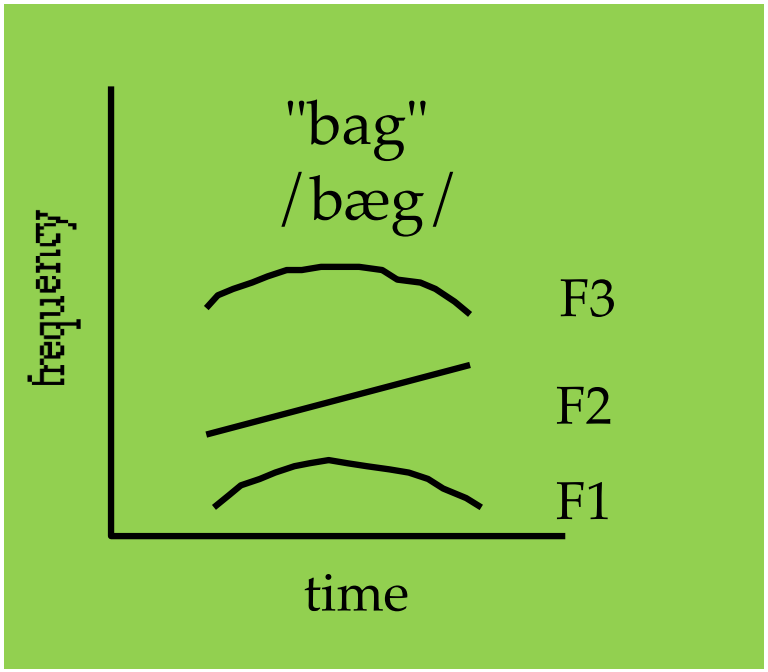


声门激励　　声道调制　　语音频谱

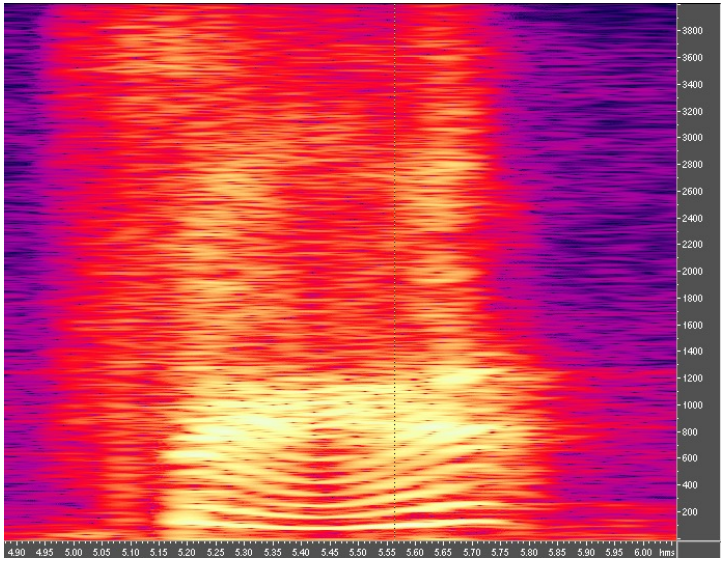
Pitch and Formants

1. Harmonics (giving pitch) produced by vocal cord vibration
2. Formant frequencies: resonances of the vocal tract
3. Formant frequencies change as you change the shape of your vocal tract

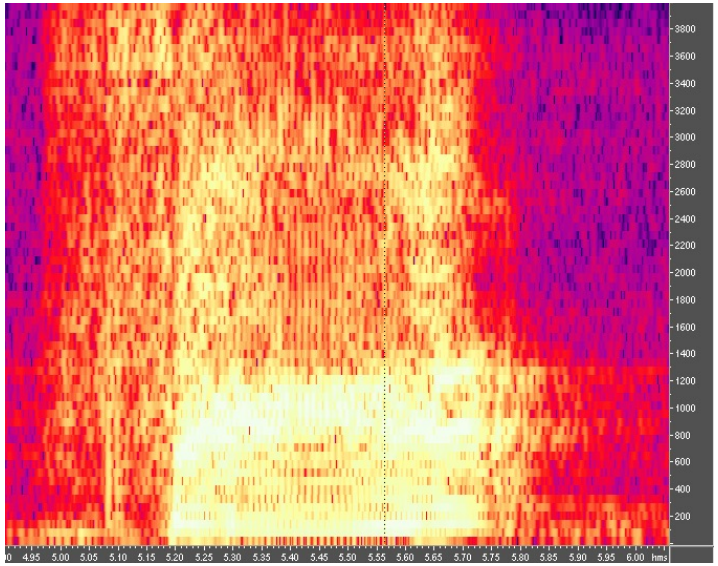




Wide and narrow band analysis

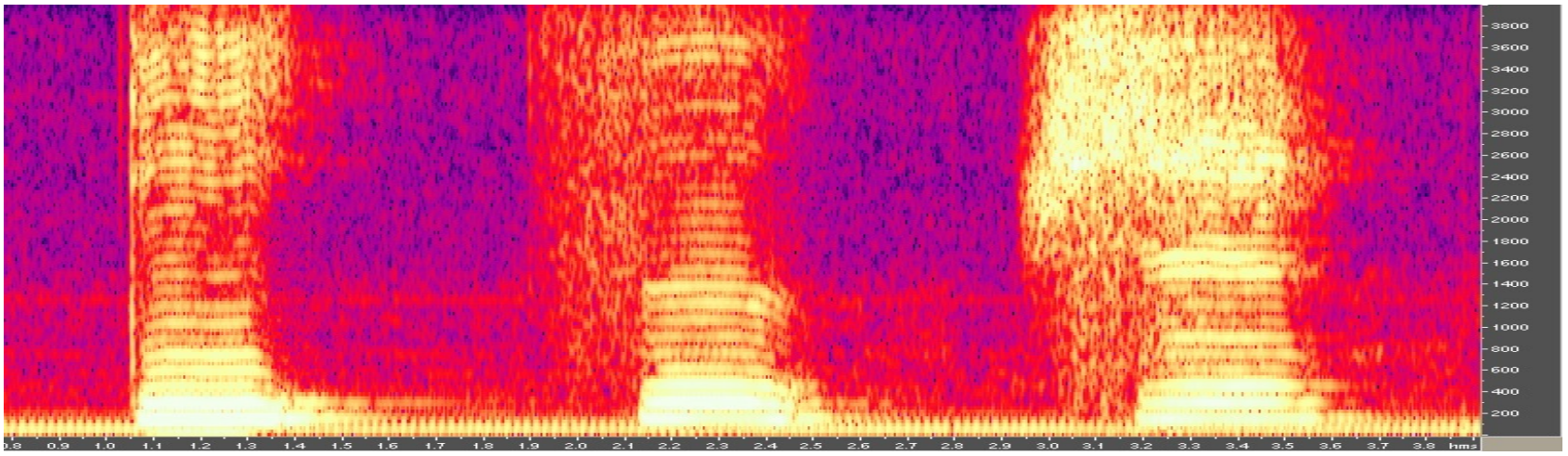


Narrow band spectrogram



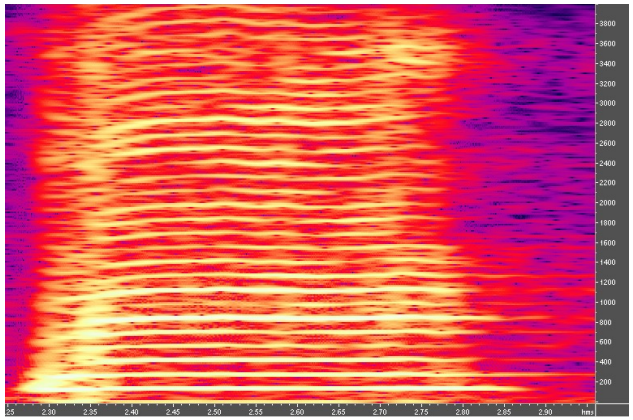
Wideband spectrogram

Consonant Production

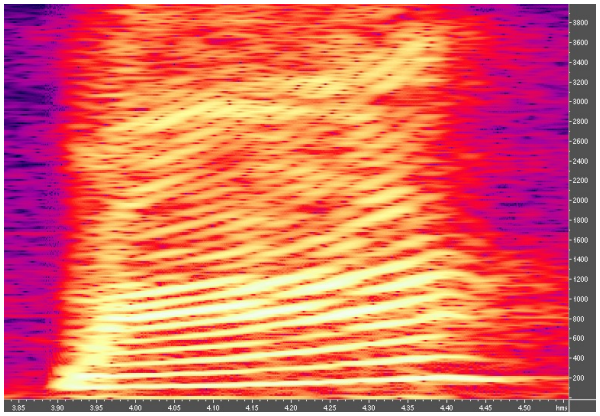


[d] [i] [s] [i] [m] [i]

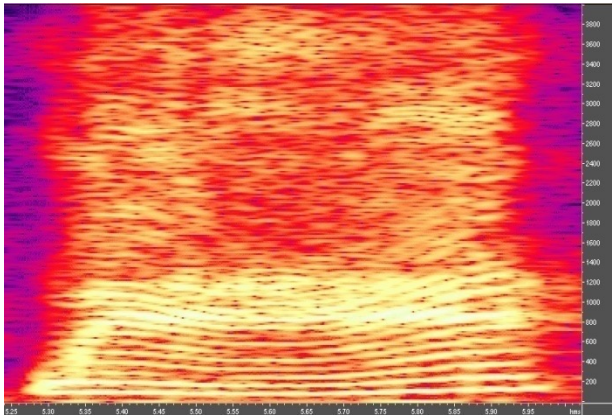
Tone pattern



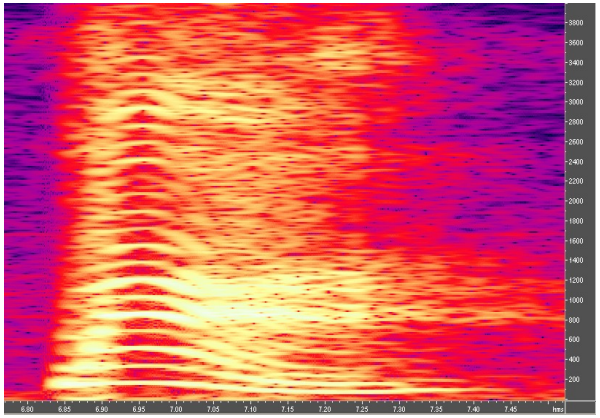
[ma55]



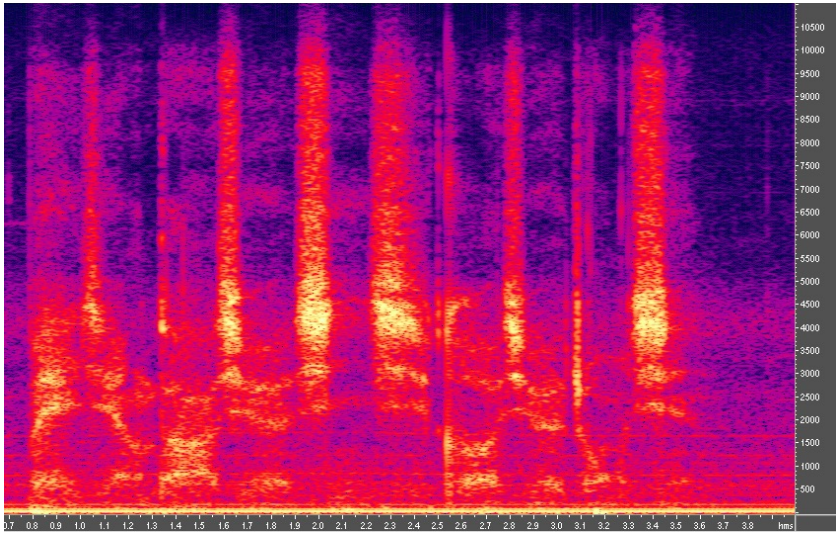
[ma35]



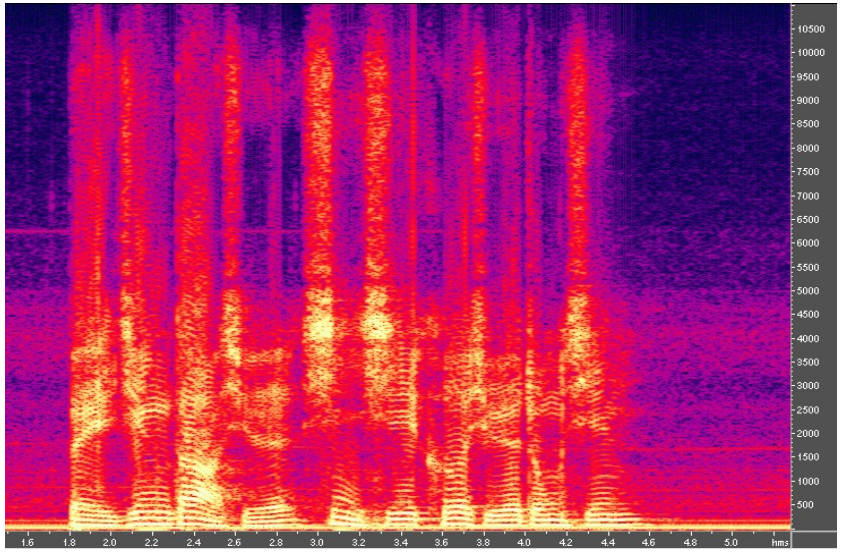
[ma214]



[ma51]



whispered



normal

Why is speech recognition hard?

Segments vs. Prosody

Segmental: consonants / vowels -> words (CV, CVC)

Prosodic: pitch contour, stress, emotion.

"I thought she was married?"
"I thought she was married."
"I thought she was married!"
"I thought she was married!"

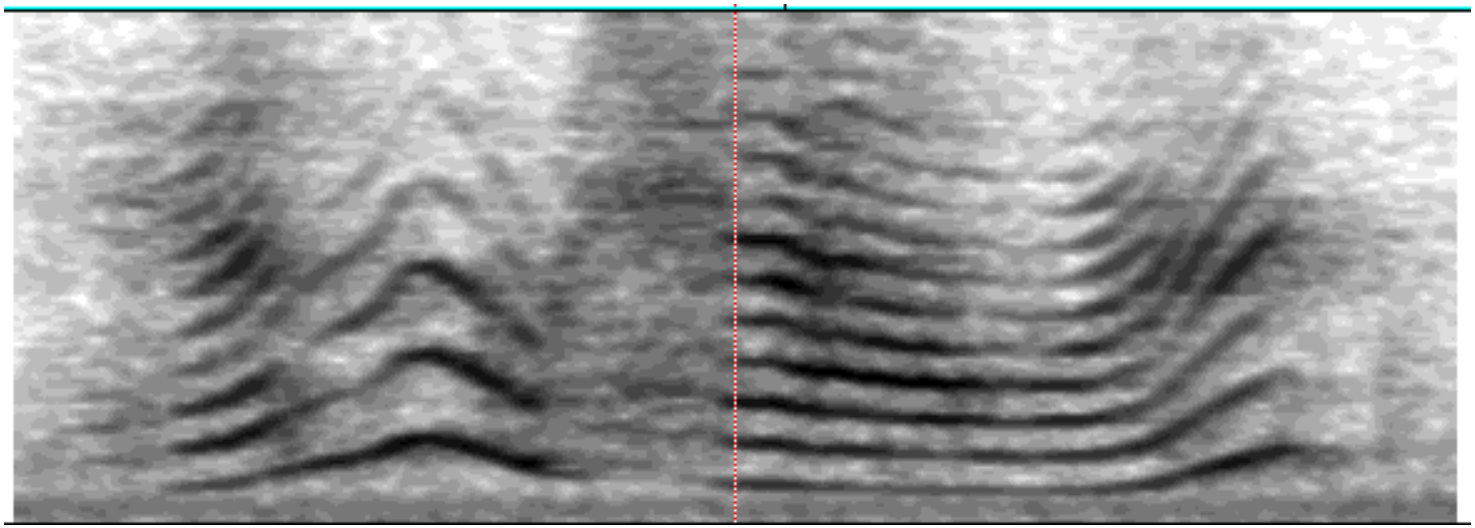
yes!

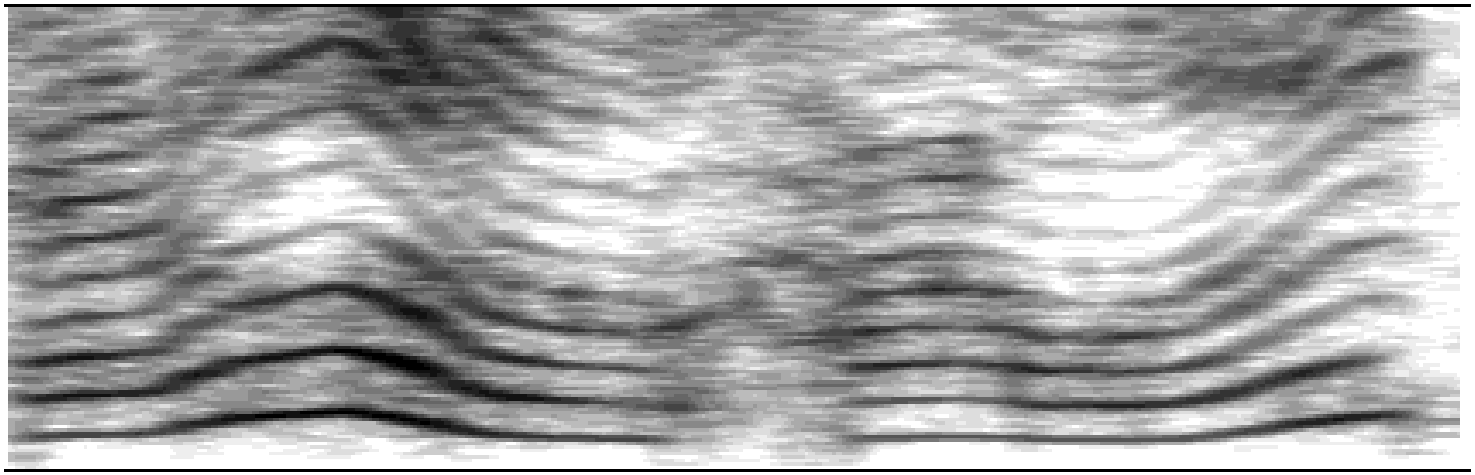
Are bird/mammal animal systems like human prosody?

generally use different pitch contours

Mynah bird speech

Klatt &Stefanski (1974) How does a mynah bird imitate human speech? J Acoust Soc Amer, 55, 822-832.





Mynah vs. Grey parrot

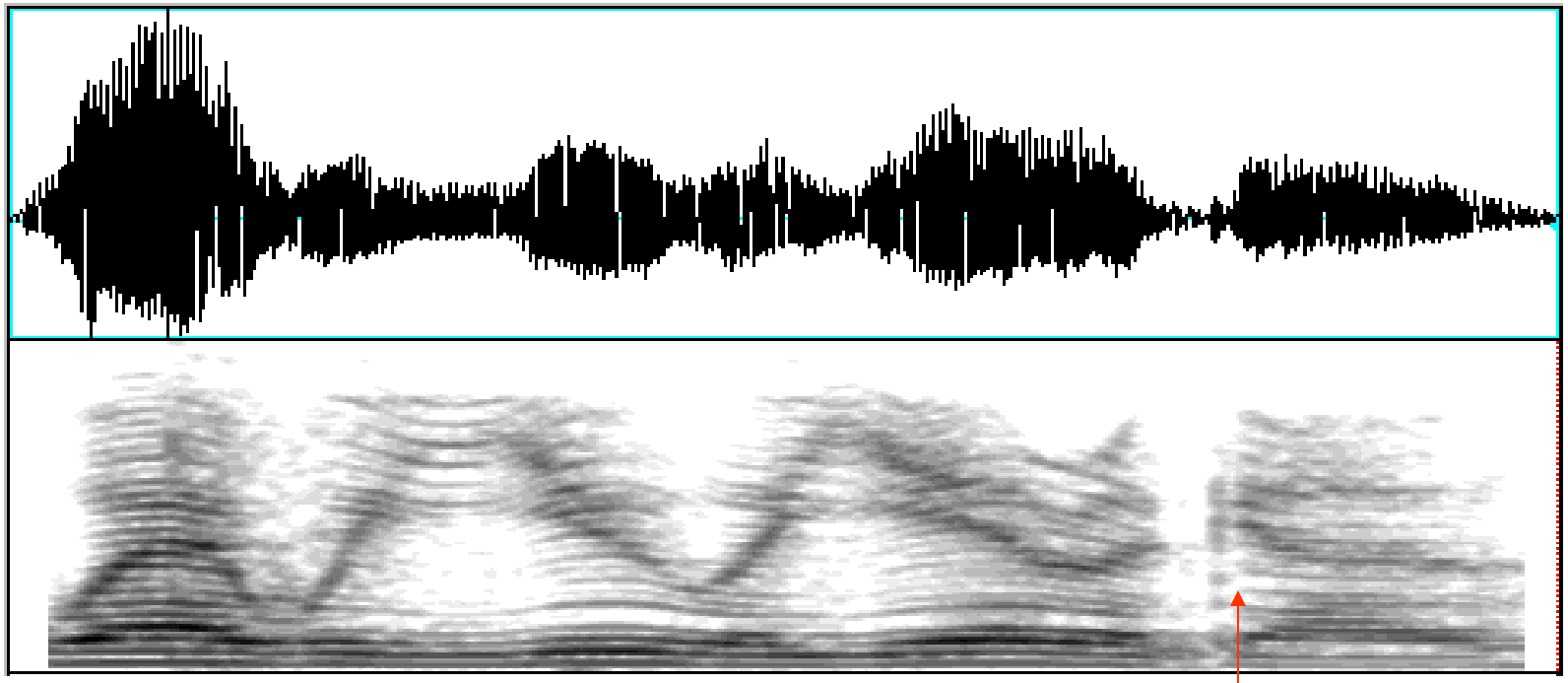
- Mynah produces "formants" but probably through changing syrinx resonances, not through changing vocal tract shape.

  (Klatt & Stefanski, 1974, J Acoust Soc Amer)

- Grey parrot has a longer vocal tract and may use changes in its shape to produce formant variation, more like human speech.

  (Warren, Patterson, Pepperburg, 1996, Auk)

Characteristics of speech



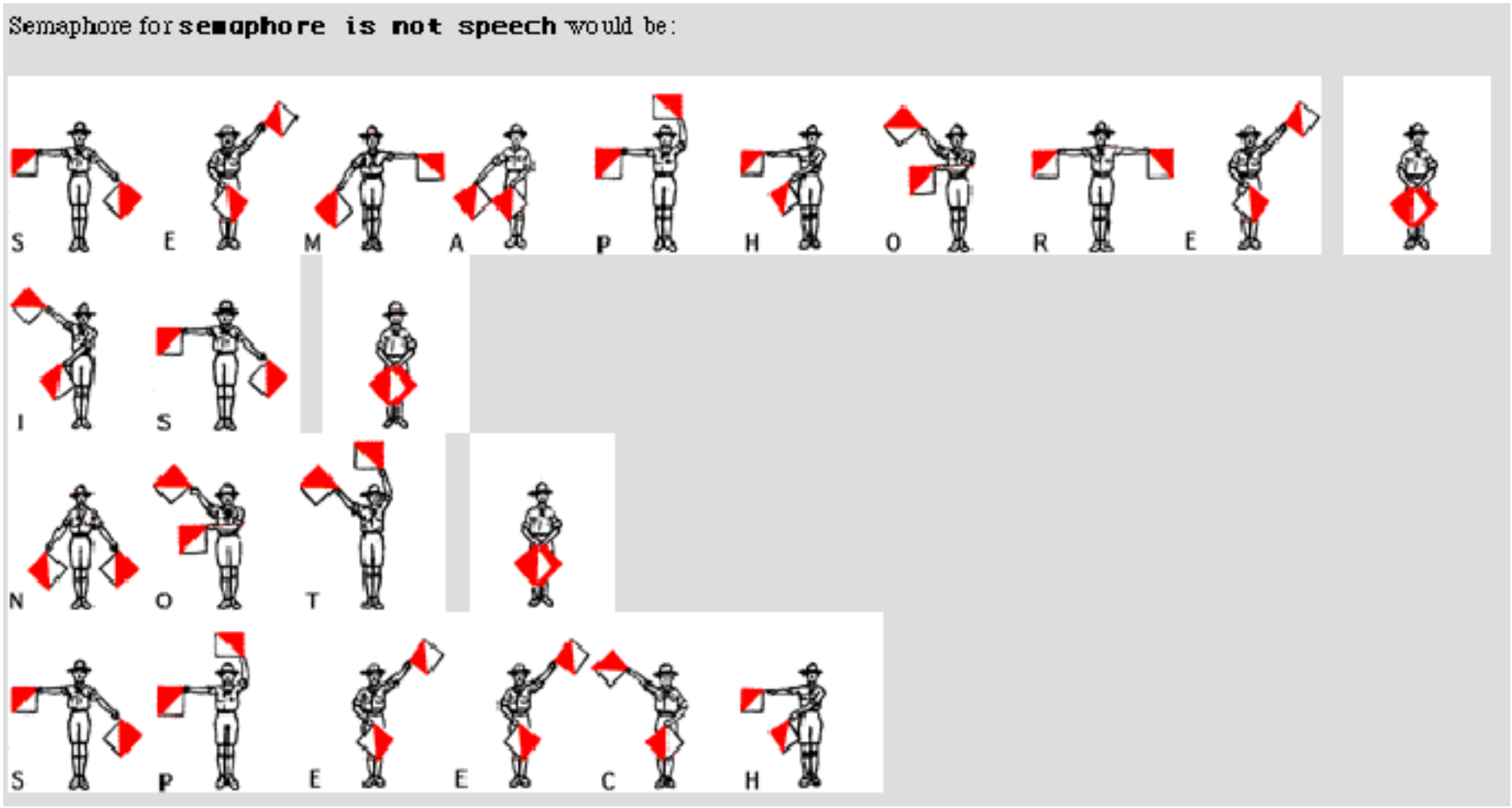
Only silence is /g/ of “ago”

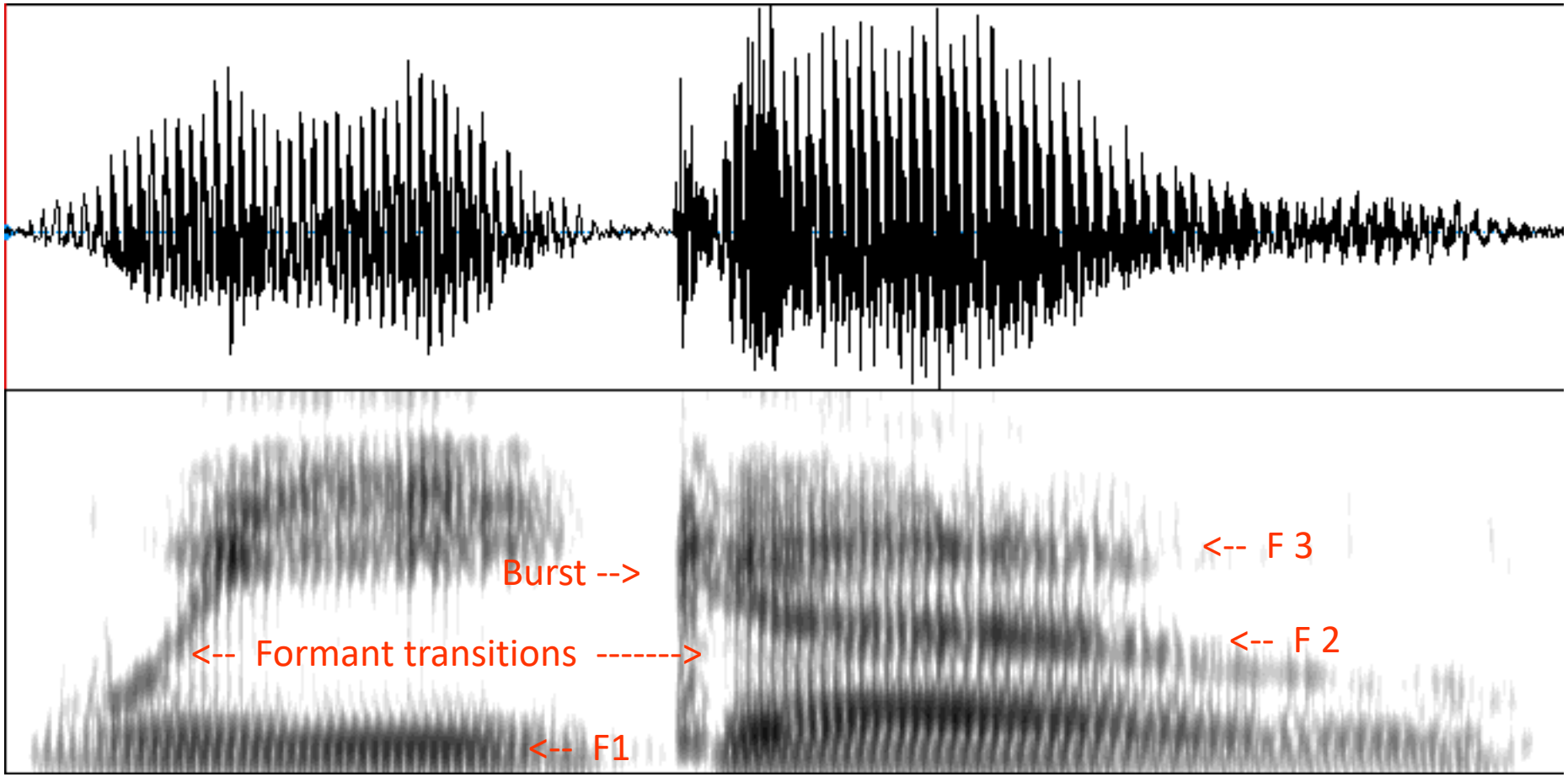
- No gaps between words
- Smoothly changing sound from one speech sound to the next
- So you can’t just shuffle the acoustic “words”

Speech is more like semaphore than like music

- Music: discrete targets giving discrete acoustic events
- Semaphore: discrete targets with transitions between targets
- Speech: articulatory transitions between targets

Semaphore

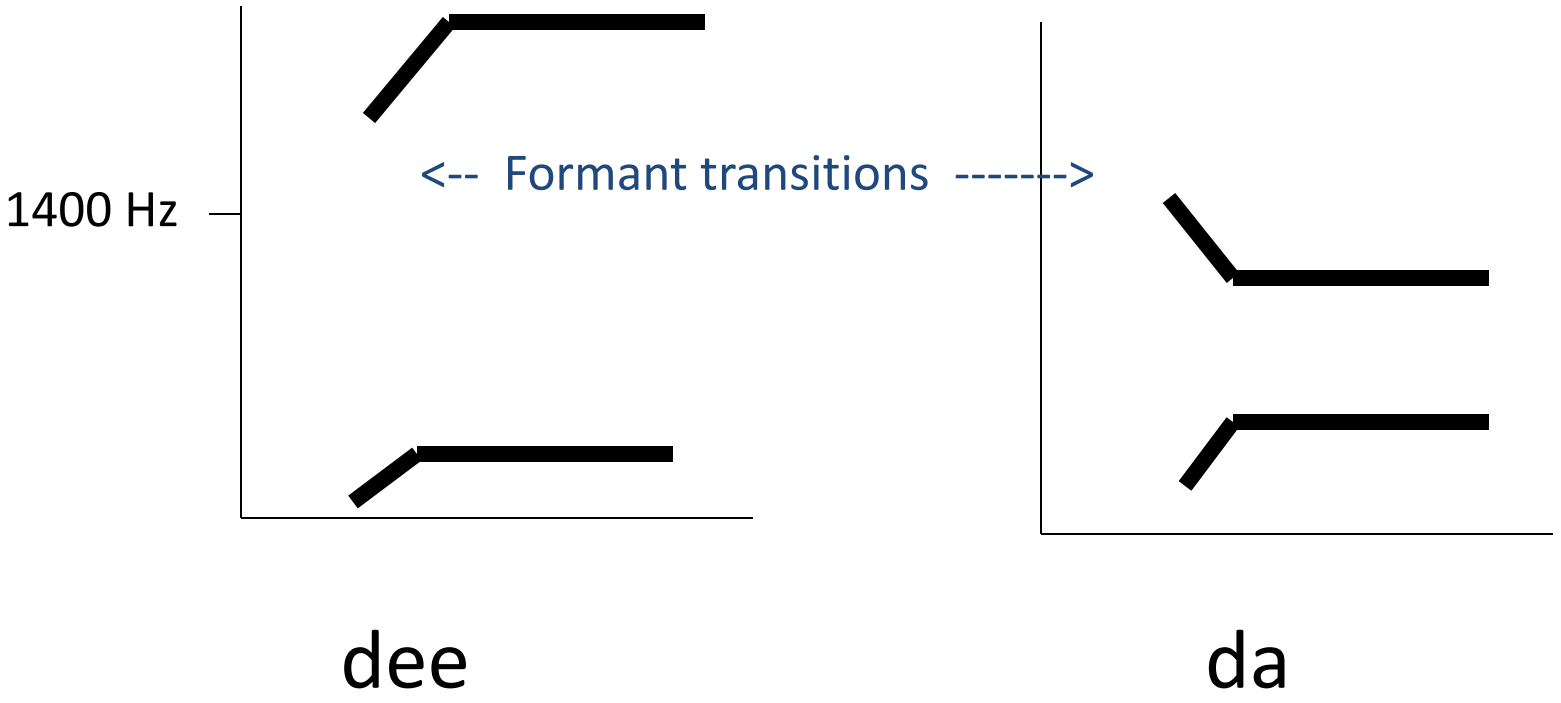

Semaphore for semaphore is not speech would be:
S E M A P H O R E
I S
N O T
S P E E C H



“w e g o”

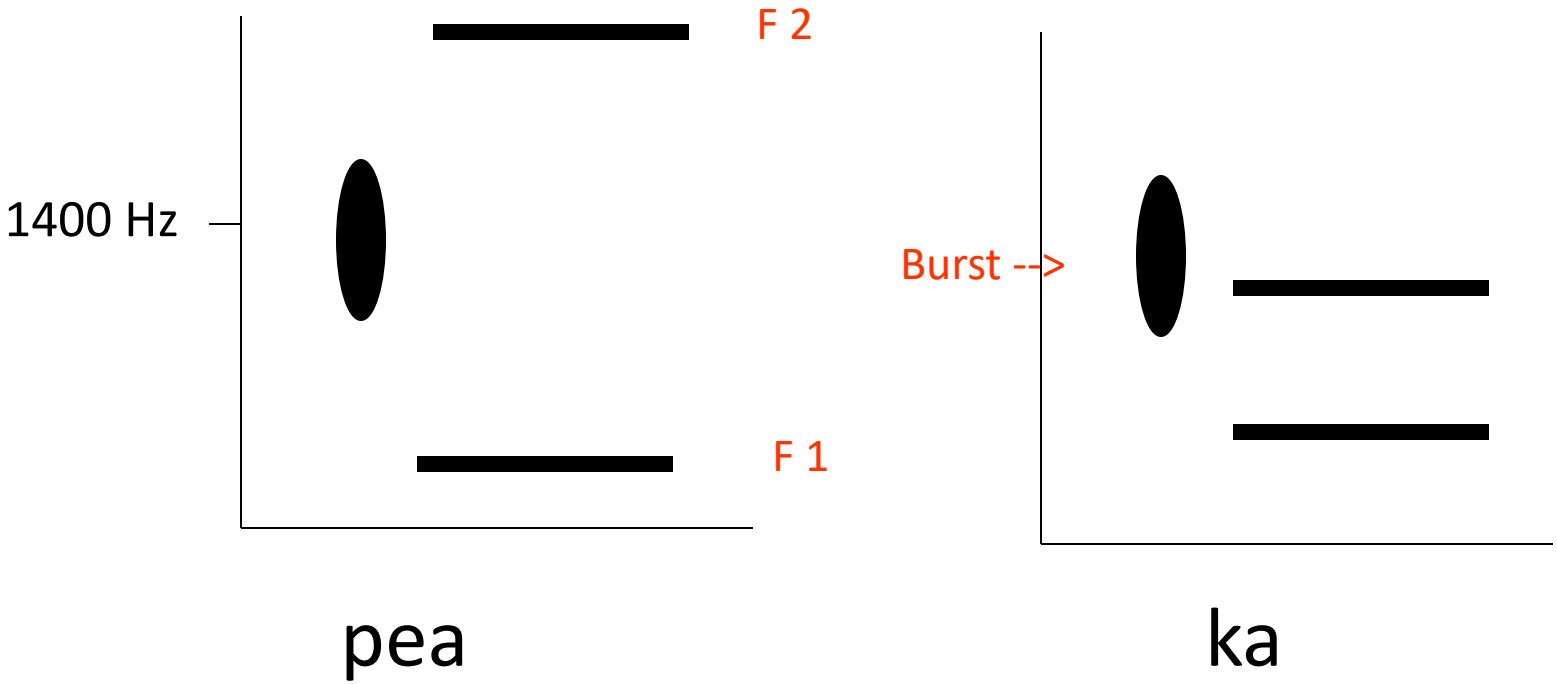
Formants in a wide-band spectrogram

Different transition - same consonant



Liberman et al. (1967) Perception of the speech code. Psych Rev 74, 431-461

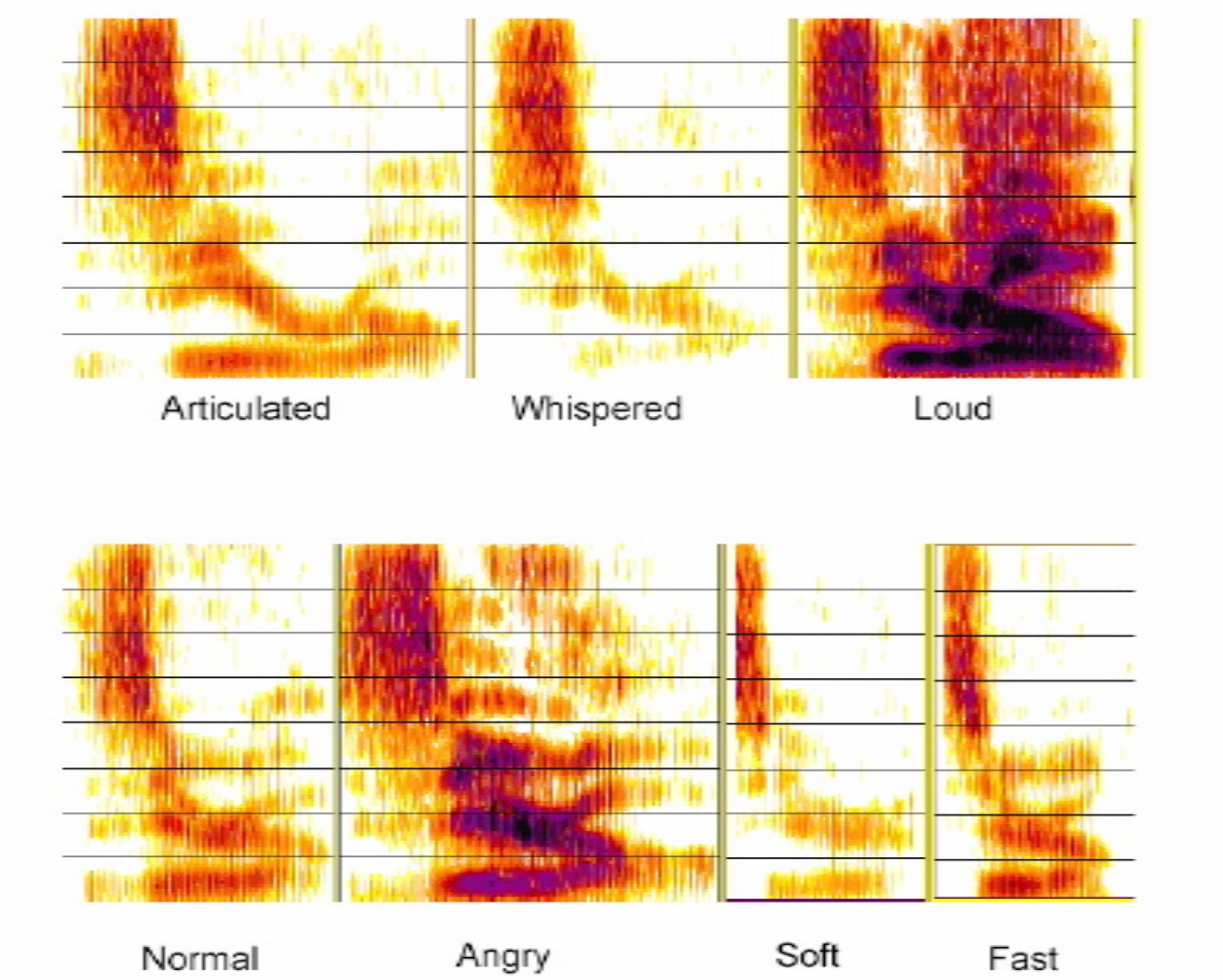
Same noise - different consonant



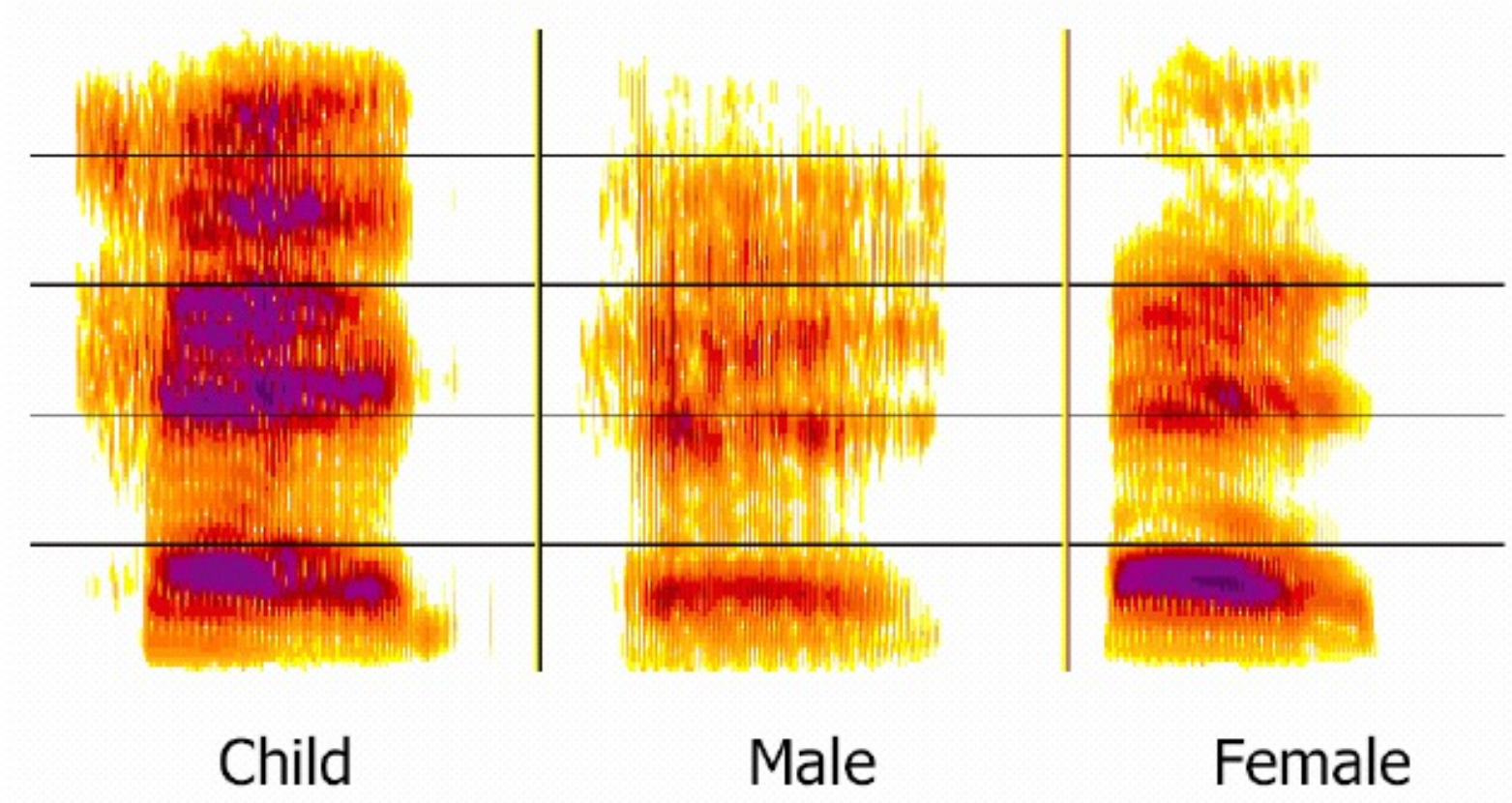
Liberman et al. (1967) Perception of the speech code. Psych Rev 74, 431-461

- The same speaker in different speaking mode

The word of “zero” spoken by same speaker



The word of “head” spoken by different speakers



Speech is more like speech than like semaphore

Speech does not have invariant acoustic targets: consonants change with the vowel.

Compare /d/ in dee

with /d/ in da

This is due to co-articulation.

Co-articulation

- Arises because (mainly) consonant gestures don't involve all the articulators:
  - eg /b/ is lips only, tongue free to take up position for next vowel
  - /d/ and /s/ just involve the tongue tip, touching the alveolar ridge, tongue body and lips free to take up position for next vowel - viz. /si/ /su/.

Two articulatory systems

Öhman suggested that articulation can be decomposed into *two semi-independent systems*:

- Slow movement from one vowel target to next, eg /i/ -> /u/
- Rapid consonantal movement superimposed eg /b/ /d/

So the /b/ in /ibu/ is not the same as in /iba/

Advantages of Co-articulation:

1. information about different segments is spread across time.
   - ✓ You know that a /u/ is coming because of the type of /s/ you have heard.
2. Liberman thought that this spreading across time makes it easier to transmit information at a fast rate.

Disadvantage of Co-articulation：

There are no constant acoustic targets in speech.

• The same phoneme can be represented as different sounds in different contexts

— eg. /s/ before /u/ or /i/

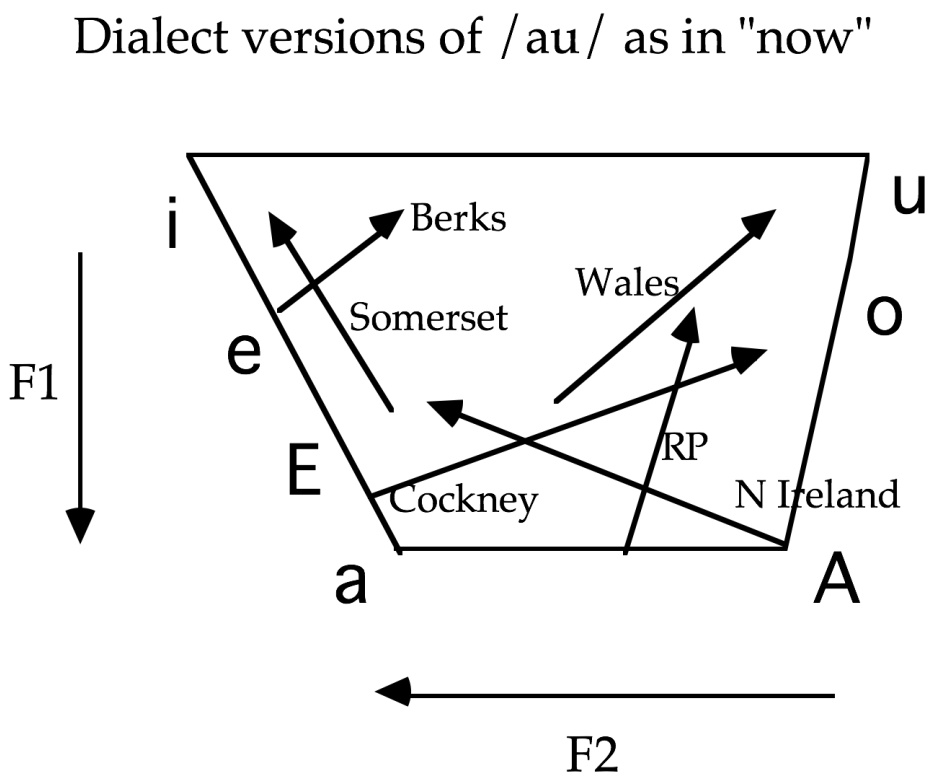
• Conversely, the same sound, can be heard as different consonants in different contexts

— eg. as /p/ before /i/ and /a/.

Speech Code

Factors that make it hard (for machines) to recognise speech

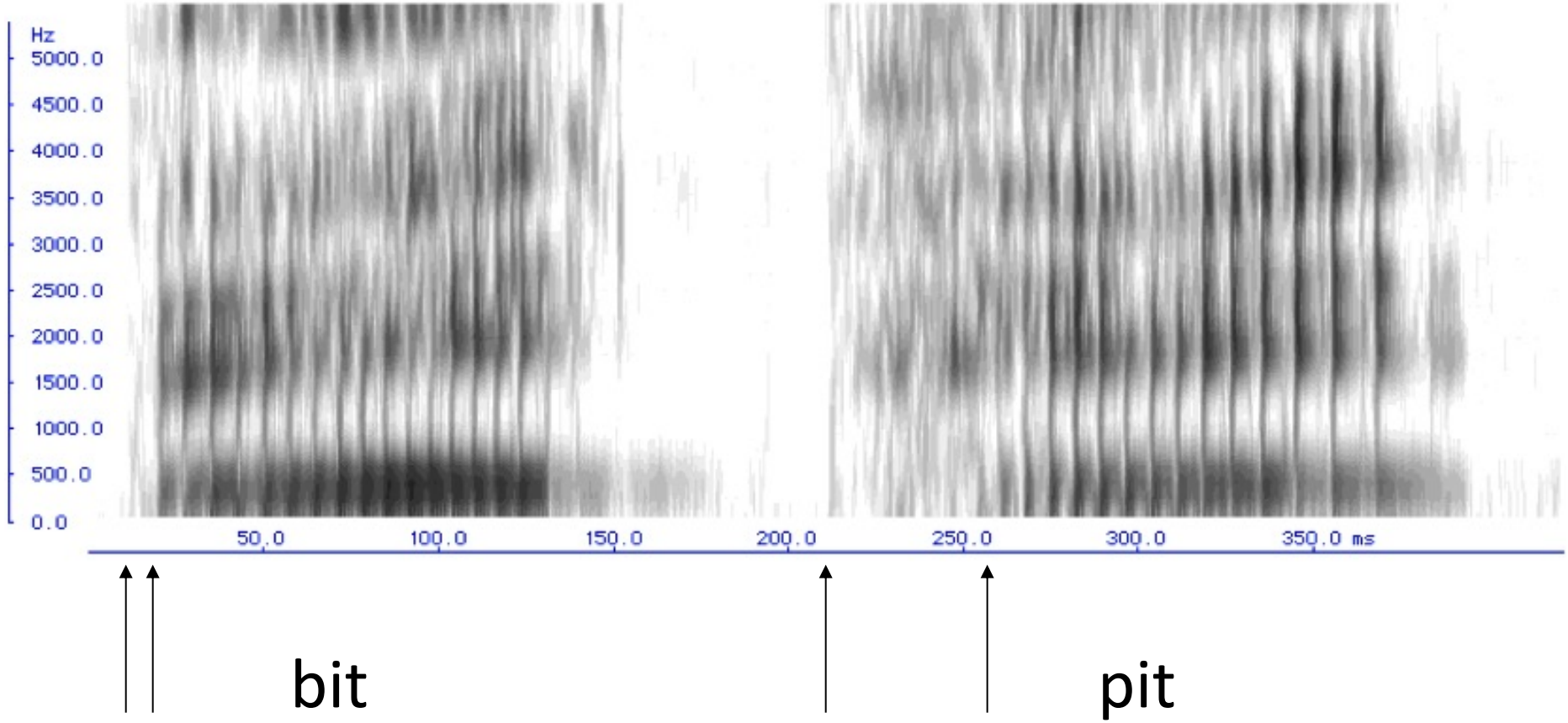
- Articulatory movement
- Co-articulation
- Rapid speech /djewonegaat/
- Different vocal-tract sizes:
  - men 15% longer than women
- Speaker variation
- Different dialects



So how do we do it?

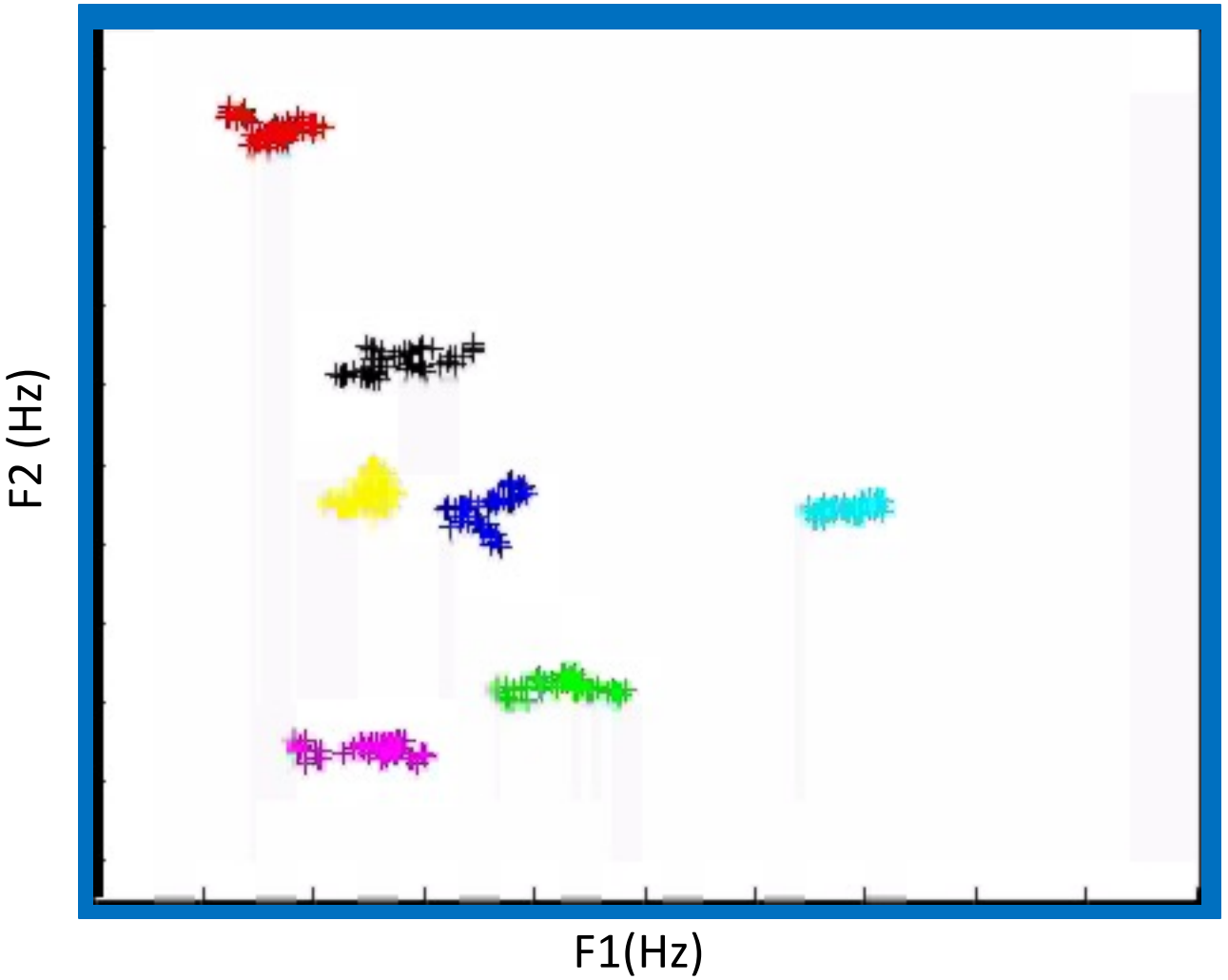
- Categorical Perception:
  - we impose categories on **physically continuous** stimuli

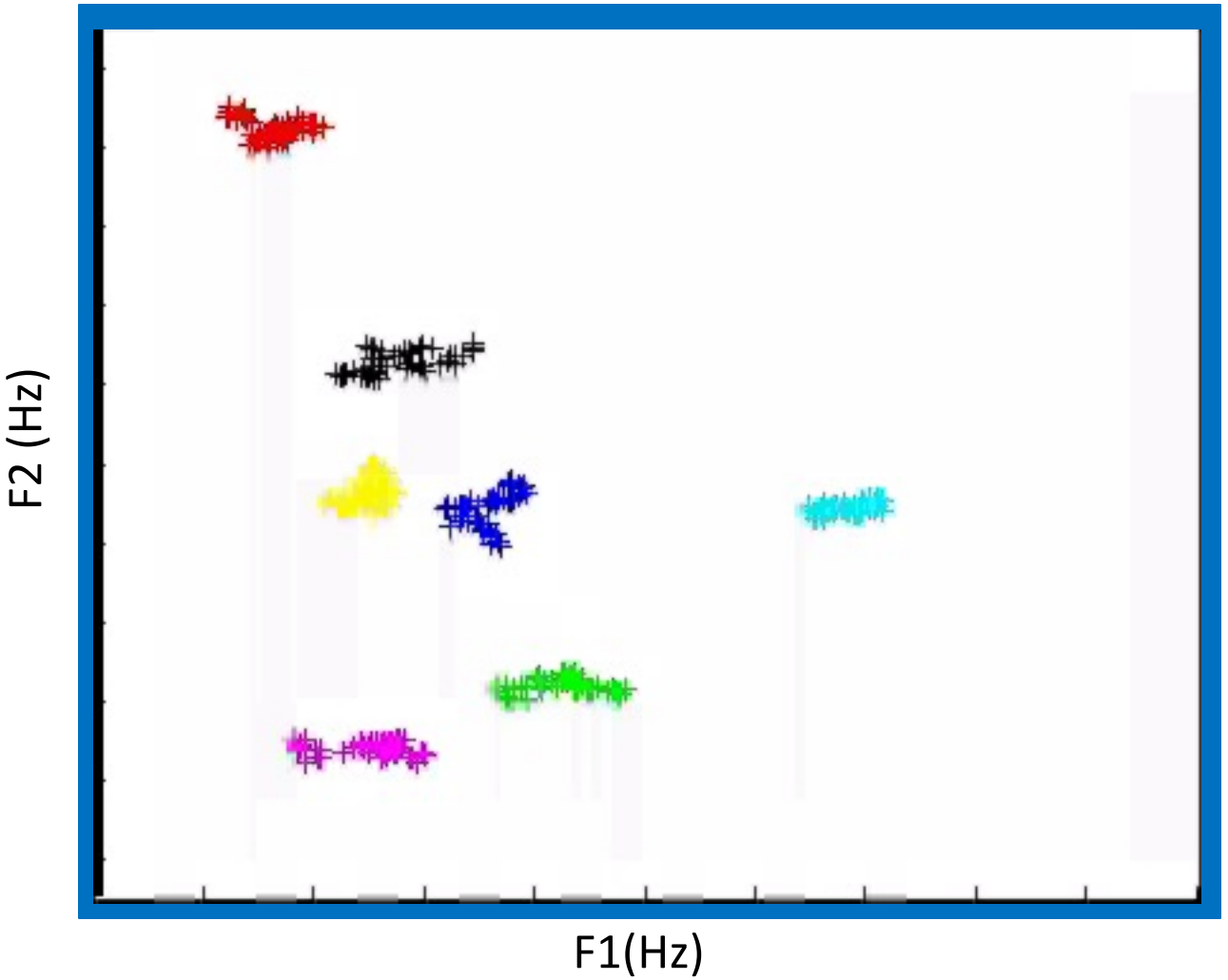
Voice-Onset Time (VOT)



- Voicing: differences in Voice Onset Time (VOT)
  - Small VOT: voiced
  - Large VOT: unvoiced

Vowel Space

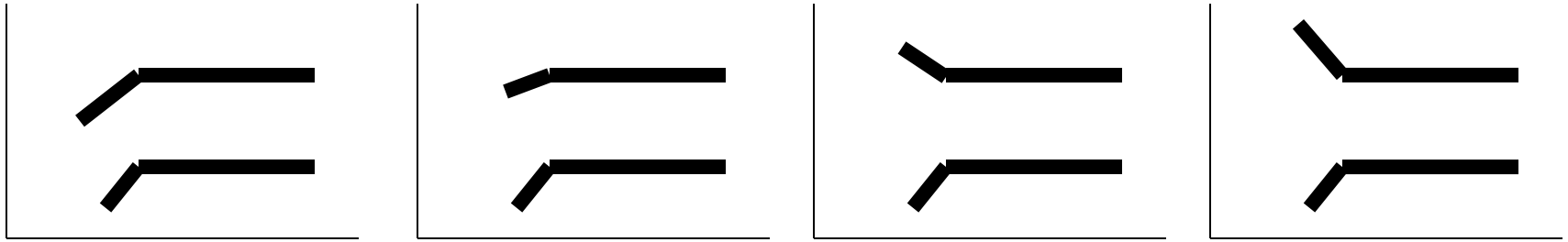



F2 (Hz)
F1(Hz)

Categorical Perception

1. Set up a continuum of sounds between two categories

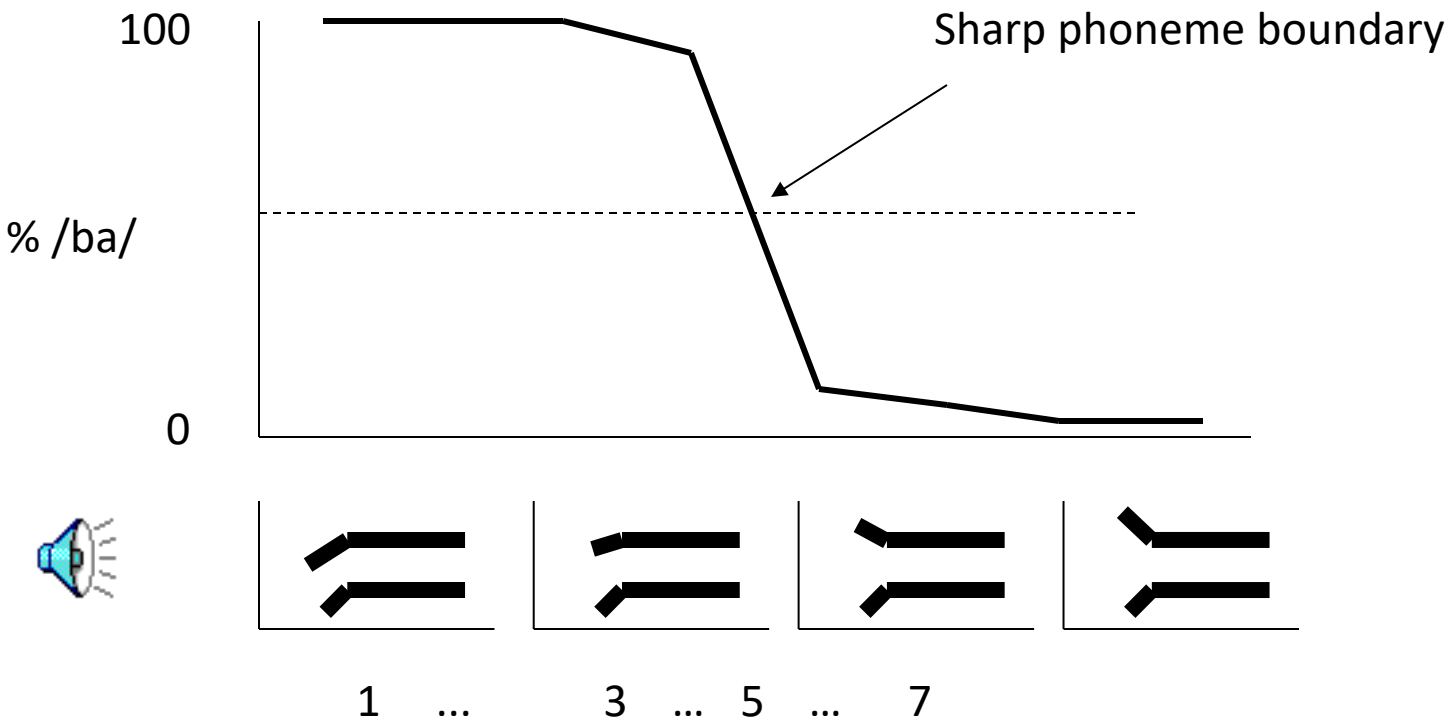
/ba/ - /da/



1 … 3 … 5 … 7

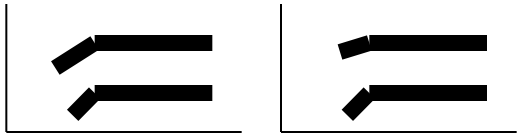
Categorical Perception

2. Run an identification experiment

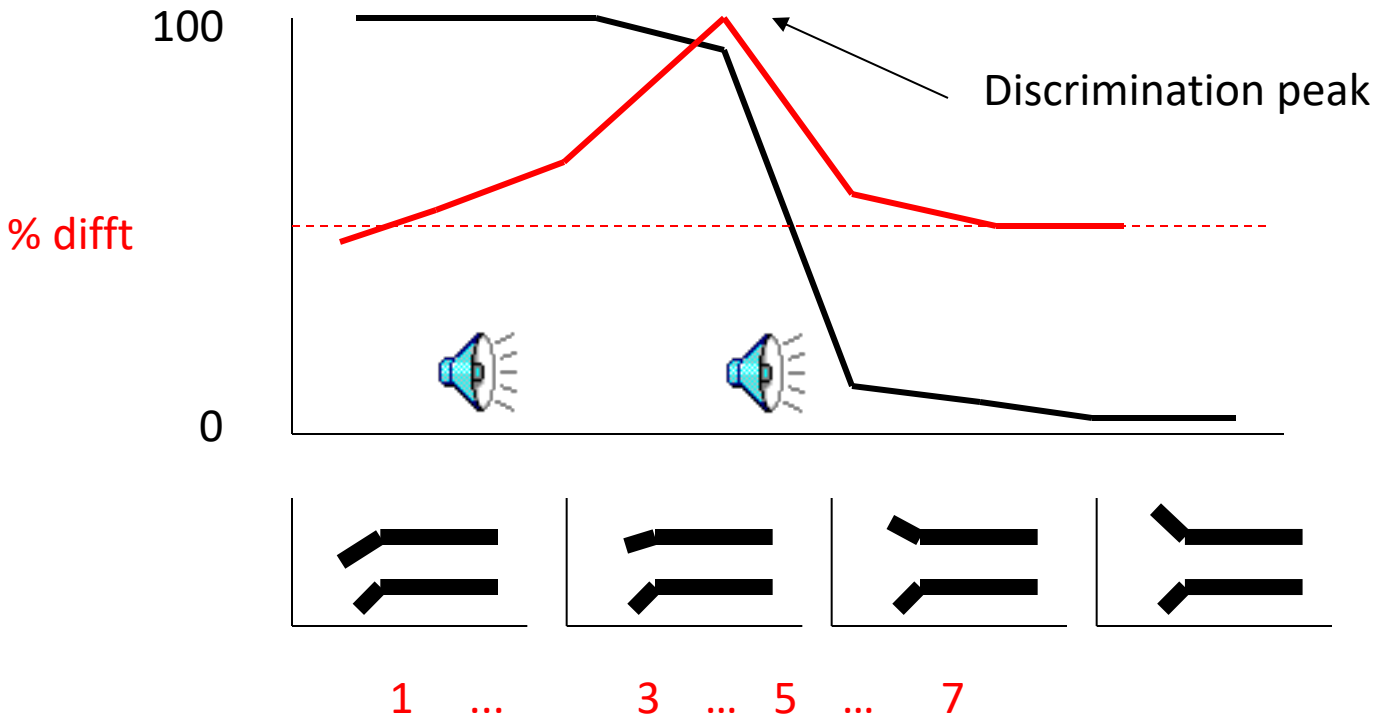


Categorical Perception

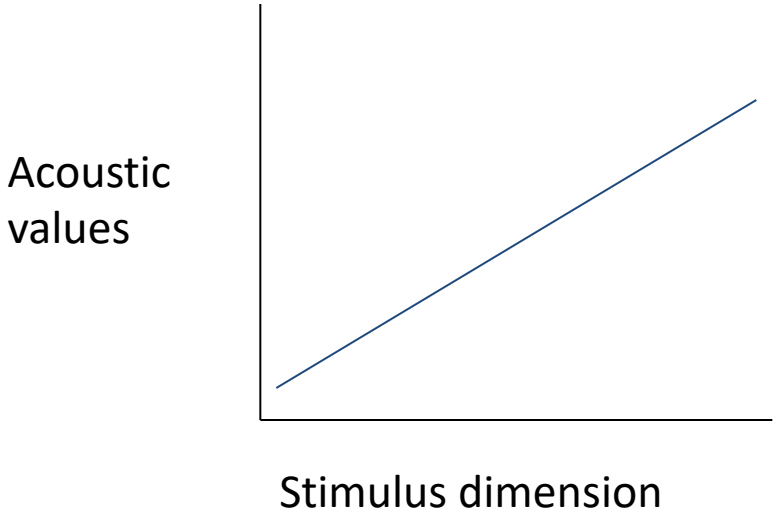
3. Run a discrimination experiment

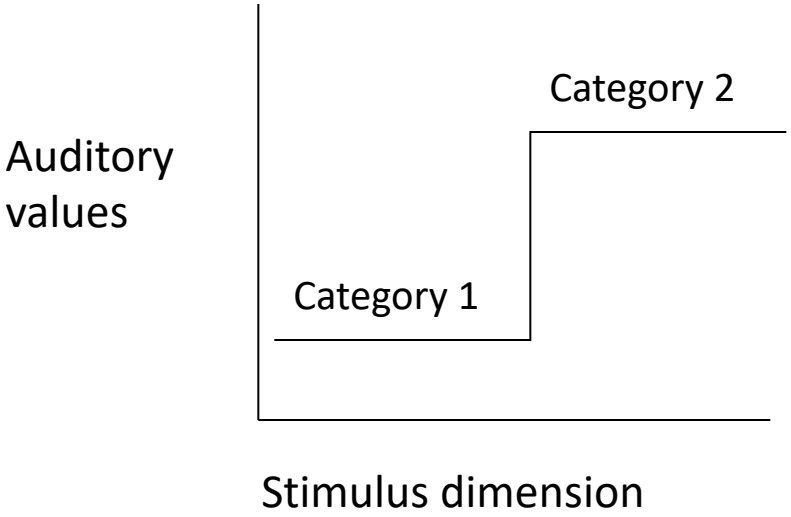


1 versus 3



Natural auditory categories

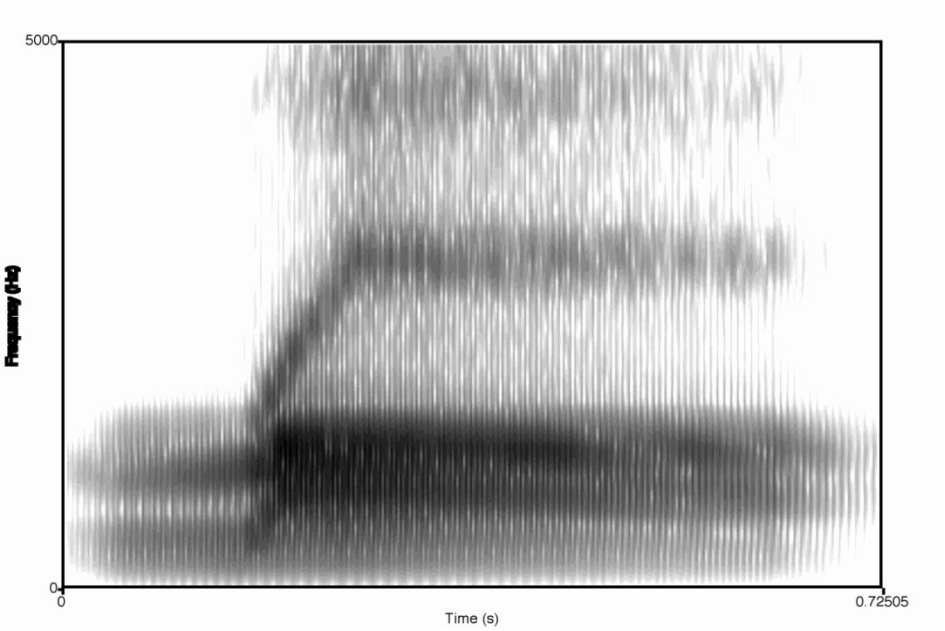




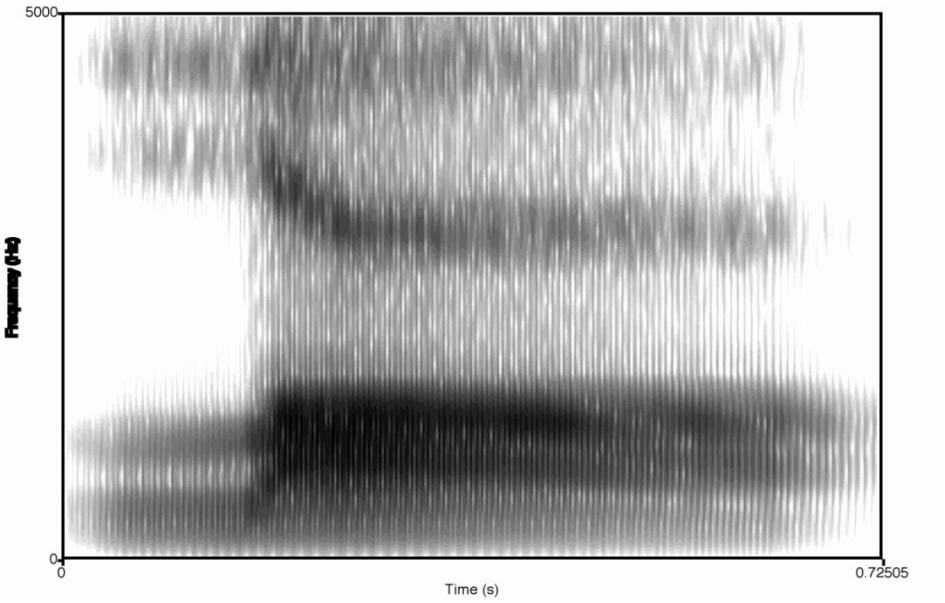
Auditory system show these natural categories.

Sinex, D. G. and McDonald, L. P. (1989). "Synchronized discharge rate representation of voice-onset time in the chinchilla auditory nerve," J. Acoust. Soc. Am. 85, 1995-2004.
Sinex, D. G., McDonald, L. P. and Mott, J. B. (1991). "Neural correlates of nonmonotonic temporal acuity for voice onset time," J. Acoust. Soc. Am. 90, 2441-9.

Synthetic Stimuli: /ra/-/la/

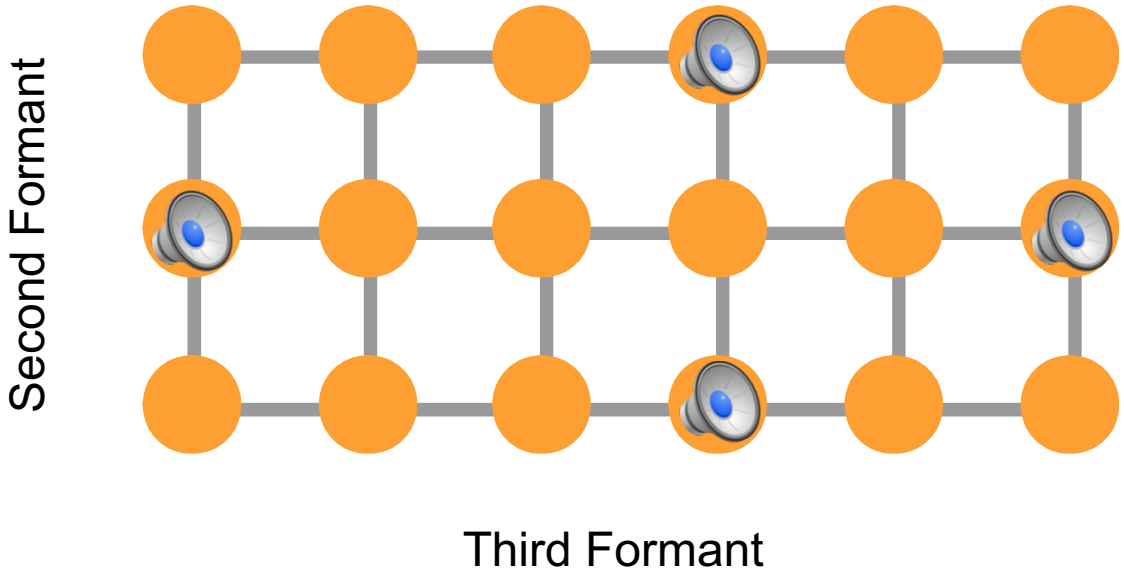


/ra/



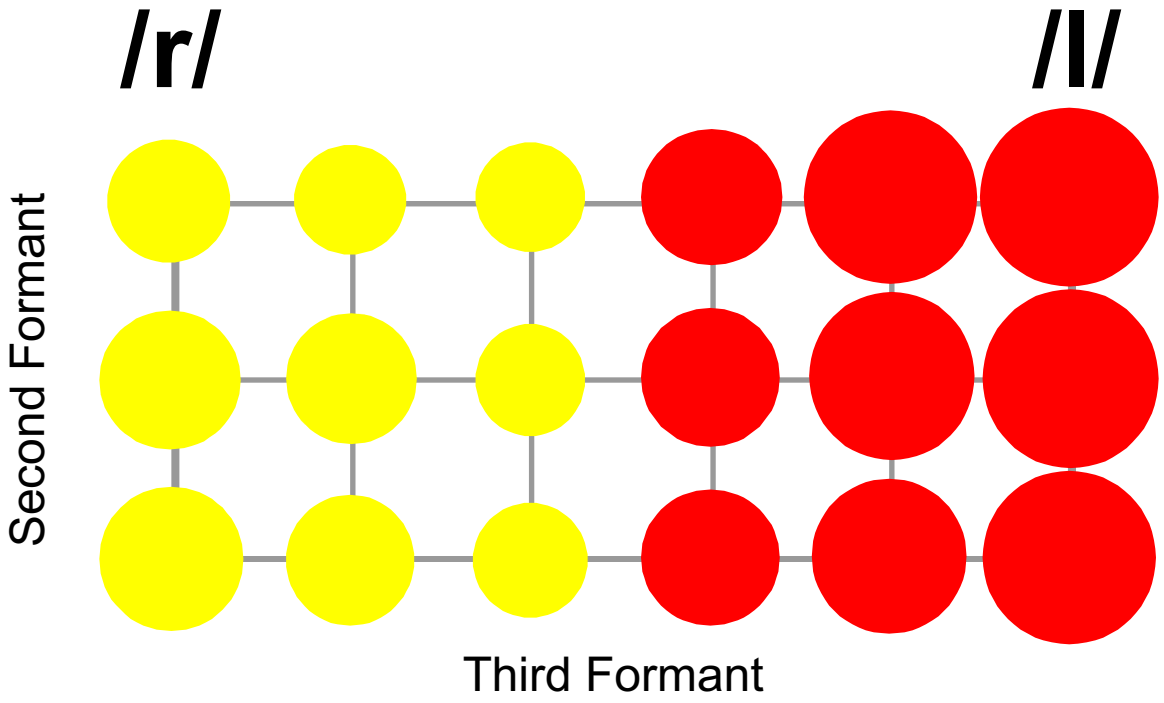
/la/

Synthetic Stimuli: /ra/-/la/



Iverson, P., et al. (2003). "A perceptual interference account of acquisition difficulties for non-native phonemes," Cognition 87, B47-57.

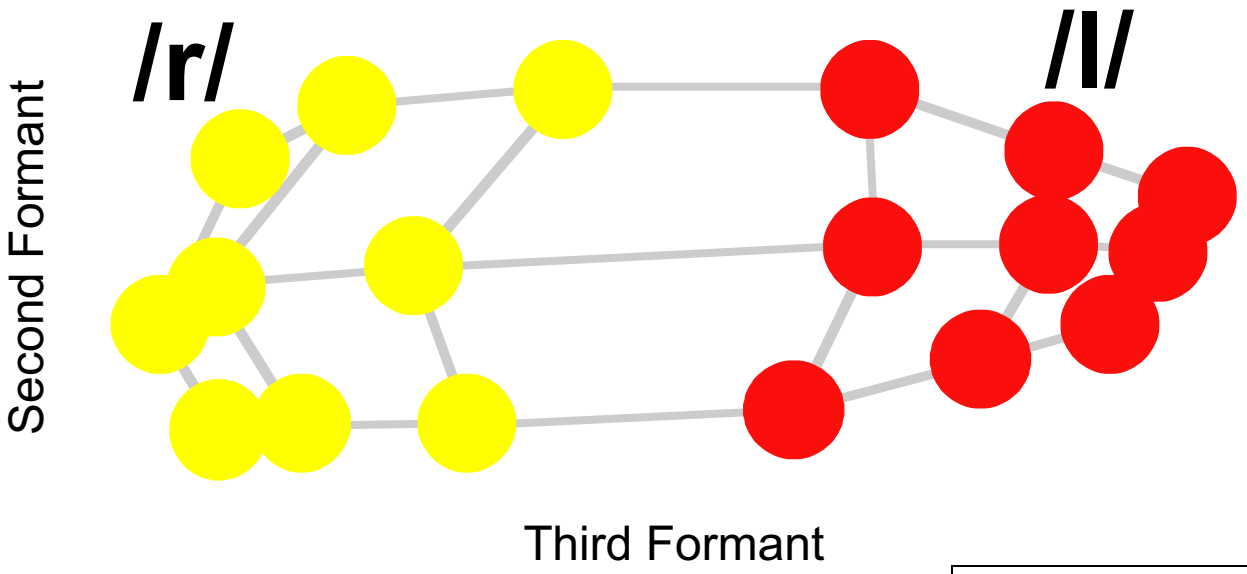
American Goodness and Identification Results



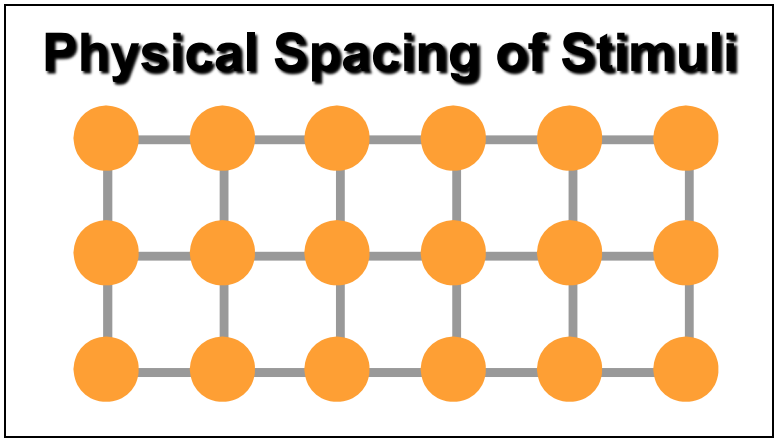
listeners identified and rated the goodness of individual stimuli...

American MDS Solution

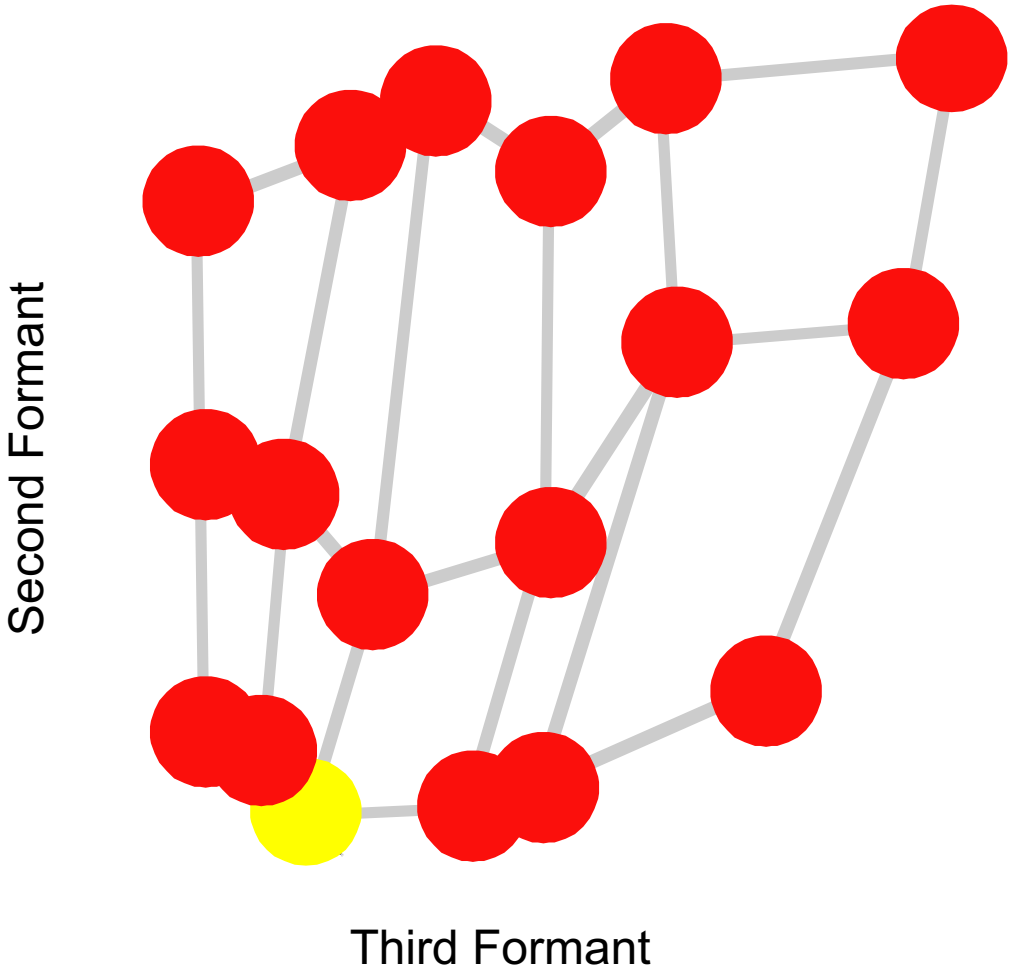
MDS：Multiple dimension scaling



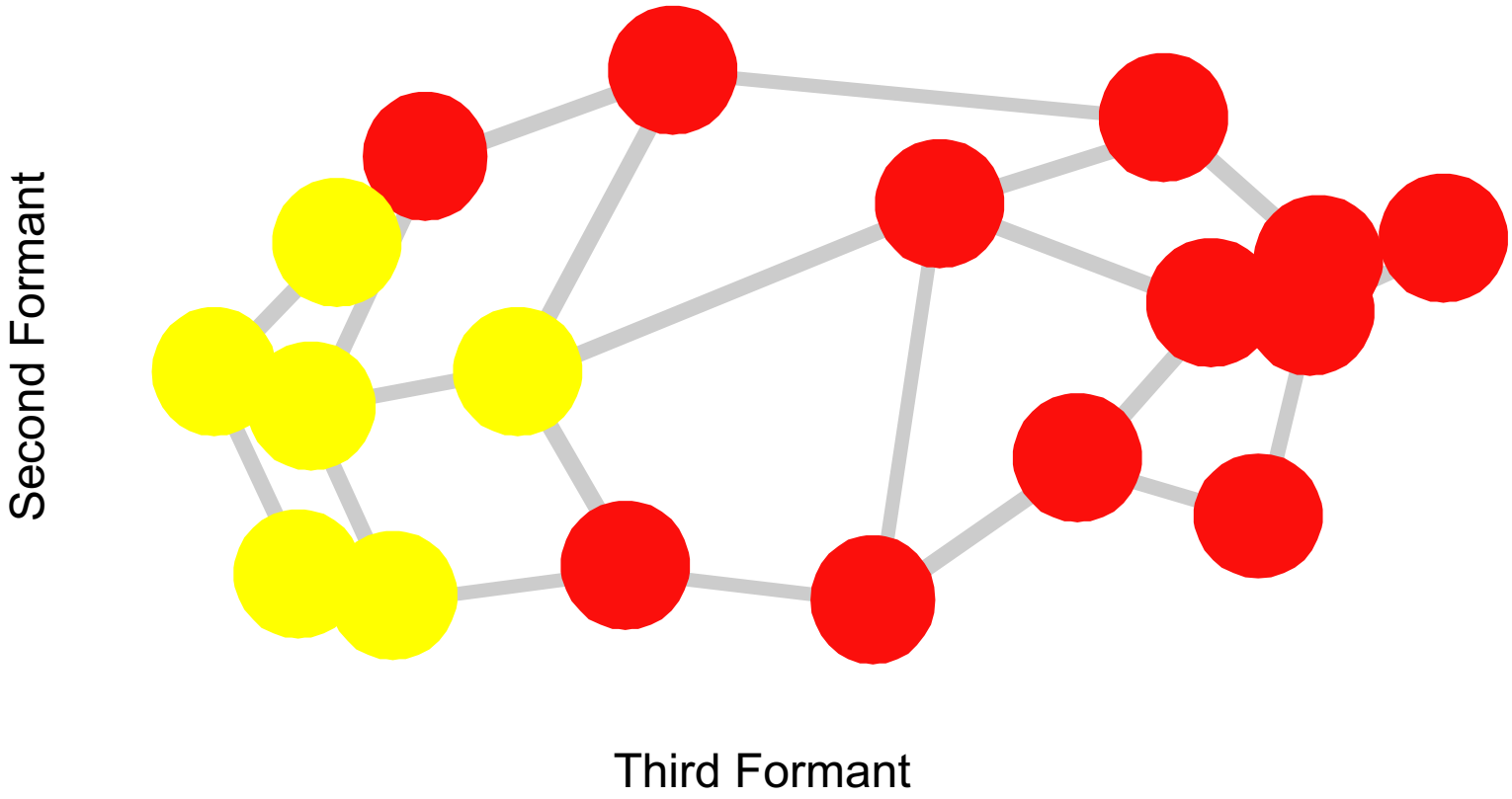
...and rated the similarity of stimulus pairs



Japanese MDS Solution

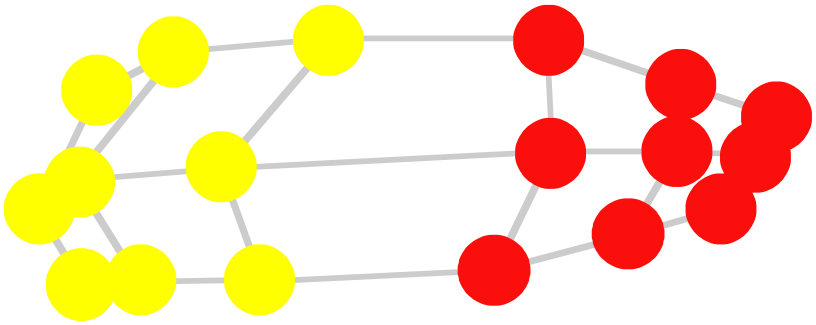


German MDS Solution

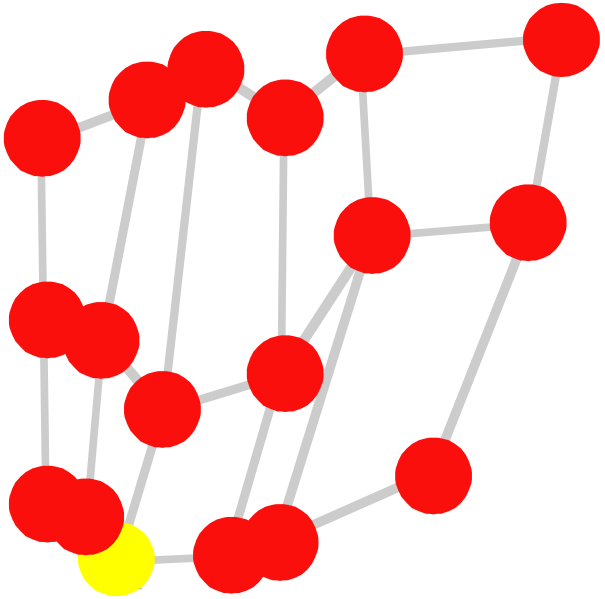


All MDS Solutions

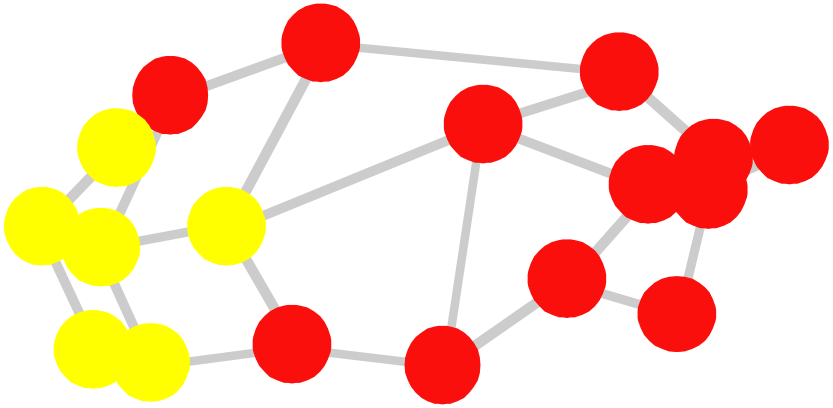
American



Japanese

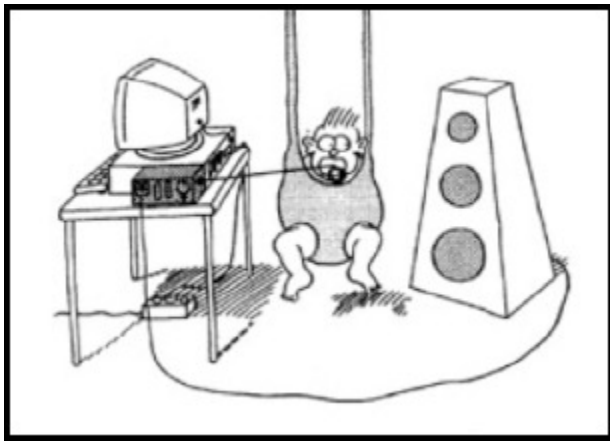


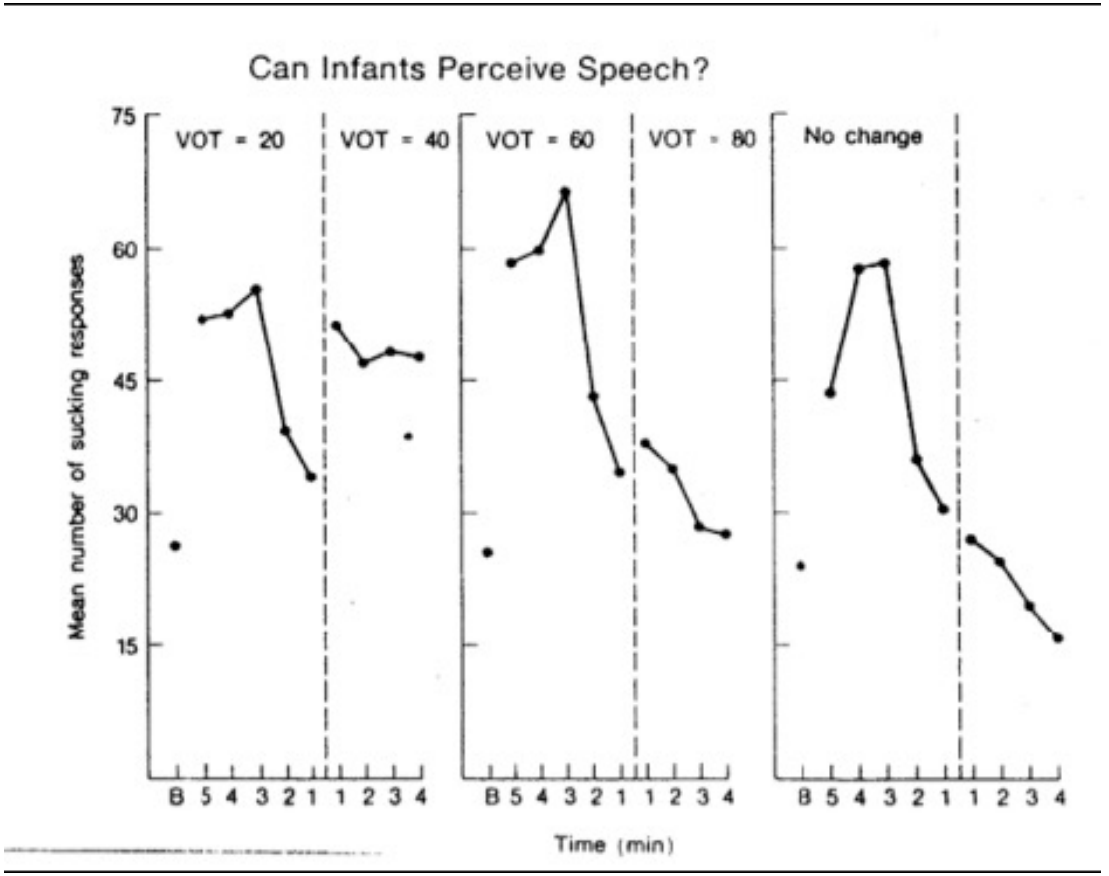
German



Is categorical perception innate?
(i.e., are category boundaries learned or "built-in"?)

- Infants do it (Eimas et al., 1971)

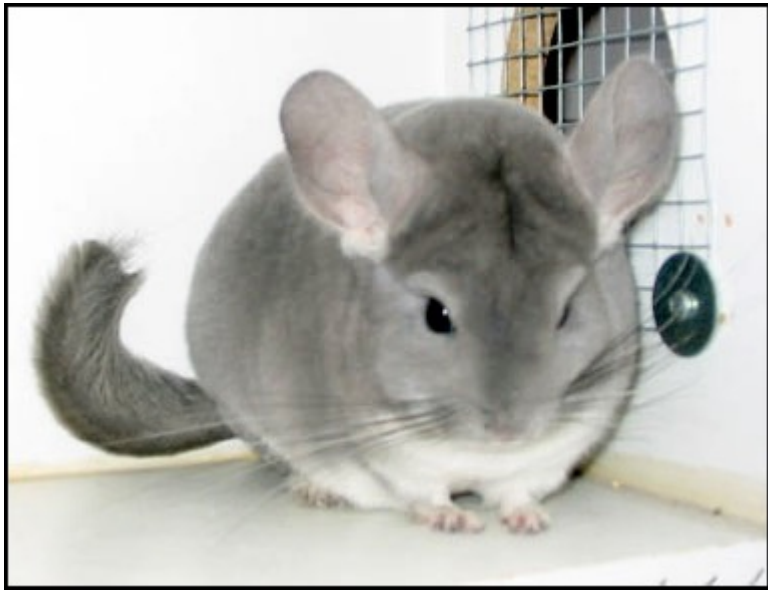


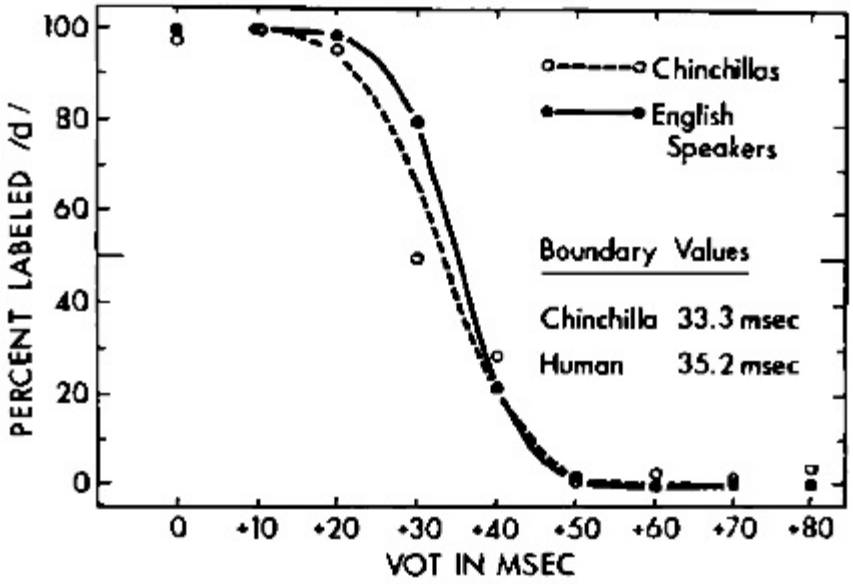


Is categorical perception innate?

(i.e., are category boundaries learned or "built-in"?)

- Infants do it (Eimas et al., 1971)
- but Chinchillas do too. (Kuhl & Miller, 1978)





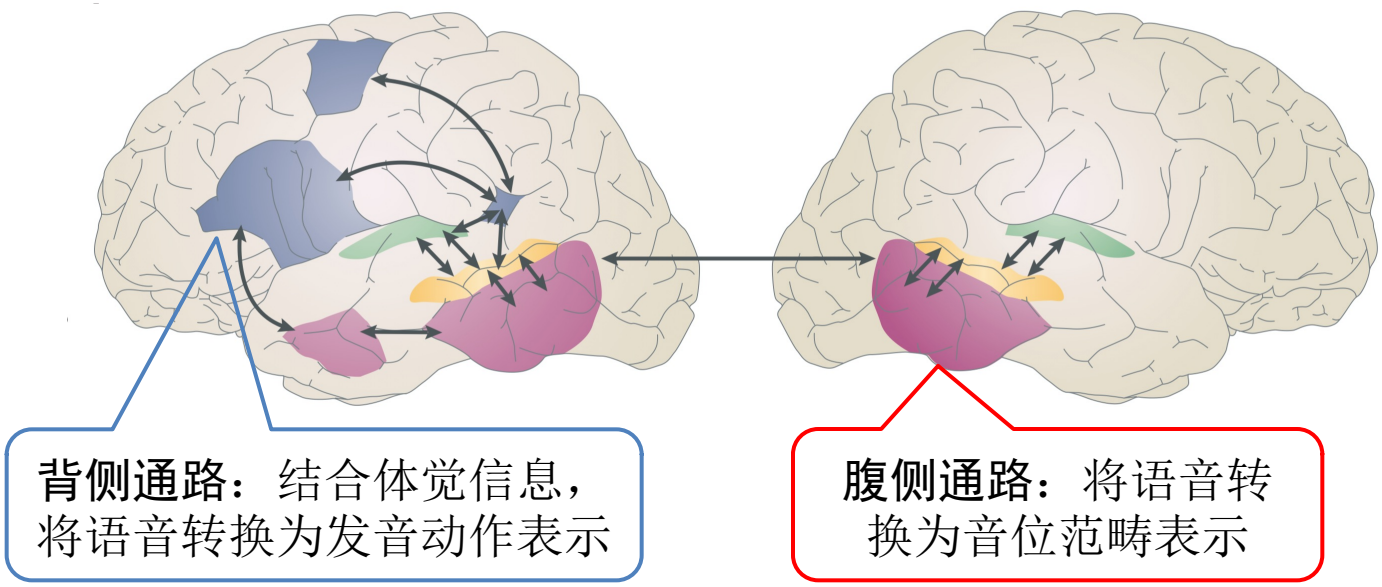
		Place of articulation		
		Labial	Alveolar	Velar
Stop	Voiced	b	d	g
Stop	Voiceless	p	t	k
Nasal	Voiced	m	n	ng

Place-of-articulation dimensions show these natural categories.

- Infants born with ability to make many discriminations — Place-of-articulation
- During second 6-months they lose the ability to make distinctions not used in their language — Language Culture

Dual-Stream Model of Auditory System

Hickok & Poeppel, 2007



人无法观察到舌、软腭等发音器官的运动状态，实质上实现的是无监督语音反演

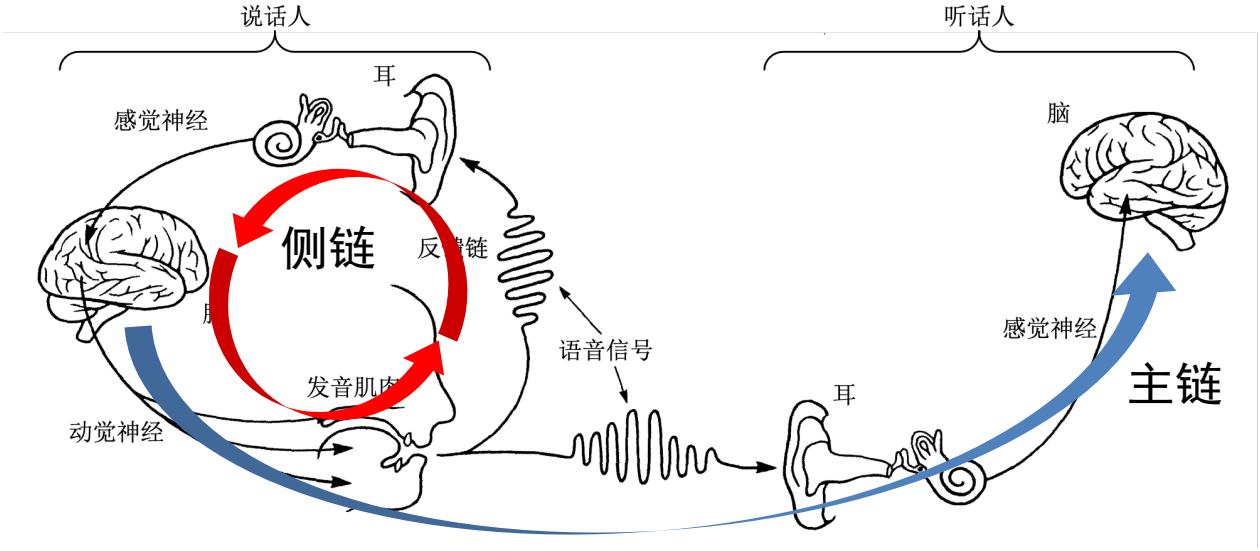
Embodied Cognition and Motor Theory

- 具身认知
  - 认知涌现于脑、身体和环境之间的持续交互作用
  - 认知的发育过程受到身体物理属性的约束
  - 身体提供了认知的最原始概念
- 言语感知的肌动理论(Motor Theory)
  - 感知言语就是感知讲话者的目标发音姿态(intended gestures)
  - 言语感知和言语生成密切关联
  - 无法有效控制发音器官的病人无法获得言语感知能力 (MacNeilage et al., 1967)

人如何实现言语感知和言语生成的交互？

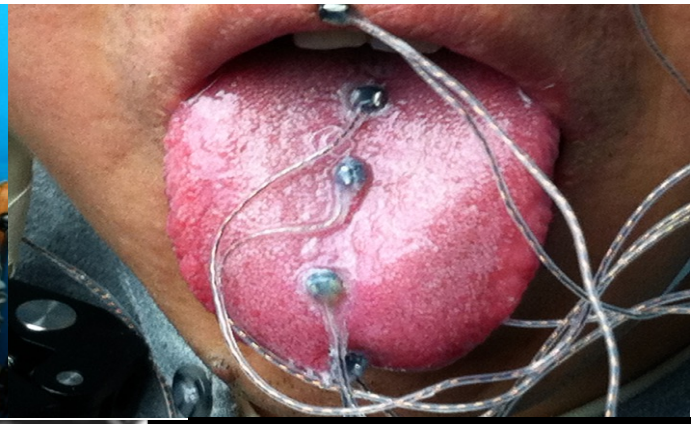
Speech Chain

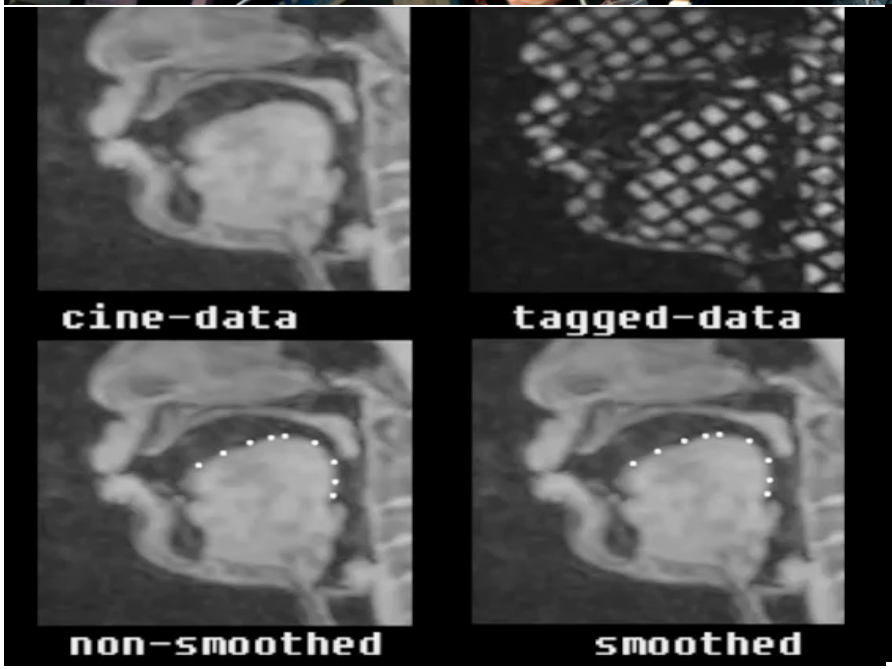
- 言语感知和言语生成的交互
- 延迟听觉反馈(Delayed Auditory Feedback，DAF)(Yates, 1963)
  - 长期耳聋会导致言语退化

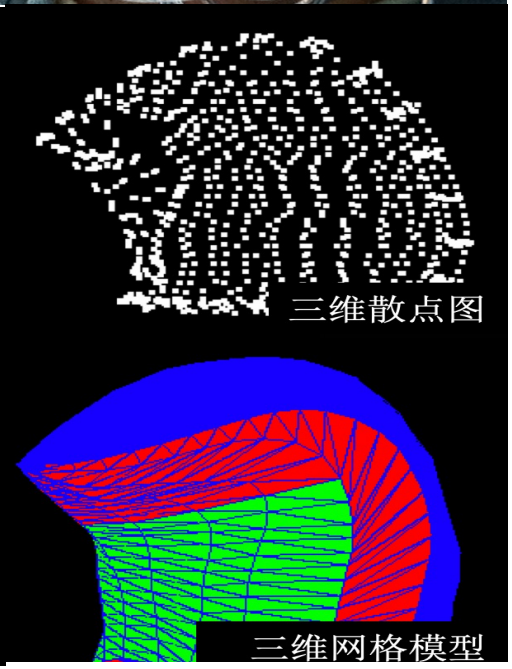


Denes et al., 1993

Recording of Speech Production

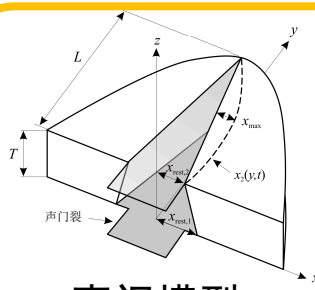




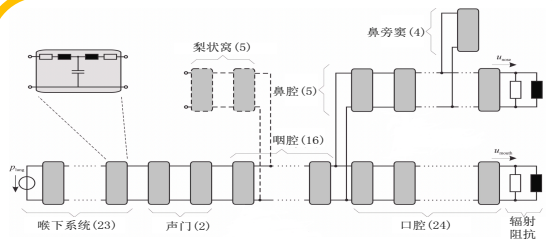


Models of Speech Production

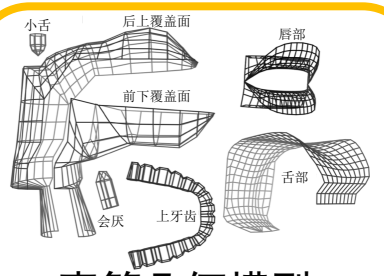
- VocalTractLab (VTL) models (Birkholz, 2013)



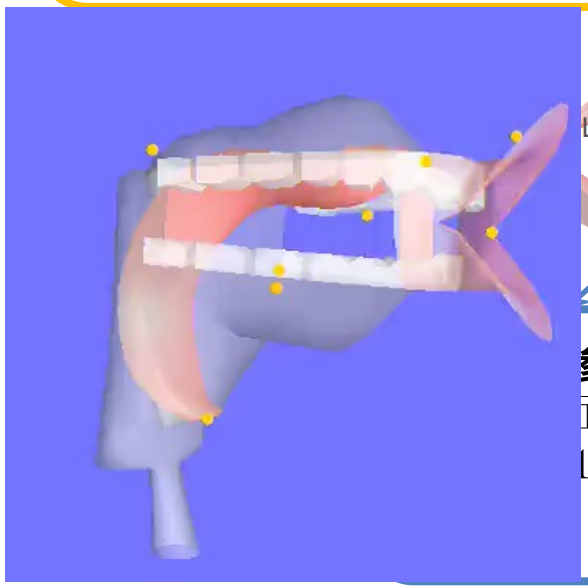
声门模型



传输线声学模型



声管几何模型

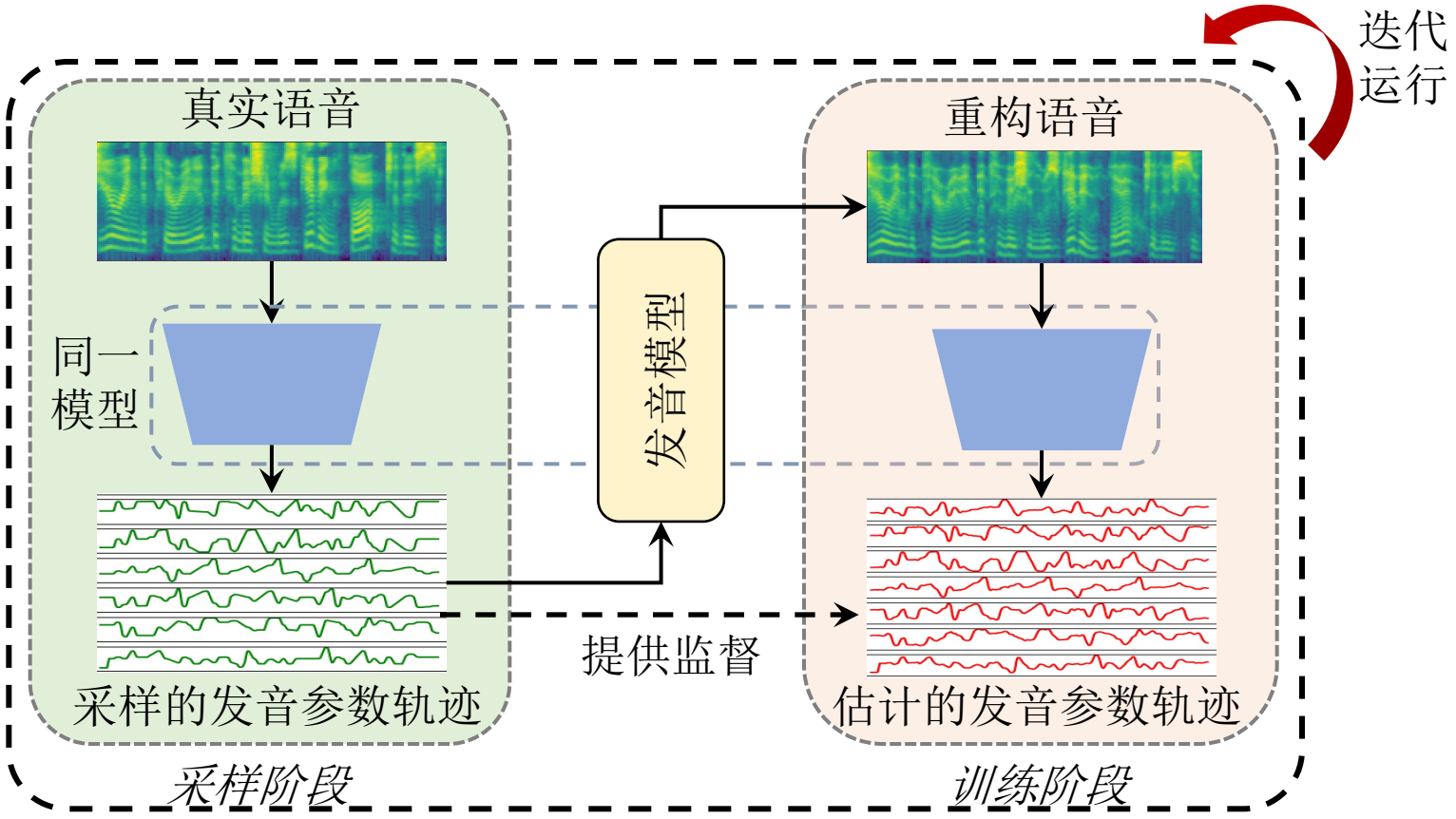


声门动作参数：基频、气压、送气强度等11维；**时变**

参数：面、舌19维；

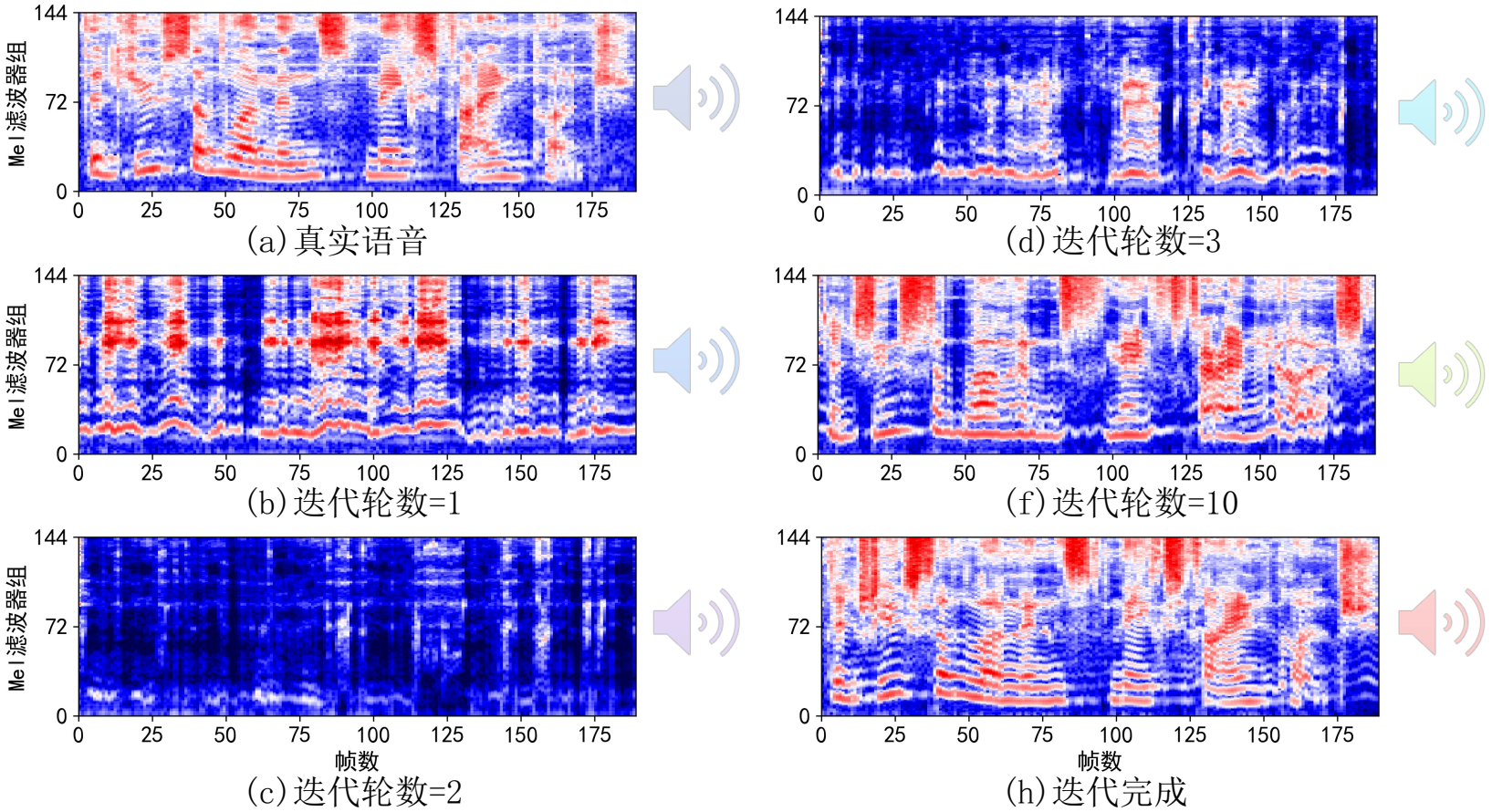
声道结构参数：唇宽、咽长、上下臼齿高度等13维；**非时变**

EmJEm: Embodied Joint Embedding



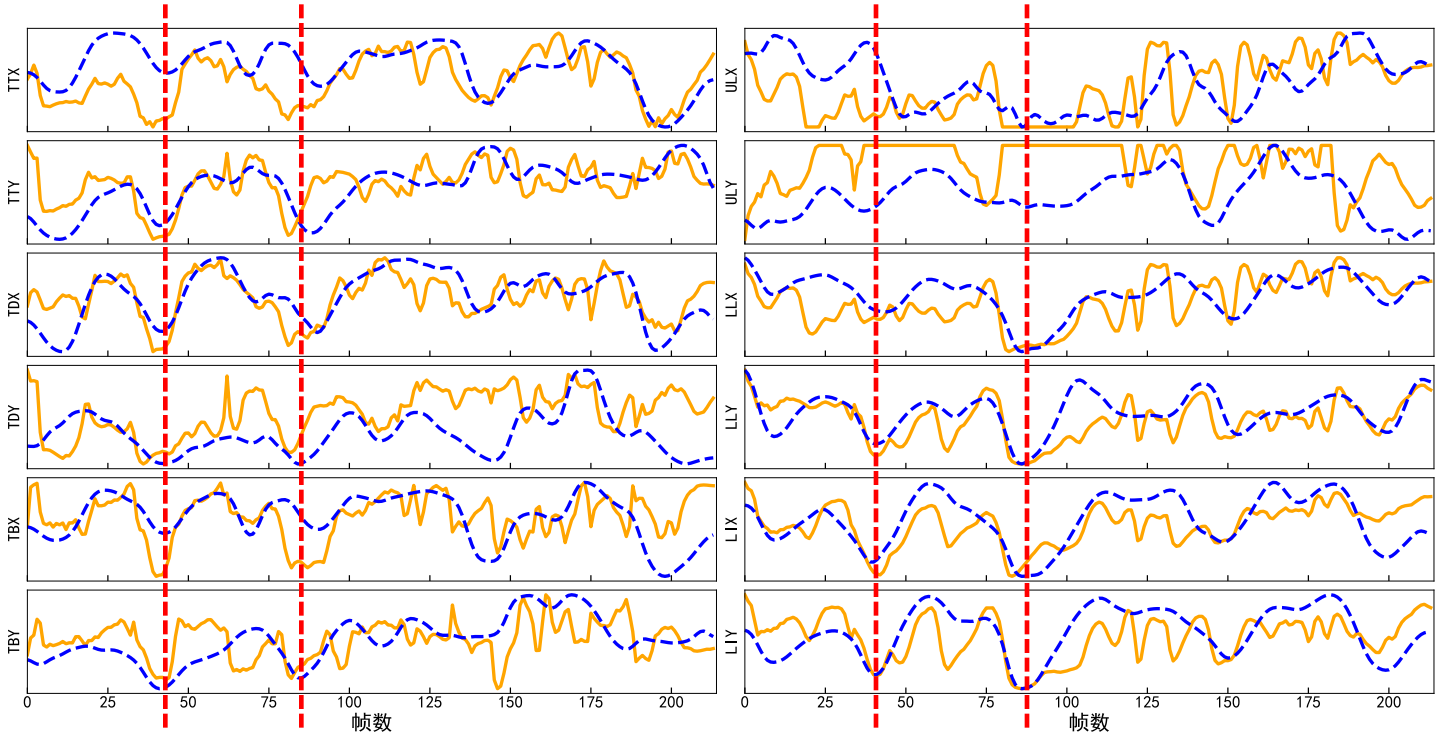
迭代式无监督语音反演方法

- 收敛过程



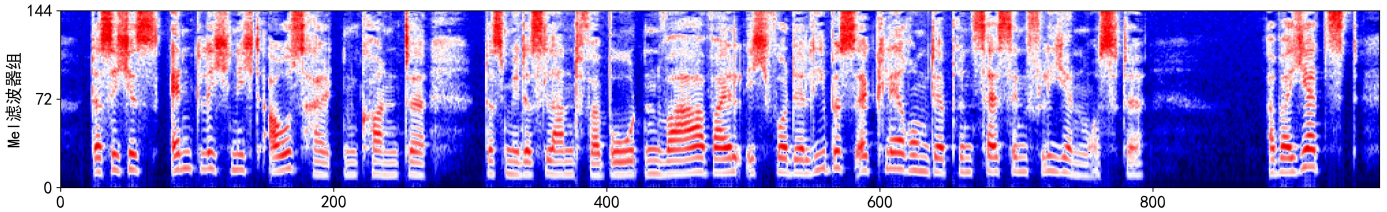
不同迭代轮次的重构语音梅尔语谱图

- 发音参数轨迹

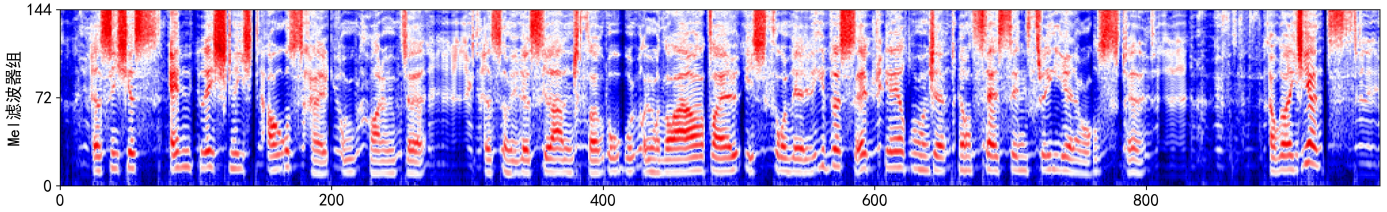


估计的发音参数轨迹(黄色实线)与实测发音轨迹（蓝色虚线）的样例，通过VTL将估计的声道动作参数转换到EMA维度

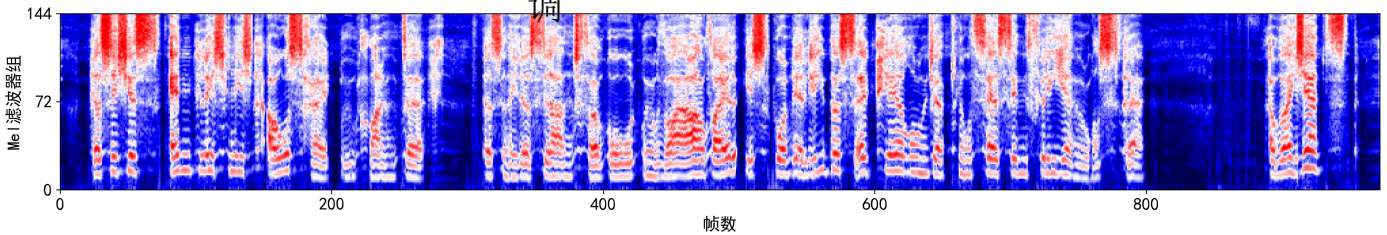
- 实验结果



(a)真实语音



(b)重构语音，无微调



(c)重构语音，微调（回放句数=20）

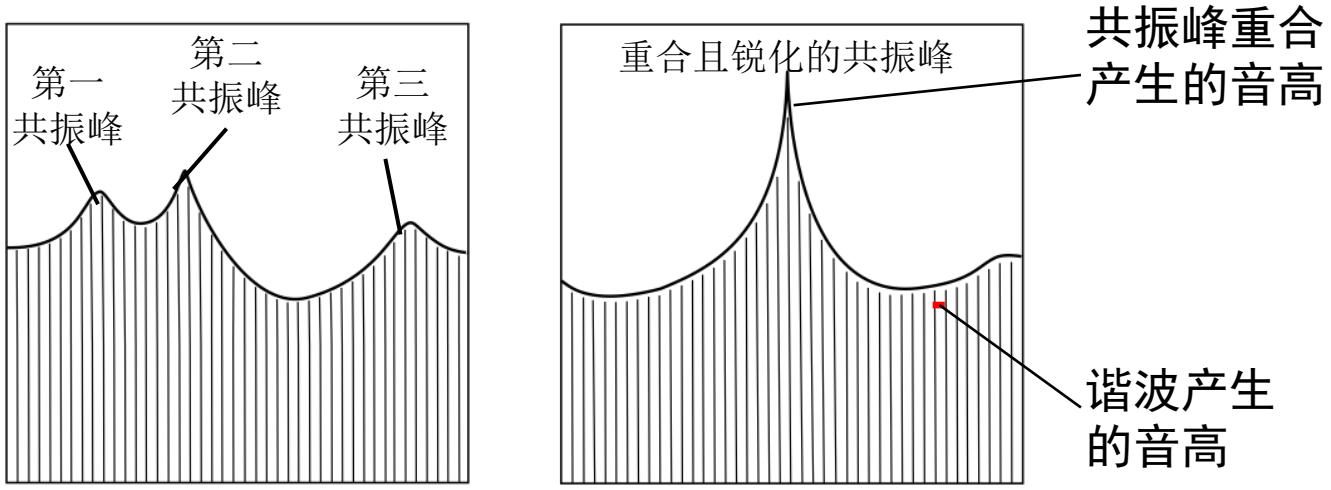
对德语（女）新说话人微调前后的样例

演示：呼麦及其产生机制

- 呼麦
  - 蒙古族、图瓦人的一种民族音乐艺术形式
  - 单个歌者同时产生两个音高

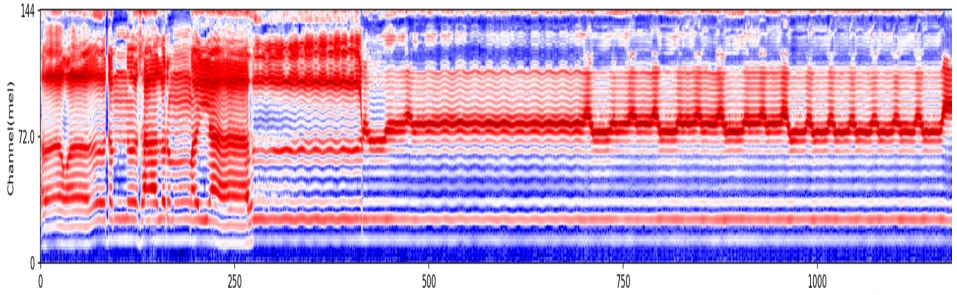
产生机制：

- 一种运用泛音的歌唱方式，第一、第二共振峰重合

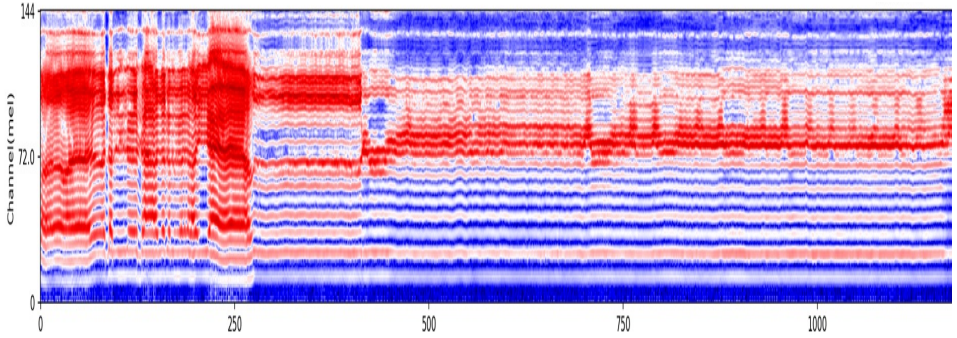


问题：发音器官采取何种姿态可使得前两个共振峰重合？

演示：呼麦反演



（a）实录呼麦的部分语谱



（b）重构呼麦的部分语谱

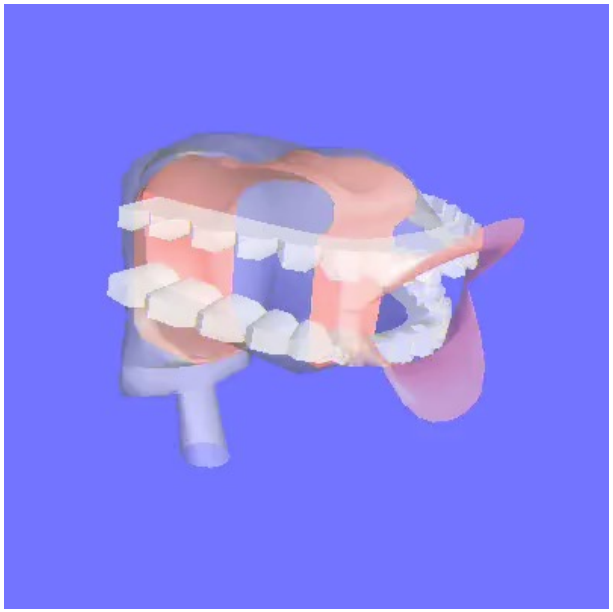


图3. 重构的呼麦与估计的发音动作

So how do we do it?

- Categorical Perception
  - we impose categories on physically continuous stimuli
- Top-Down effects

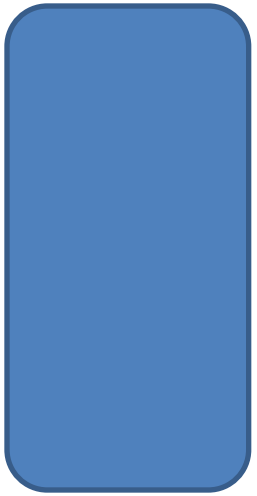
Top-Down Effects: Fence Effect

"The governor met with the state legi*lature earlier in the year."

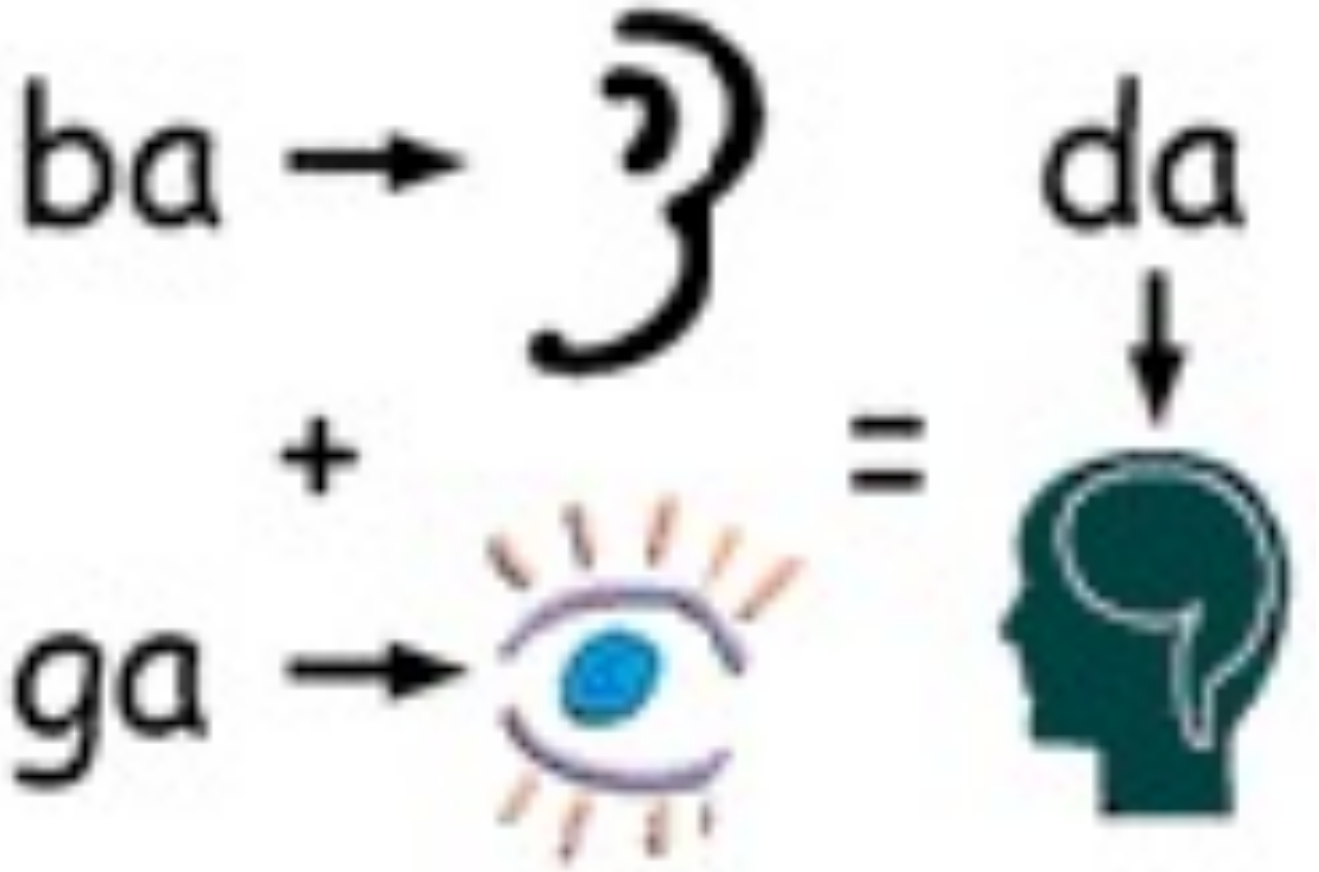
(* = cough)

Top-Down Effects: Use of Context

- It was found that the *eel was on the axle.
- It was found that the *eel was on the shoe.
- It was found that the *eel was on the orange.
- It was found that the *eel was on the table.



Top-Down Effects: McGurk Effect



So how do we do it?

- Categorical Perception
  - we impose categories on physically continuous stimuli
- Top-Down effects
  - Use of Context
  - Multi modal： McGurk Effect
  - Fence Effect

Acquisition of language - Learning words

- Speech is continuous - unreliable gaps between words
- No silences between words to mark word boundaries
- How does infant learn words？

Infants prefer extra pauses at clause boundaries

"(Cinderella lived) in a great big house, // but it was \\ sort of dark // because she had \\ this mean, mean, mean stepmother. // And oh she \\ had two stepsisters // that were so \\ ugly. // They were mean \\ too. They were…" *(Recording of mother talking to 19 month-old child)*

- 7-10 month old infants prefer the passage with extra 1 sec pauses at clause boundaries (//) than within clauses (\\).
- But no similar result when using speech directed to an adult. (Jusczyk, 1989)
- So the exaggerated prosody of child-directed speech is important.

Using prosodic information

- Newborn infants have learned about the rhythmic properties of their mother's language (similar abilities to the non-human primate tamarins).

- They can use this knowledge:
  - to divide speech up into clauses
  - to help them to segment words

  (But how do they know that this strategy works since it is language specific? Eg strong-weak more common in English than weak-strong)

Infant word segmentation

- Knowing some words helps to identify where the others are.

- Infants boot-strap themselves into speech recognition.

Conclusion

- The Structure of Speech
  - ✓Articulatory & Acoustic Phonetics
- Why is speech recognition hard?
  - ✓lack of invariance & segmentation problems
- So how do we do it?
  - ✓ Categorical perception
  - ✓ Top-down processing
- Acquisition of language

Q & A