Cognitive Media Processing #12

Nobuaki Minematsu

Menu of the last four lectures

Robust processing of easily changeable stimuli

- Robust processing of general sensory stimuli
- Any difference in the processing between humans and animals?

Human development of spoken language

- Infants' vocal imitation of their parents' utterances
- What acoustic aspect of the parents' voices do they imitate?

Speaker-invariant holistic pattern in an utterance

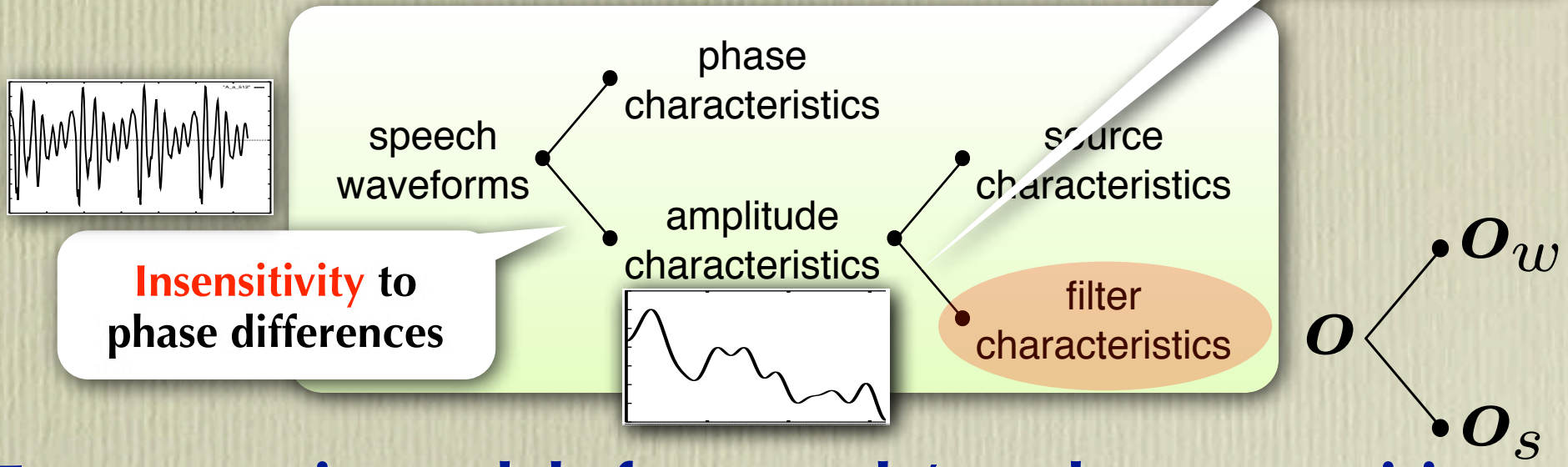
- Completely transform-invariant features -- *f*-divergence --
- Implementation of word Gestalt as relative timbre perception
- Application of speech structure to robust speech processing

Radical but interesting discussion

- A hypothesis on the origin and emergence of language
- What is your definition of "human-like" machines?

Feature separation to find specific info.

De facto standard acoustic analysis of s



Two acoustic models for speech/speaker recognition

- Speaker-independent acoustic model for **w**ord recognition
  - $P(o|w) = \sum_s P(o,s|w) = \sum_s P(o|w,s)P(s|w) \sim \sum_s \underline{P(o|w,s)}P(s)$
- Text-independent acoustic model for **s**peaker recognition
  - $P(o|s) = \sum_w P(o,w|s) = \sum_w P(o|w,s)P(w|s) \sim \sum_w \underline{P(o|w,s)}P(w)$
- Require intensive collection
  - $o \rightarrow o_w + o_s$ is possible or not?

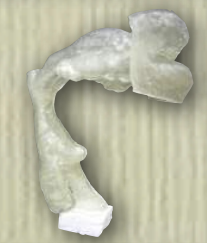
“Separately brought up identical twins”

The parents get divorced immediately after the birth.

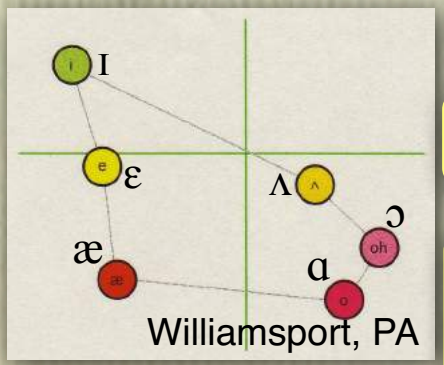
- The twins were brought up separately by the parents.
- What kind of pron. will the twins have acquired 5 years later?





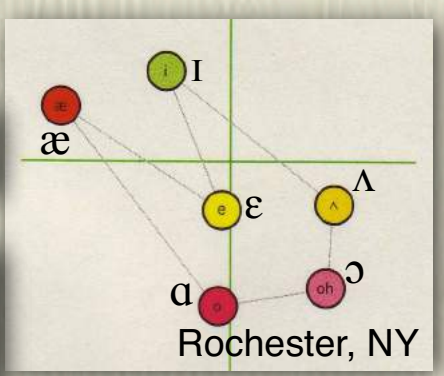



Diff. of VTL = Diff. of timbre



Diff. of regional accents = Diff. of timbre

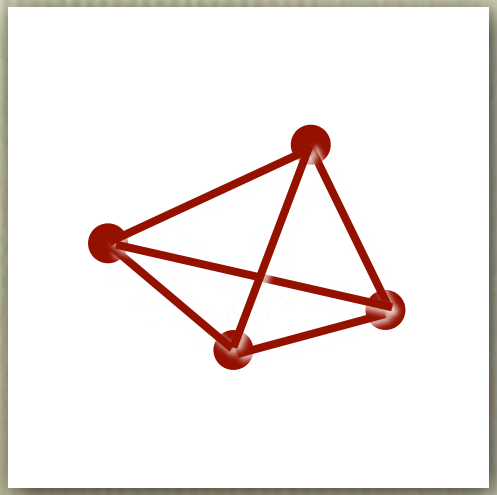
Machines that don’t learn what infants don’t learn.

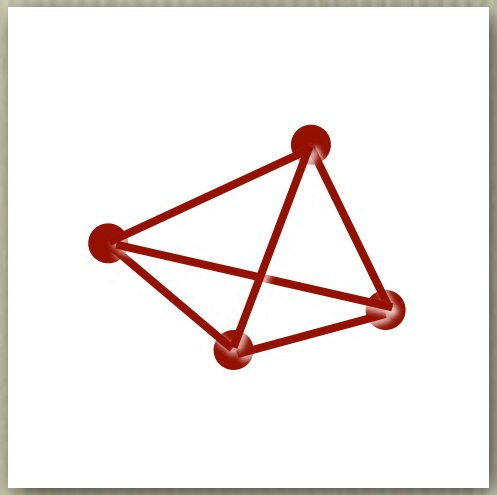


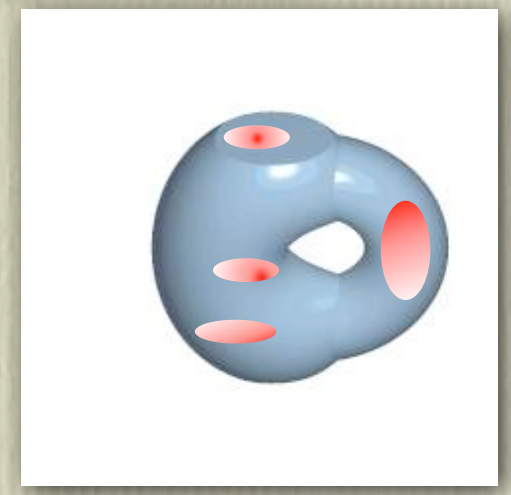
Invariance in variability

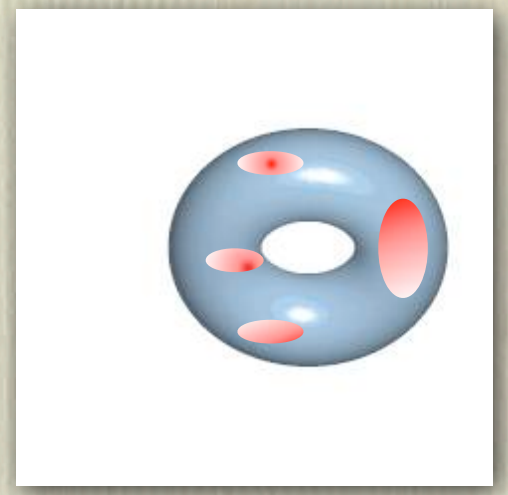
Topological invariance [Minematsu'09]

- Topology focuses on invariant features wrt. any kind of deformation.





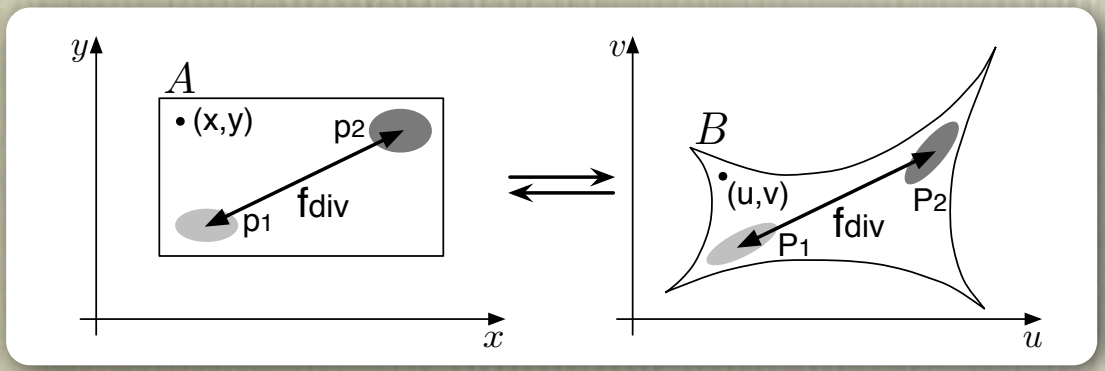




Complete transform-invariance

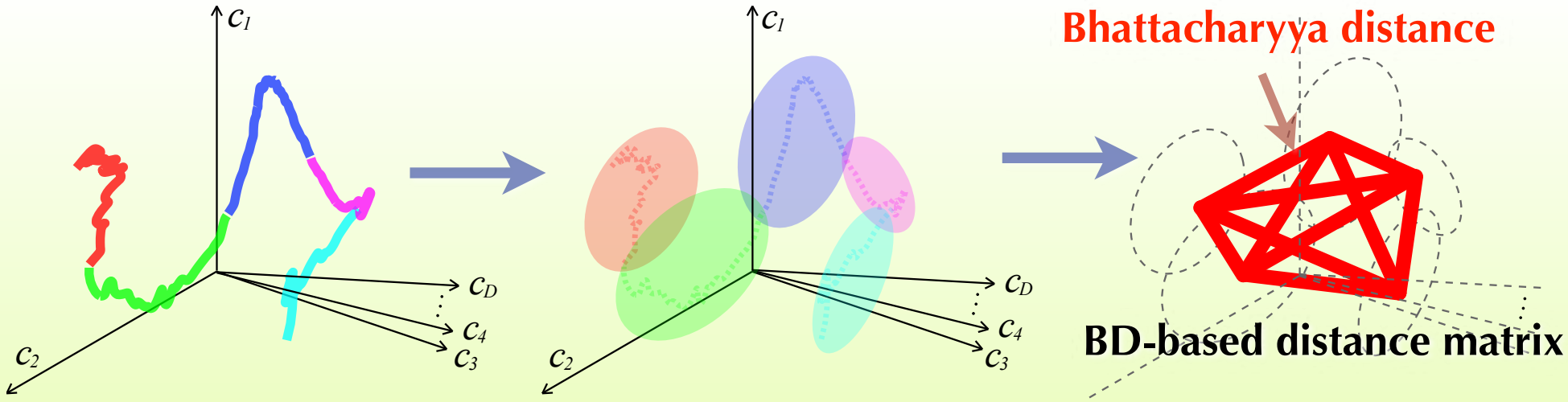
Any general expression for invariance? [Qiao'10]

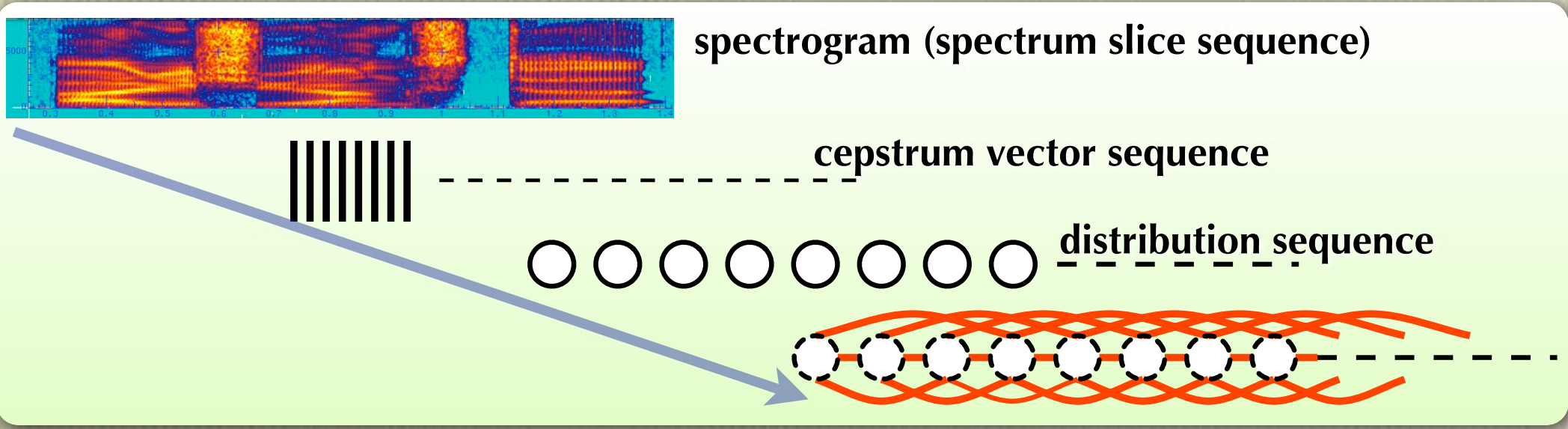
- BD is just one example of invariant contrasts.
- f-divergence is invariant with any kind of transformation.
  - $f_{div}(p_1, p_2) = \int p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}$
  - $g(t) = t\log(t) \rightarrow f_{div} = \mathrm{KL-div.} \qquad g(t) = \sqrt{t} \rightarrow -\log(f_{div}) = \mathrm{BD}$
  - $f_{div}(p_1, p_2) = f_{div}(P_1, P_2)$
- Invariant features have to be f-divergence.
  - If $\oint M(p_1(\boldsymbol{x}), p_2(\boldsymbol{x}))d\boldsymbol{x}$ is invariant with any transformation,
  - The following condition has to be satisfied. $M = p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right)$



Invariant speech structure

Utterance to structure conversion using *f*-div. [Minematsu'06]



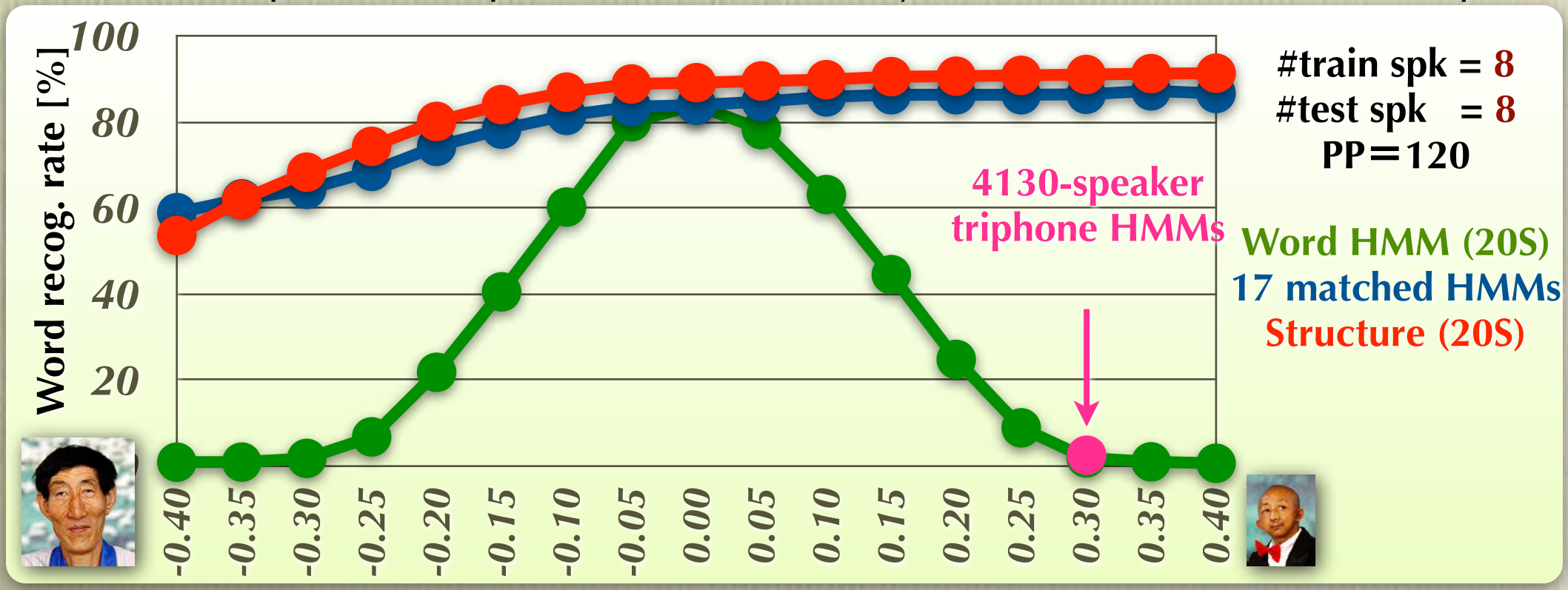


- An event (distribution) has to be much smaller than a phoneme.

Application of structures to ASR

Isolated word recognition using warped utterances

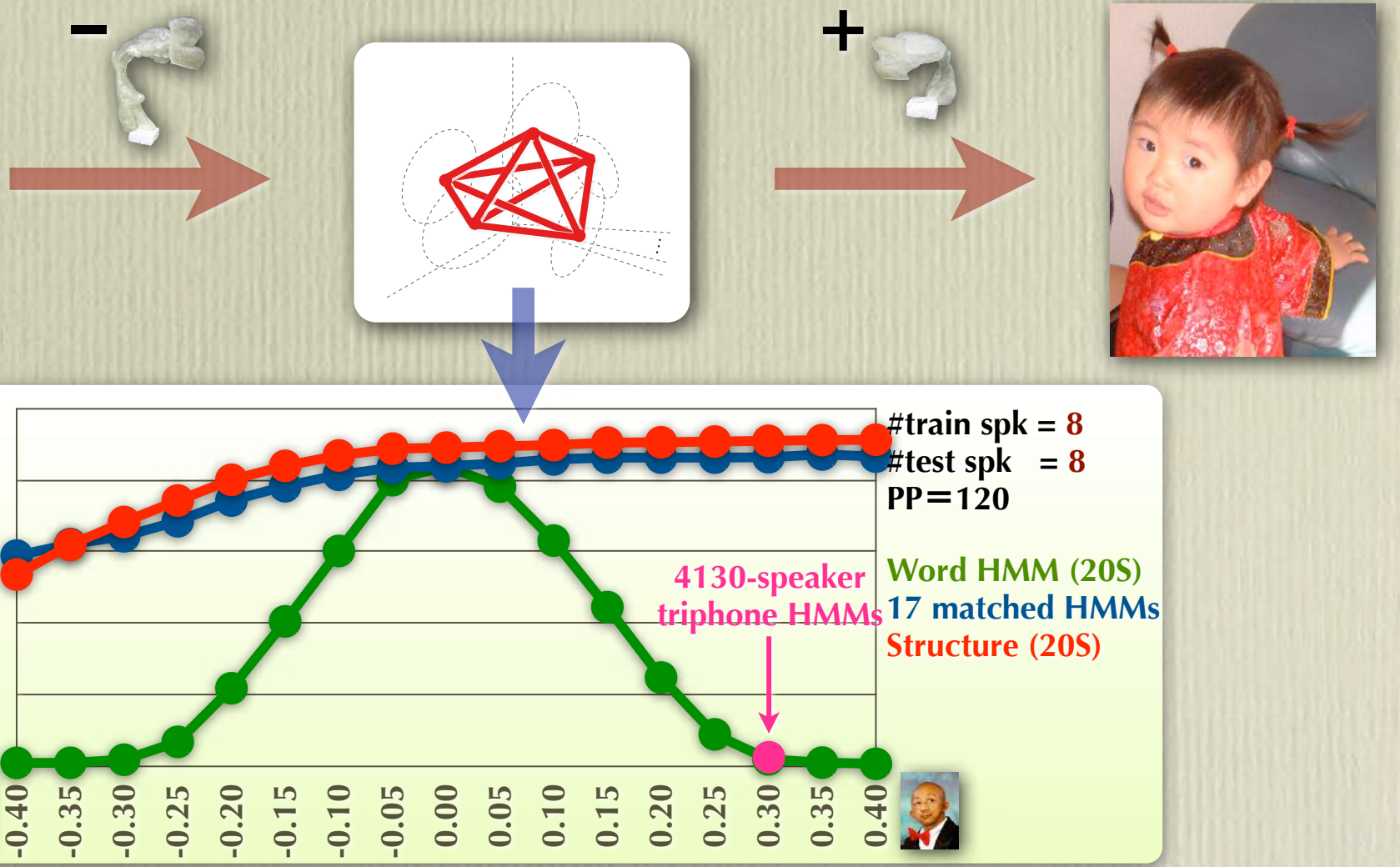
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used as baseline model in Japan



An experiment with real vocal imitation

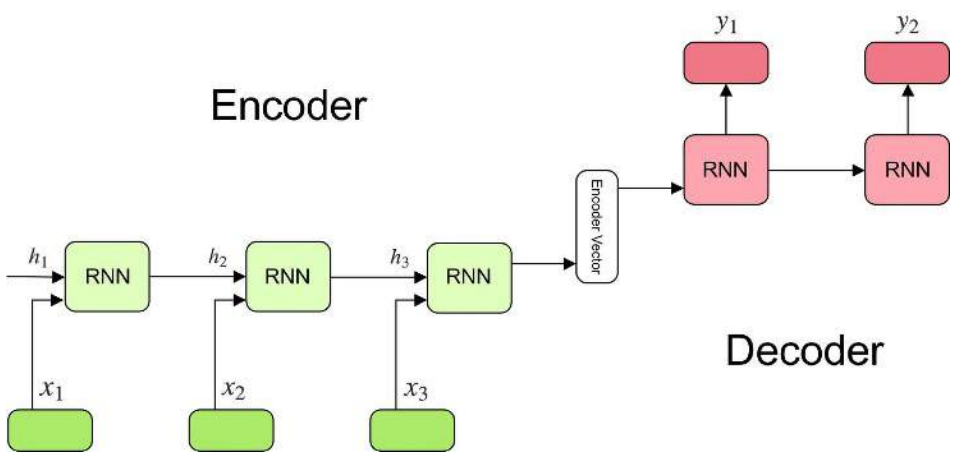
Demonstration with my wife and daughter

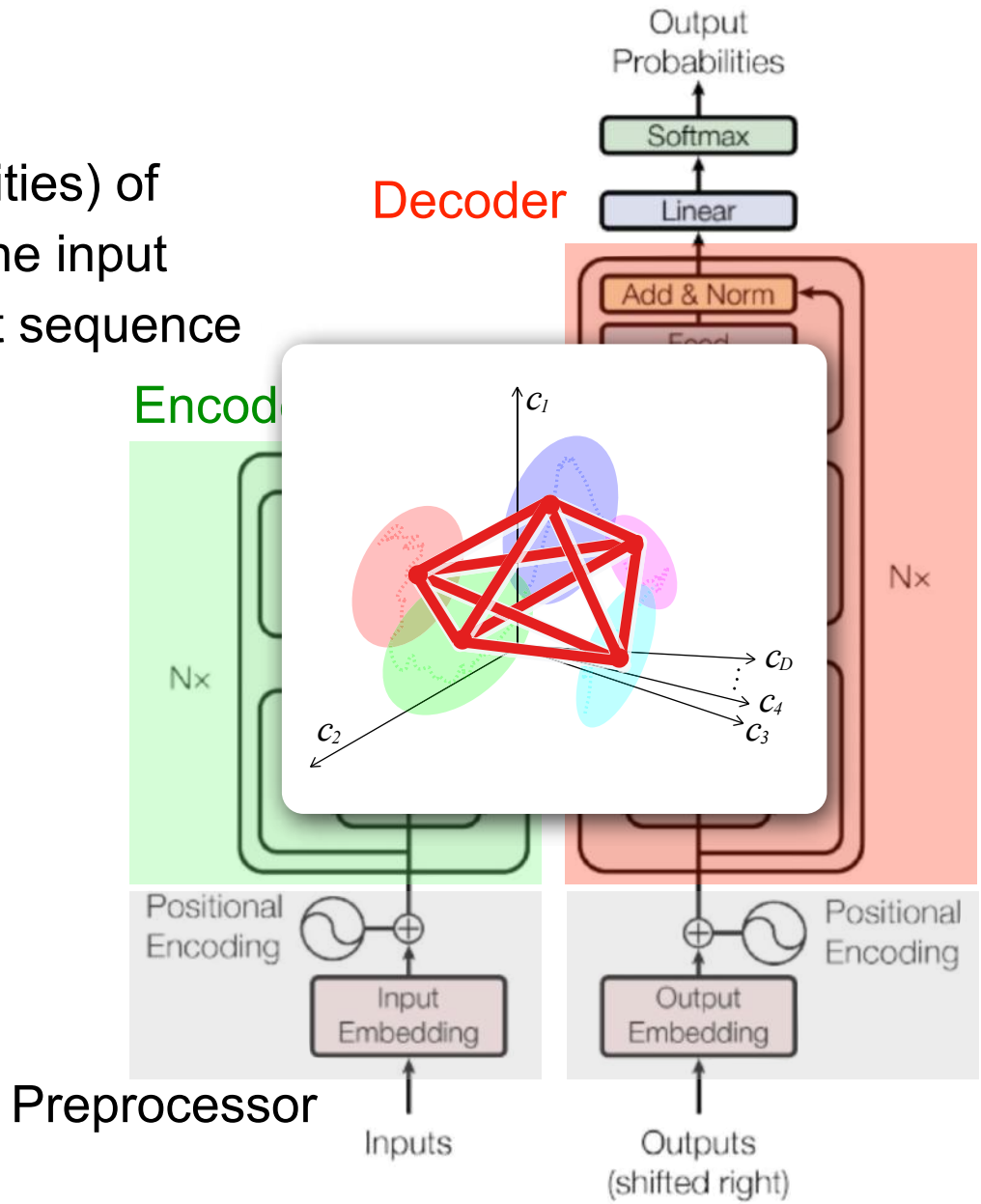
- Constraint conditions are given by my wife.
- Initial conditions are given by my daughter.



Transformer model

- **Attention is all you need !!**
  - https://arxiv.org/abs/1706.03762
  - Explicit modeling of the relations (similarities) of the current input token to other ones in the input sequence and to the tokens in the output sequence generated so far.
    - Self-attention mechanism

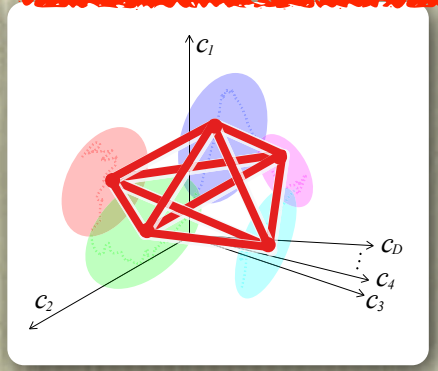




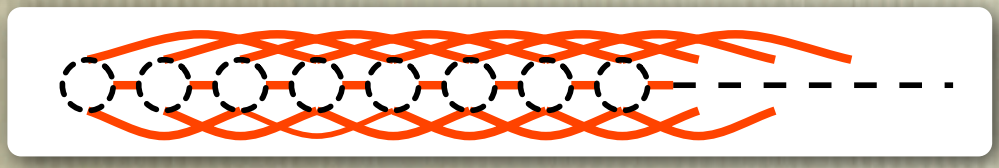
More classical claims in linguistics

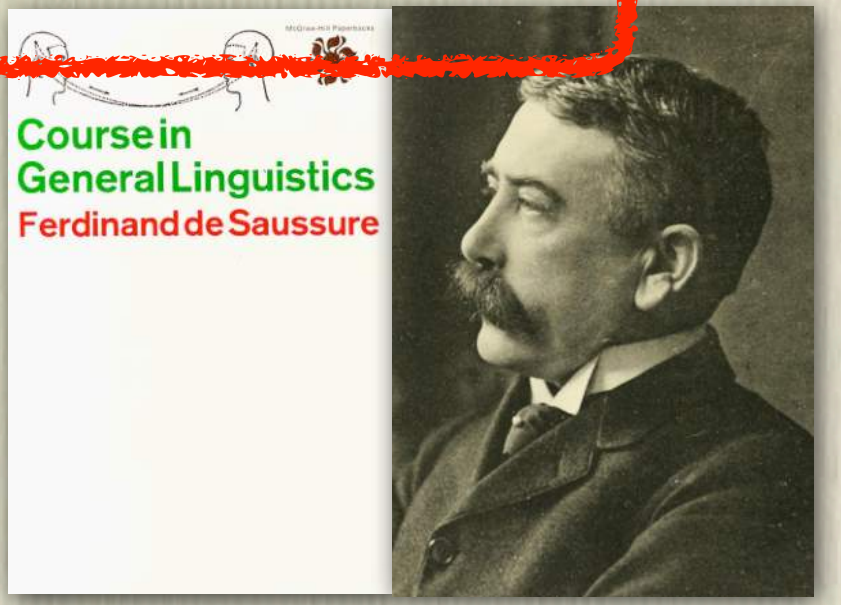
Ferdinand de Saussure (1857-1913)

- Father of modern linguistics
- "Course in General Linguistics" (1916)
- What defines a linguistic element, conceptual or phonic, is the relation in which it stands to the other elements in the linguistic system.
- The important thing in the word is not the sound alone but the phonic differences that make it possible to distinguish this word from the others.
- Language is a system of only conceptual differences and phonic differences.



$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ d_{31} & & & \\ \vdots & & & \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix}$$





A big solution for CALL development

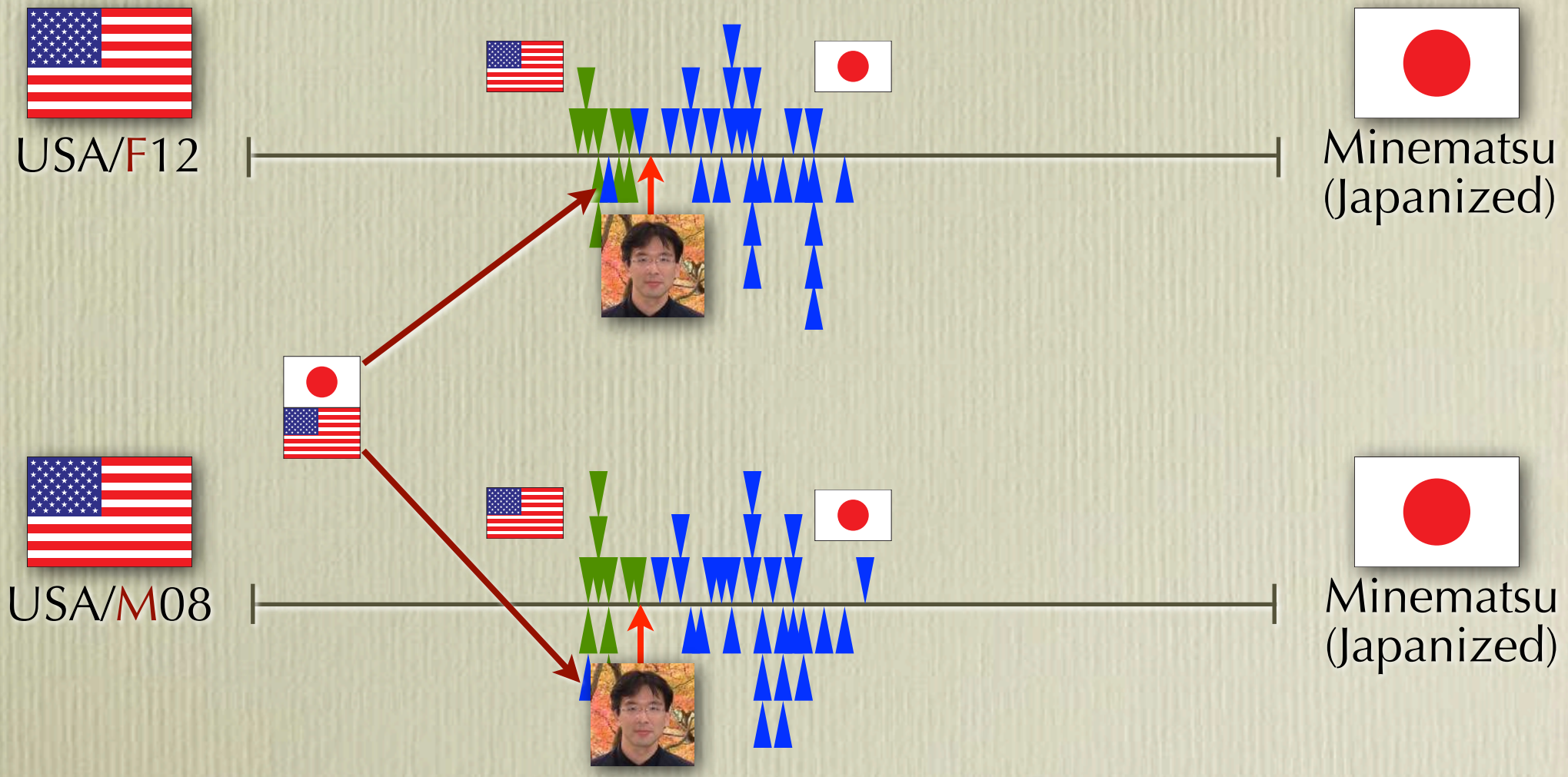
To which does Minematsu's normal English sound closer ?

speaker	USA/F12	✗	Minematsu	○	Minematsu
gender	female	✗	male	○	male
age	?	✗	37	○	37
mic	Sennheiser	✗	cheap mic	○	cheap mic
room	recording room	✗	living room	○	living room
AD	SONY DAT	✗	PowerBook	○	PowerBook
proficiency	perfect	△	good	✗	Japanized

(Minematsu@ICSLP 2004)

A big solution for CALL development

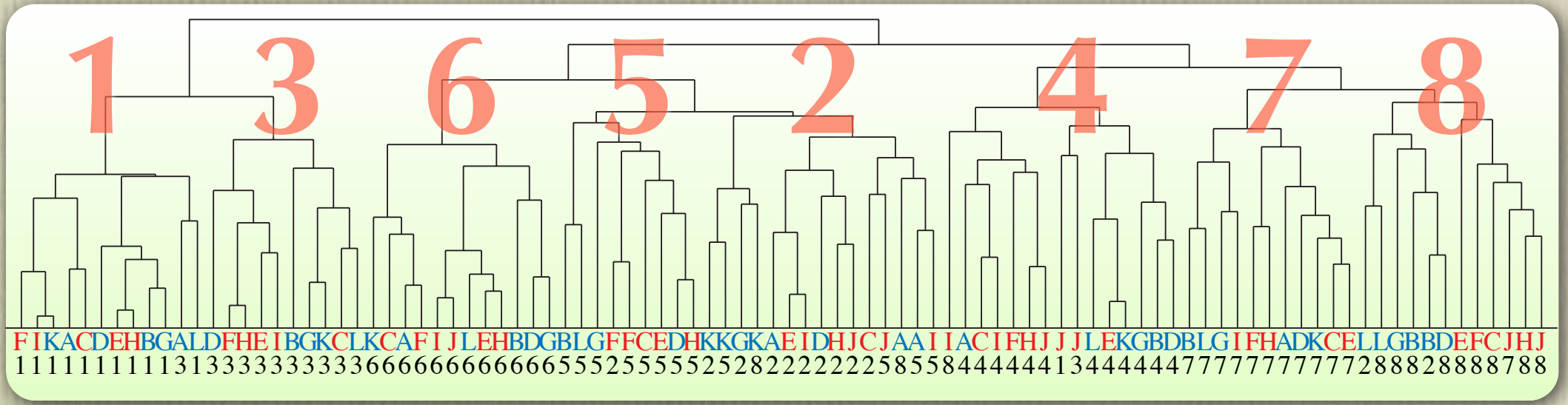
Proficiency estimation based on structural distance



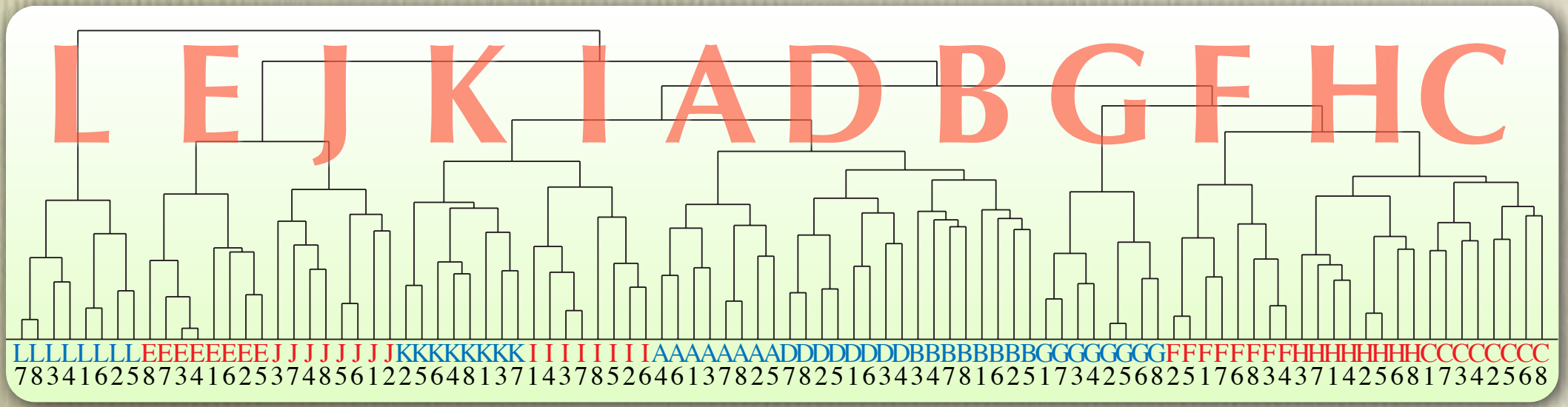
(Minematsu@ICSLP 2004)

Clustering of learners

Contrast-based comparison



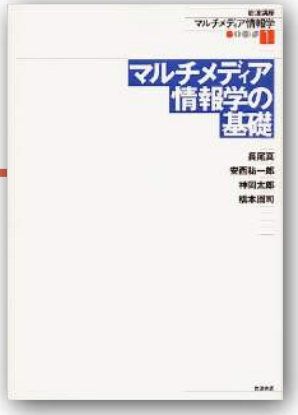
Substance-based comparison



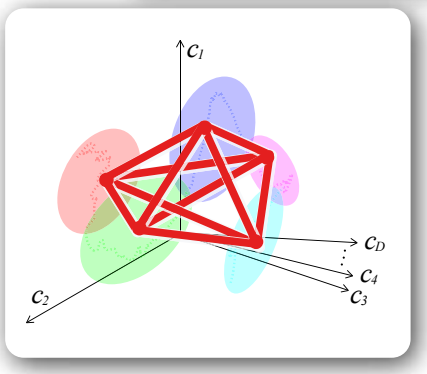
A new framework for “human-like” speech machines #4

Nobuaki Minematsu

Title of each lecture



- Theme-1
  - ~~Multimedia information and humans~~
  - ~~Multimedia information and interaction between humans and machines~~
  - ~~Multimedia information used in expressive and emotional processing~~
  - ~~A wonder of sensation - synesthesia -~~
- Theme-2
  - ~~Speech communication technology - articulatory & acoustic phonetics -~~
  - ~~Speech communication technology - speech analysis -~~
  - ~~Speech communication technology - speech recognition -~~
  - ~~Speech communication technology - speech synthesis -~~
- Theme-3
  - ~~A new framework for "human-like" speech machines #1~~
  - ~~A new framework for "human-like" speech machines #2~~
  - ~~A new framework for "human-like" speech machines #3~~
  - A new framework for "human-like" speech machines #4



Menu of the last four lectures

Robust processing of easily changeable stimuli

- Robust processing of general sensory stimuli
- Any difference in the processing between humans and animals?

Human development of spoken language

- Infants' vocal imitation of their parents' utterances
- What acoustic aspect of the parents' voices do they imitate?

Speaker-invariant holistic pattern in an utterance

- Completely transform-invariant features -- *f*-divergence --
- Implementation of word Gestalt as relative timbre perception
- Application of speech structure to robust speech processing

Radical but interesting discussion

- A hypothesis on the origin and emergence of language
- What is your definition of "human-like" machines?

A MODULATION-DEMODULATION MODEL FOR SPEECH COMMUNICATION AND ITS EMERGENCE

NOBUAKI MINEMATSU

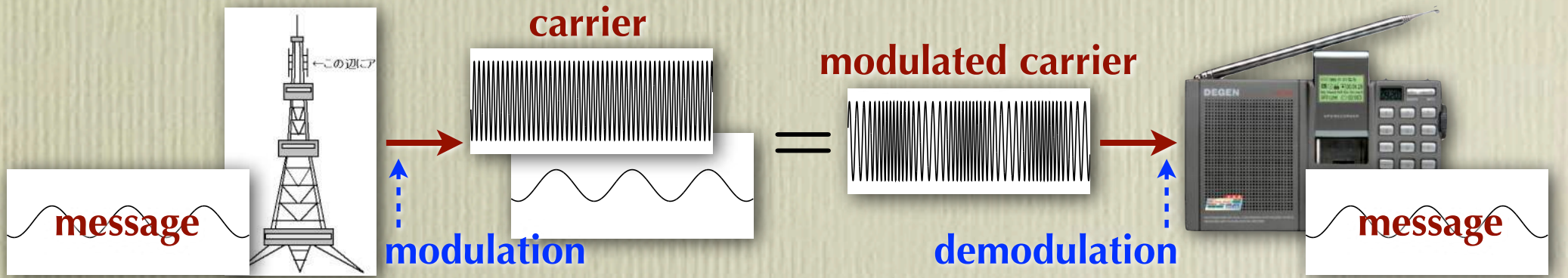
Graduate School of Info. Sci. and Tech., The University of Tokyo, Japan,
mine@gavo.t.u-tokyo.ac.jp

Perceptual invariance against large acoustic variability in speech has been a long-discussed question in speech science and engineering (Perkell & Klatt, 2002), and it is still an open question (Newman, 2008; Furui, 2009). Recently, we proposed a candidate answer based on mathematically-guaranteed relational invariance (Minematsu et al., 2010; Qiao & Minematsu, 2010). Here, transform-invariant features, f-divergences, are extracted from the speech dynamics in an utterance to form an invariant topological shape which characterizes and represents the linguistic message conveyed in that utterance. In this paper, this representation is interpreted from a viewpoint of telecommunications, linguistics, and evolutionary anthropology. Speech production is often regarded as a process of modulating the baseline timbre of a speaker's voice by manipulating the vocal organs, i.e., spectrum modulation. Then, extraction of the linguistic message from an utterance can be viewed as a process of spectrum *de*modulation. This modulation-demodulation model of speech communication has a strong link to known morphological and cognitive differences between humans and apes.

Modulation used in telecommunication

From Wikipedia

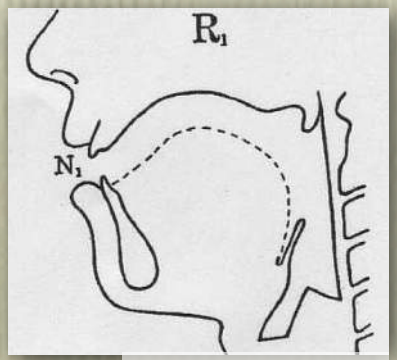
A musician modulates the tone from a musical instrument by varying its volume, timing and pitch. The three key parameters of a carrier sine wave are its amplitude ("volume"), its phase ("timing") and its frequency ("pitch"), all of which can be modified in accordance with a content signal to obtain the modulated carrier.



A way of characterizing speech production

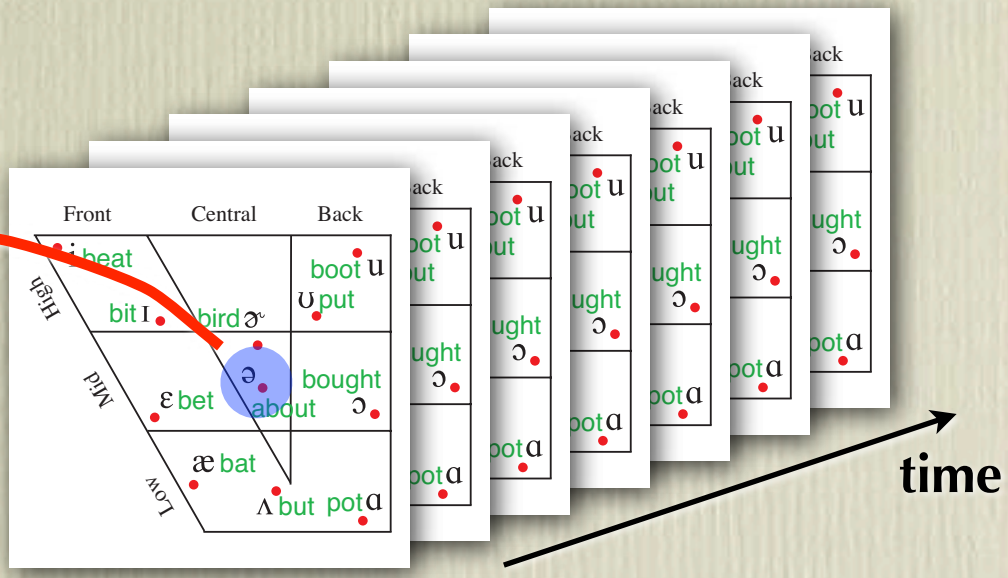
Speech production as spectrum modulation

- Modulation in frequency (FM), amplitude (AM), and phase (PM)
  - = Modulation in pitch, volume, and timing (from Wikipedia)
  - = Pitch contour, intensity contour, and rhythm (= prosodic features)
- What about a fourth parameter, which is **spectrum (timbre)**?
  - = Modulation in spectrum (timbre) [Scott'07]
  - **= Another prosodic feature?**



Tongue = modulator

Schwa
= most lax
= most frequent
= home position
= spk.-specific baseline timbre

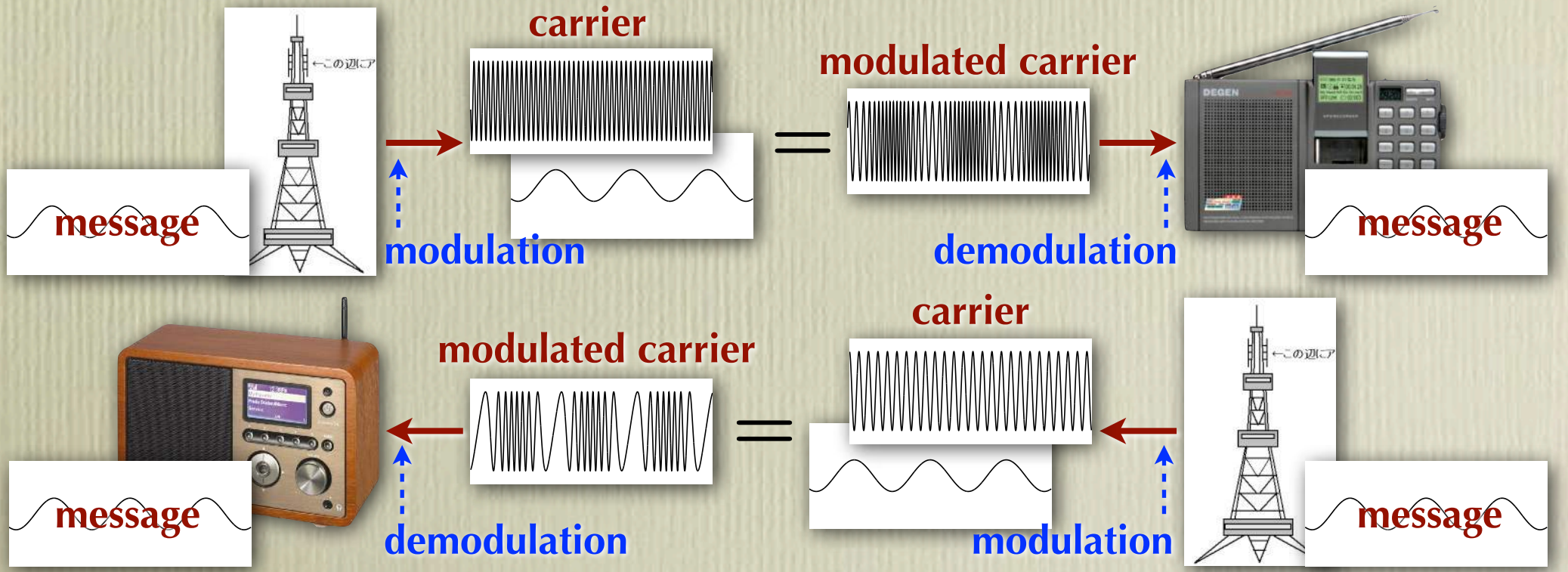


Demodulation used in telecommunication

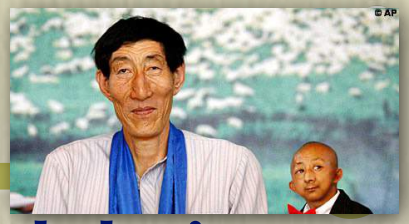
Demodulation in frequency, amplitude, and phase

- Demodulation = a process of extracting the message intactly by removing the carrier component from the modulated carrier signal.
  - Not by extensive collection of samples of modulated carriers
  - (Not by hiding the carrier component by extensive collection)

$$P(o|w) = \sum_s P(o, s|w)$$



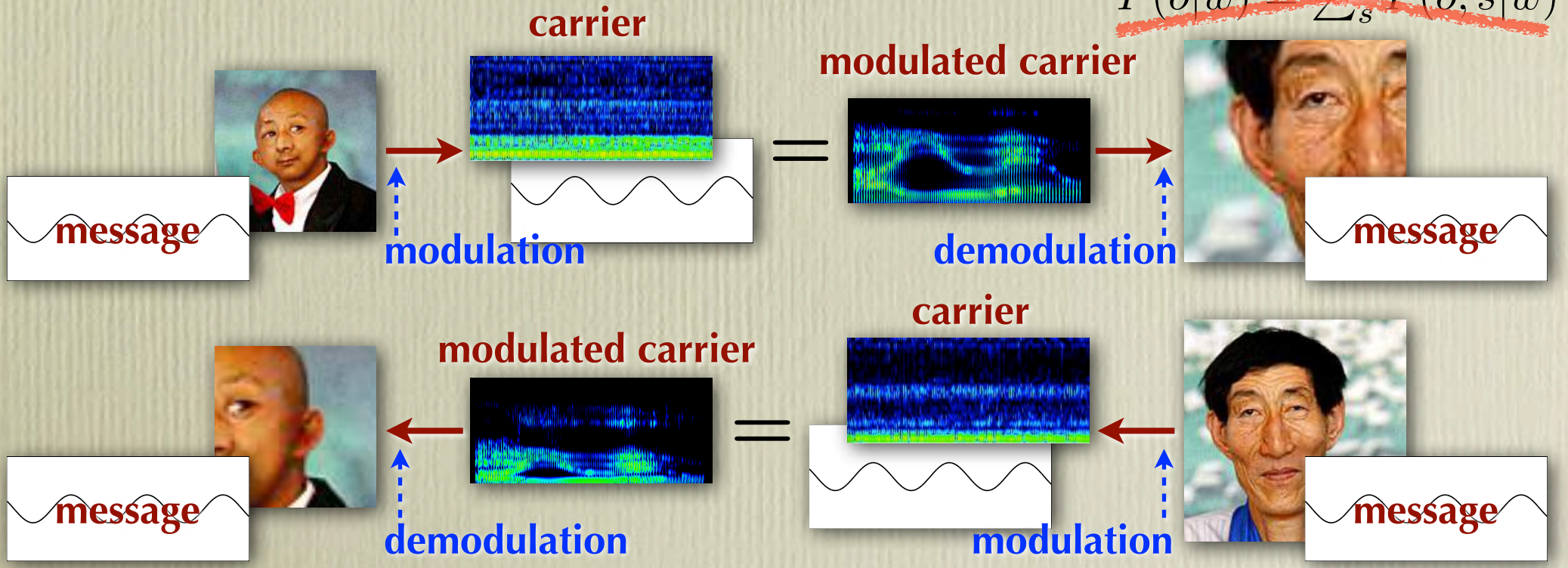
Spectrum demodulation



Speech recognition = spectrum (timbre) demodulation

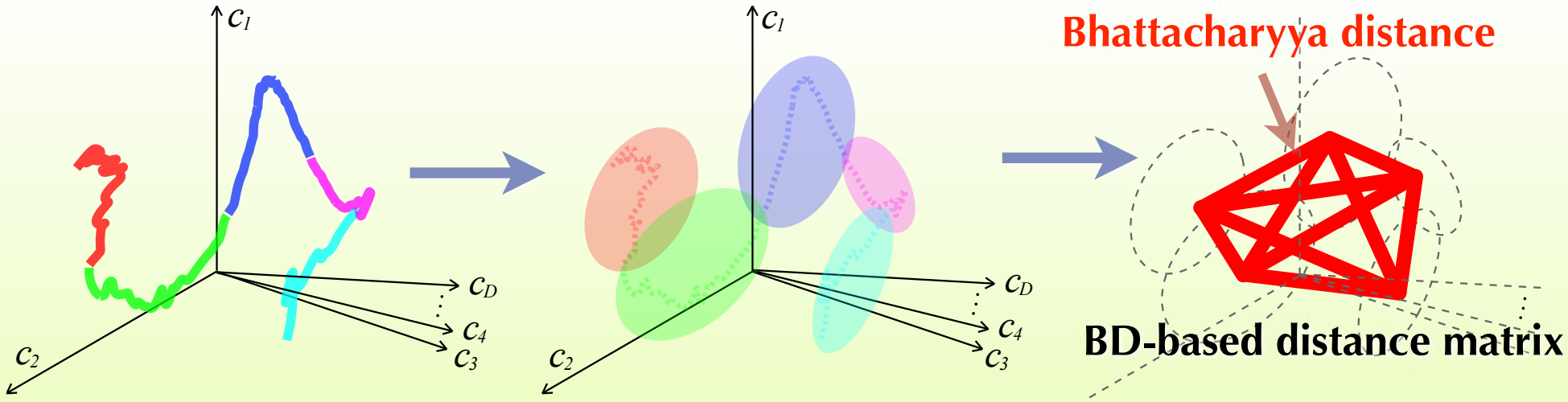
- Demodulation = a process of extracting a message intactly by removing the carrier component from the modulated carrier signal.
  - By removing speaker-specific baseline spectrum characteristics
  - Not by extensive collection of samples of modulated carriers
  - (Not by hiding the carrier component by extensive collection)

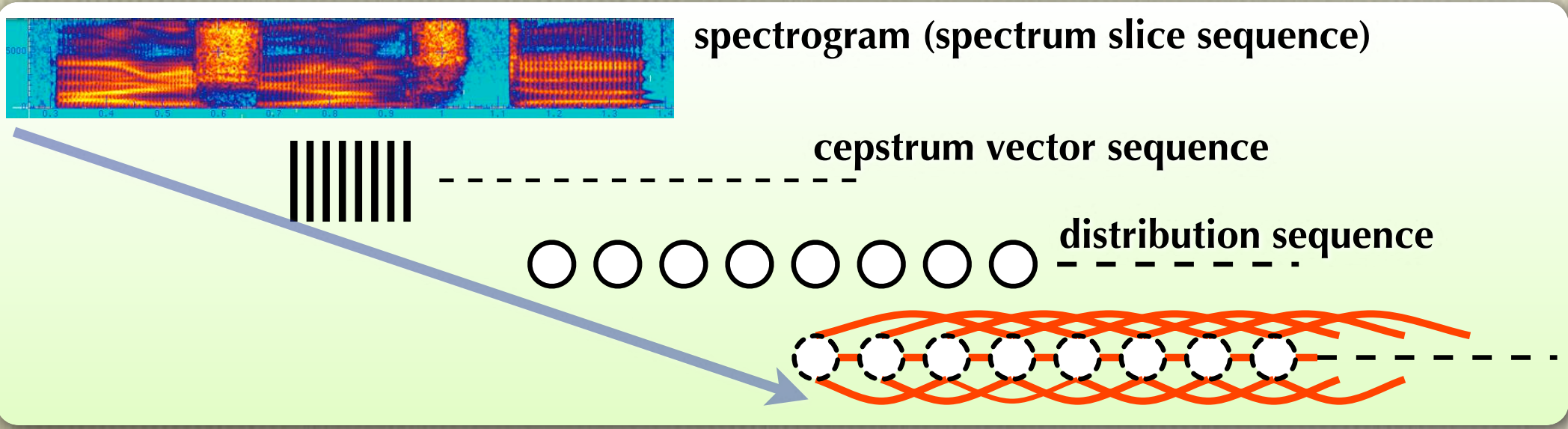
$$P(o|w) = \sum_s P(o, s|w)$$



Invariant speech structure

Utterance to structure conversion using *f*-div. [Minematsu'06]





- An event (distribution) has to be much smaller than a phoneme.

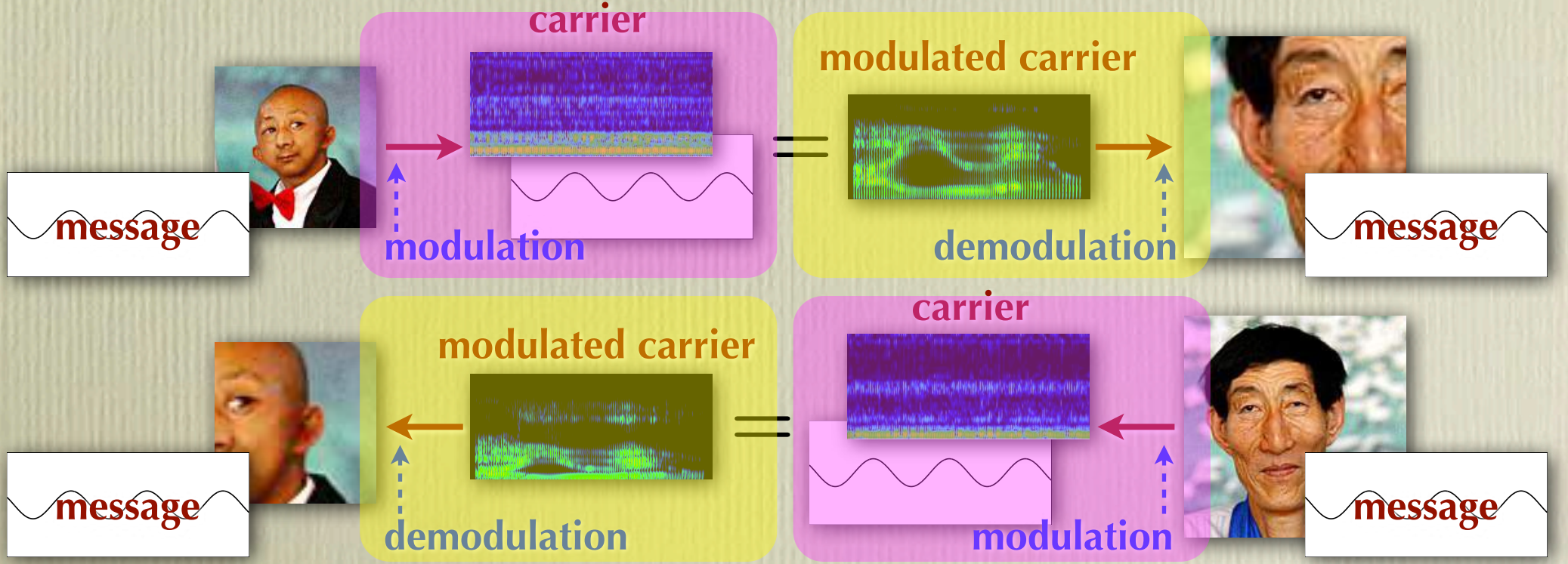
Two questions

Q1: Does an ape have a good modulator?

- Does the tongue of an ape work as a good modulator?

Q2: Does an ape have a good demodulator?

- Does the ear (brain) of an ape extract the message intactly?



Structural diff. in the mouth and the nose

フィオレンツォ・ファッキーニ著「人類の起源」同朋社出版　P114～115の図を改変

鼻腔
硬口蓋
舌
下顎骨
舌骨
喉頭蓋
喉頭室と声帯
pharynx
larynx
lun

チンパンジー

喉頭蓋は軟口蓋とほんの少しだけ離れてい

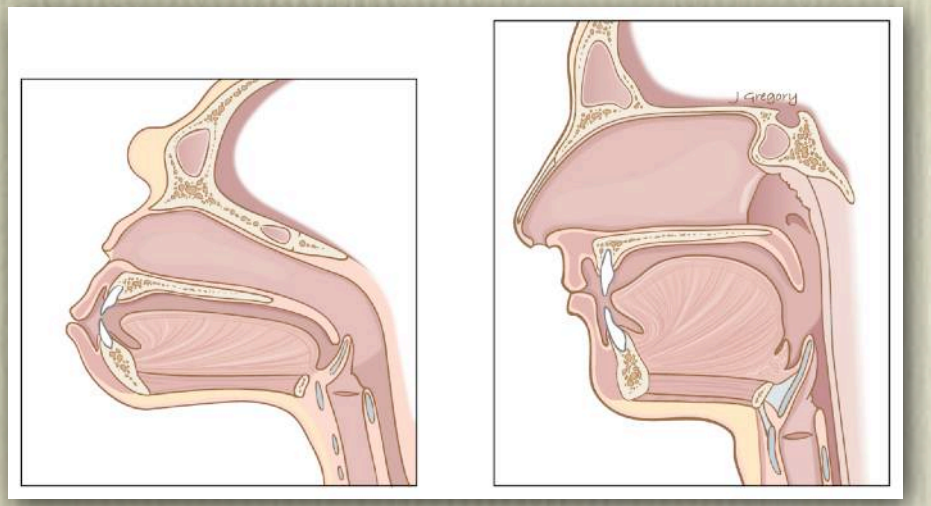
喉頭が下がっているため、喉頭蓋
って声帯でつくられた音が共鳴す
いる。ヒトが発する声音の大部分
ることでつくられる。

鼻腔
軟口蓋
pharynx
喉頭蓋
larynx
lung
stomach

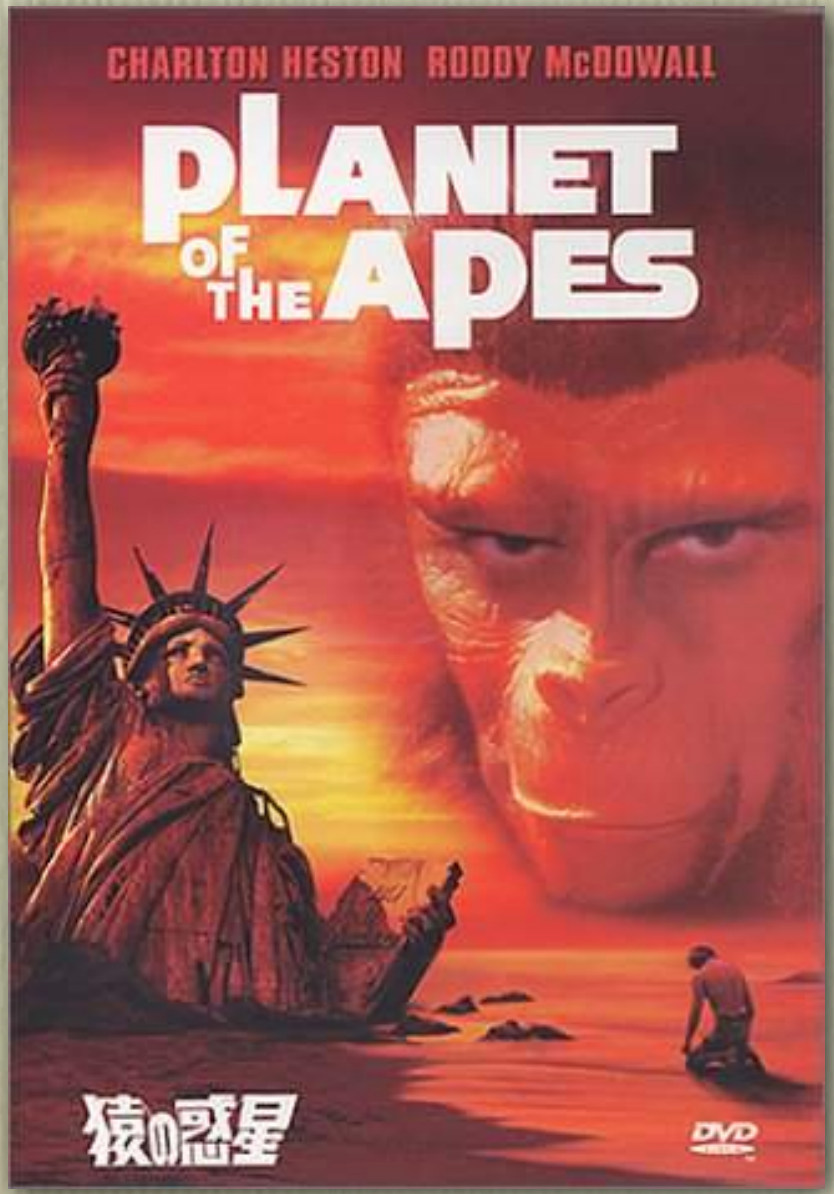
Flexibility of tongue motion

The chimp's tongue is much stiffer than the human's.

- "Morphological analyses and 3D modeling of the tongue musculature of the chimpanzee" (Takemoto'08)
  - Less capability of manipulating the shape of the tongue.



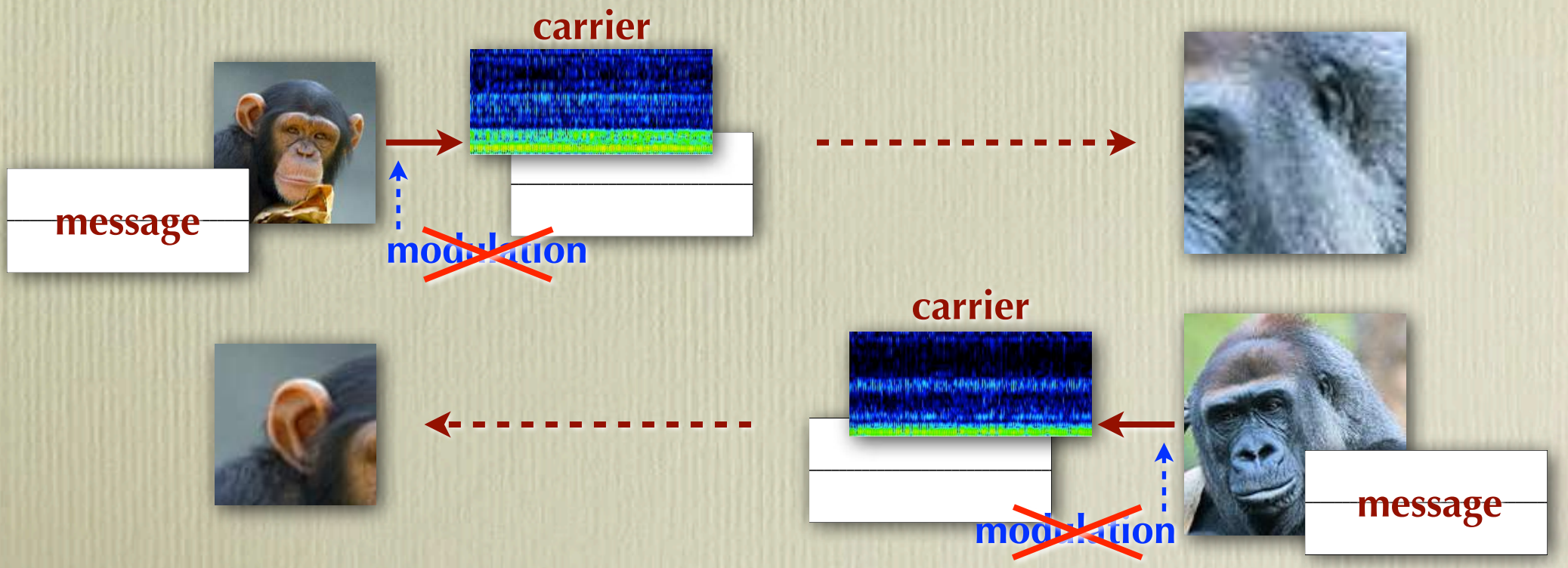
Old and new “Planet of the Apes”



Q1: Does the ape have a good modulator?

Morphological characteristics of the ape's tongue

- Two (almost) independent tracts [Hayama'99]
  - One is from the nose to the lung for breathing.
  - The other is from the mouth to the stomach for eating.
- Much lower ability of deforming the tongue shape [Takemoto'08]
  - The chimp's tongue is stiffer than the human's.



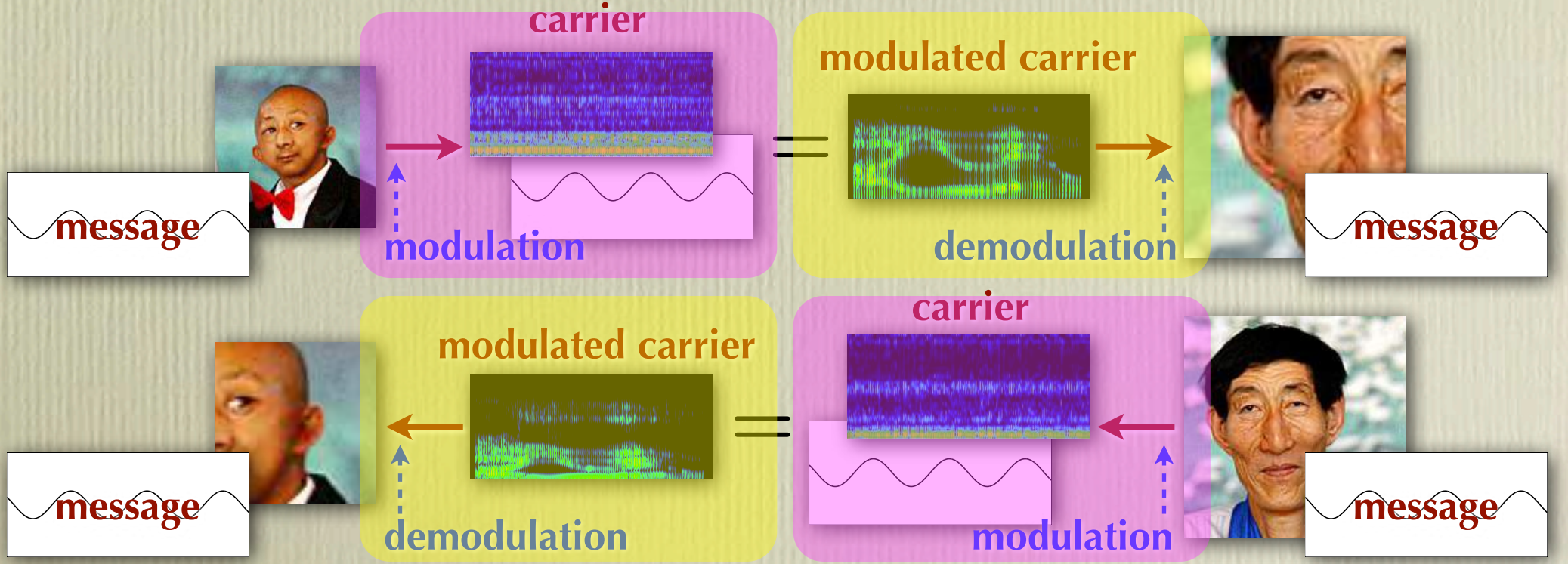
Two questions

Q1: Does the ape have a good modulator?

- Does the tongue of the ape work as a good modulator?

Q2: Does the ape have a good demodulator?

- Does the ear (brain) of the ape extract the message intactly?

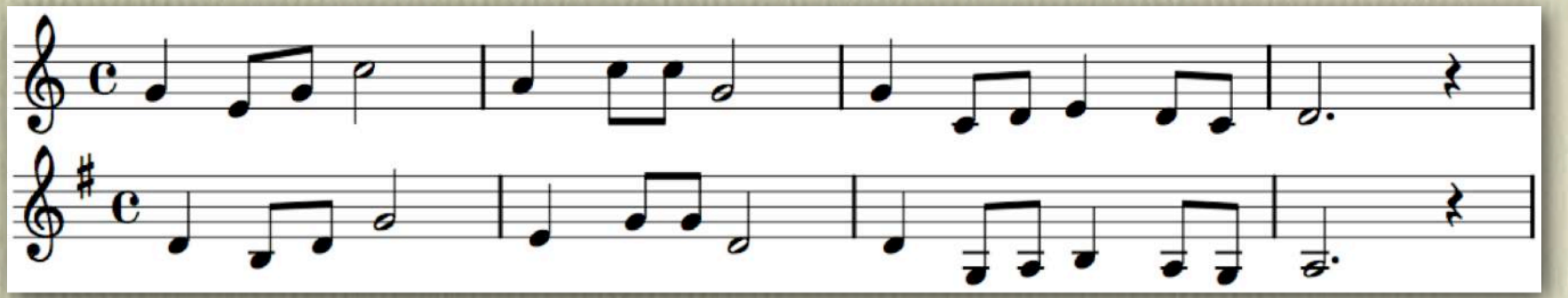


The nature's solution for static bias?

How old is the invariant perception in evolution? [Hauser'03]

1

2



1 = 2

At least, frequency (pitch) demodulation seems difficult.








Language acquisition through vocal imitation

VI = children's active imitation of parents' utterances

- Language acquisition is based on vocal imitation [Jusczyk'00].
- VI is very rate in animals. No other primate does VI [Gruhn'06].
- Only small birds, whales, and dolphins do VI [Okanoya'08].

A's VI = acoustic imitation but H's VI ≠ acoustic = ??

- Acoustic imitation performed by myna birds [Miyamoto'95]
  - They imitate the sounds of cars, doors, dogs, cats as well as human voices.
  - Hearing a very good myna bird say something, one can guess its owner.
- Beyond-scale imitation of utterances performed by children
  - No one can guess a parent by hearing the voices of his/her child.
  - Very weird imitation from a viewpoint of animal science [Okanoya'08].







A Disney file about an autistic boy

To make him recover from autistics, all the family members pretended (sounded) to be Disney characters.

Interview

How Disney gave voice to a boy with autism

Saskia Baron

As a young boy, Owen Suskind suddenly stopped talking. Diagnosed with autism, he remained largely silent until an obsession with Disney movies unexpectedly gave him a voice

Owen Suskind. Photograph: Courtesy of the Suskind family

Near the beginning of the new documentary Life, Animated, there is a home movie filmed by Cornelia Suskind in November 1993. Her husband, Ron, is playing in the garden of their old house

https://bit.ly/32YHKBa

The Suskinds have never found out what caused Owen to lose so many skills, but rather than dwell on possible causes they devoted themselves to exploring every therapy on offer. Ron's new position on the Wall Street Journal meant that Cornelia, also a journalist, could just about afford not to work. Instead, she organised and took part in a range of therapies for Owen. She also home educated him for a couple of years when the right school proved elusive. The family assembled a team of specialists to support them and give advice. Progress was painfully slow.

"Owen was just shy of seven and we realised that he was using movies to interpret our world

Many children with autism have favourite interests or activities that they never tire of repeating and which can appear to get in the way of them learning new skills or engaging with others. In Owen's case, his obsession was Disney. Despite his motor problems, he mastered the remote control for the family's video recorder and loved to watch the same films over and over

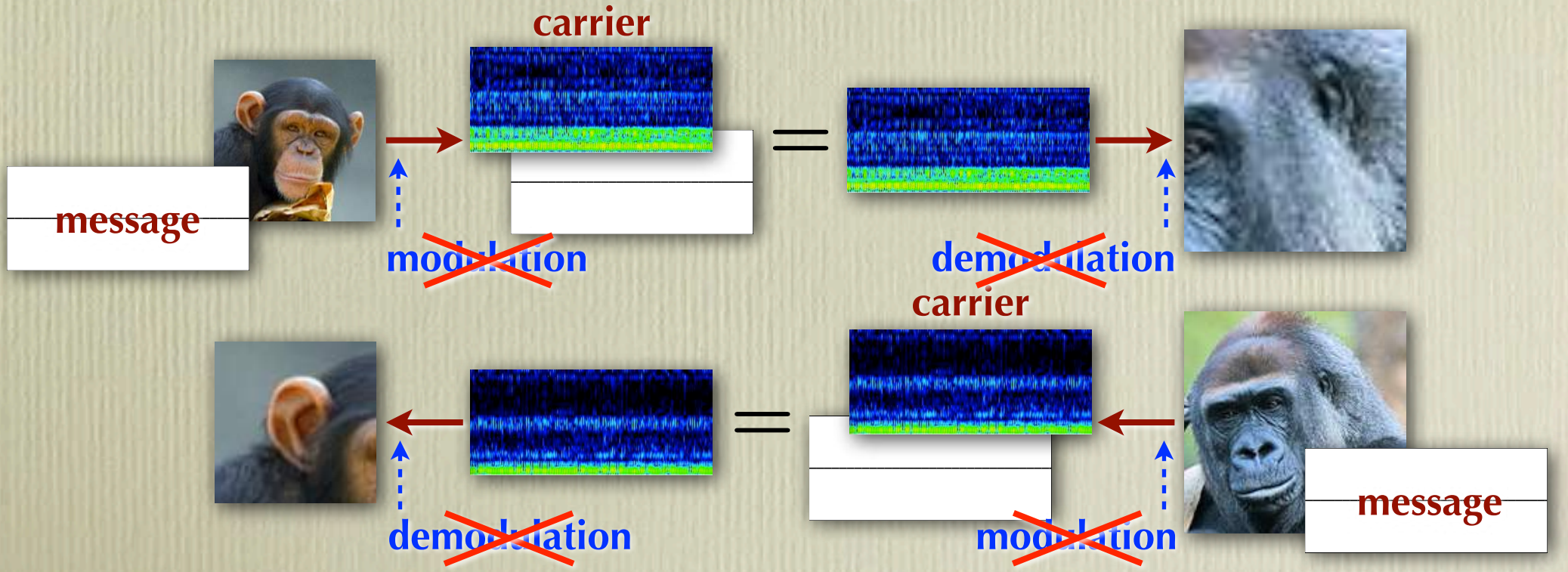
Q2: Does the ape have a good demodulator?

Cognitive difference bet. the ape and the human

- Humans can extract embedded messages in the modulated carrier.
- It seems that animals treat the (modulated) carrier as it is.

From the (modulated) carrier, what can they know?

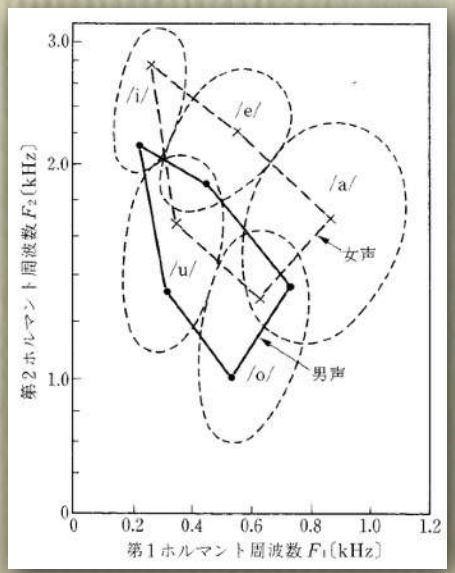
- The apes can identify individuals by hearing their voices.
  - Lower/higher formant frequencies = larger/smaller apes

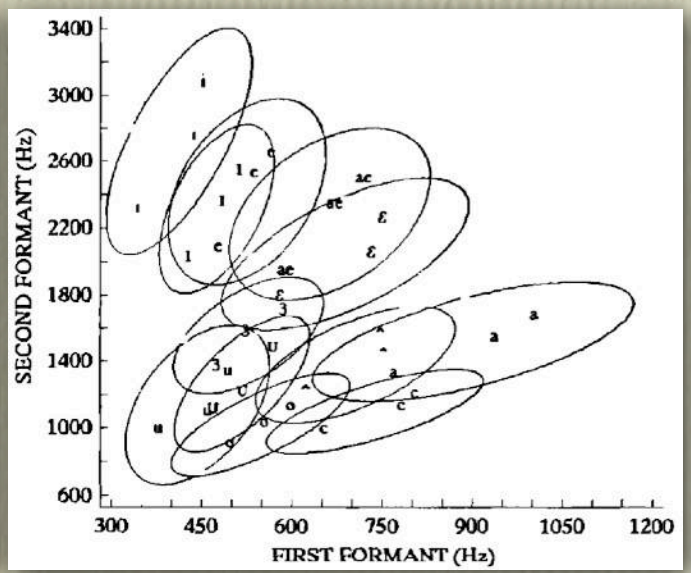


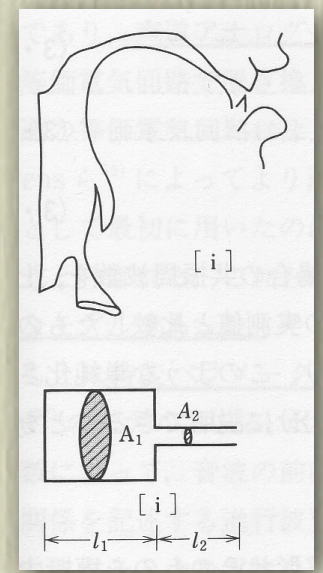
Function of the voice timbre

What is the original function of the voice timbre?

- For apes
  - The voice timbre is an acoustic correlate with the identity of apes.
- For speech scientists and engineers
  - They had started research by correlating the voice timbre with messages conveyed by speech stream such as words and phonemes.
    - Formant frequencies are treated as acoustic correlates with vowels.
  - “Speech recognition” started first, then, “speaker recognition” followed.







$$f_n = \frac{c}{2l_1} n$$

$$f_n = \frac{c}{2l_2} n$$

$$f = \frac{c}{2\pi} \left[\frac{A_2}{A_1 l_1 l_2} \right]^{1/2}$$

Function of the voice timbre

What is the original function of the voice timbre?

- For apes
  - The voice timbre is an acoustic correlate with the identity of apes.
- For speech scientists and engineers
  - They had started research by correlating the voice timbre with messages conveyed by speech stream such as words and phonemes.
    - Formant frequencies are treated as acoustic correlates with vowels.
  - "Speech recognition" started first, then, "speaker recognition" followed.

But the voice timbre can be changed easily.

- Speaker-independent acoustic model for word recognition
  - $P(o|w) = \sum_s P(o, s|w) = \sum_s P(o|w, s)P(s|w) \sim \sum_s P(o|w, s)P(s)$
- Speaker-adaptive acoustic model for word recognition
  - HMMs, even DNNs, are always modified and adapted to users.
- These methods don't remove "speaker" components in speech.

Menu of the last four lectures

Robust processing of easily changeable stimuli

- Robust processing of general sensory stimuli
- Any difference in the processing between humans and animals?

Human development of spoken language

- Infants' vocal imitation of their parents' utterances
- What acoustic aspect of the parents' voices do they imitate?

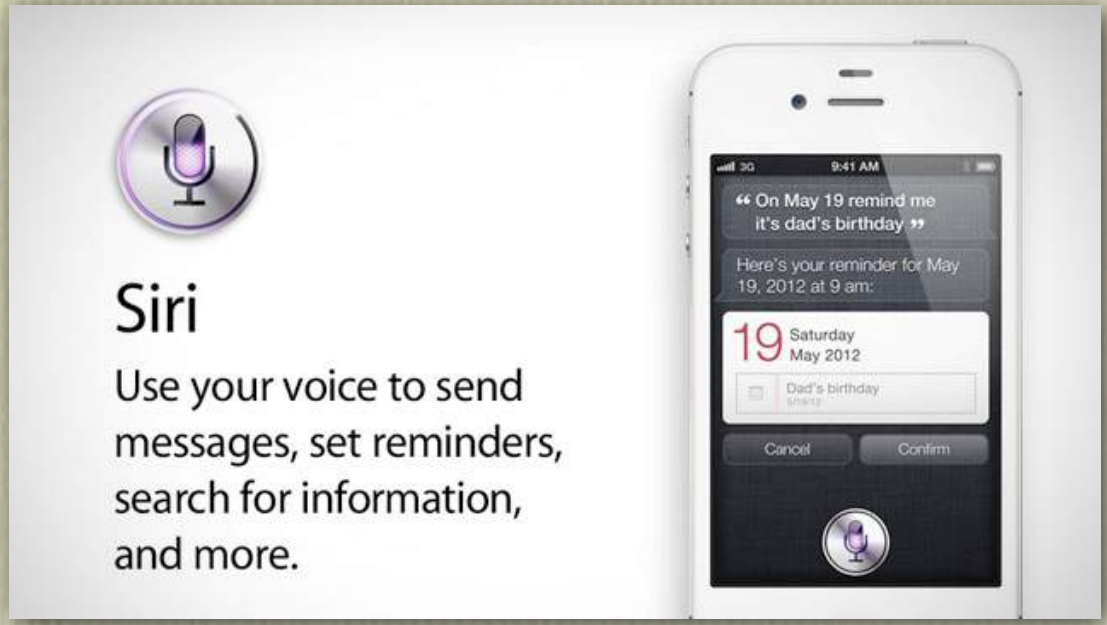
Speaker-invariant holistic pattern in an utterance

- Completely transform-invariant features -- *f*-divergence --
- Implementation of word Gestalt as relative timbre perception
- Application of speech structure to robust speech processing

Radical but interesting discussion

- A hypothesis on the origin and emergence of language
- What is your definition of "human-like" machines?

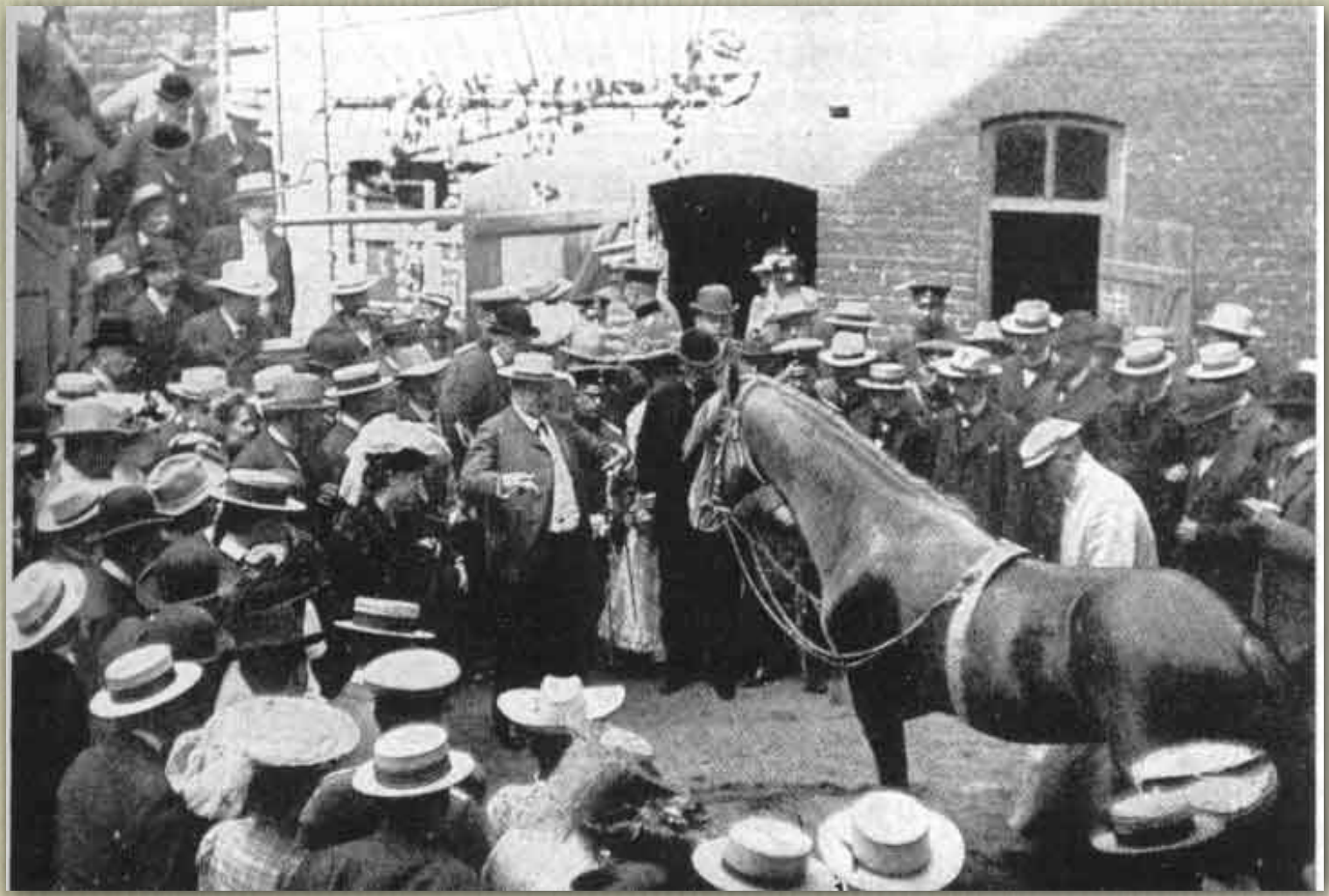
What is the goal of speech engineering?



Clever Hans

A horse who can "calculate"

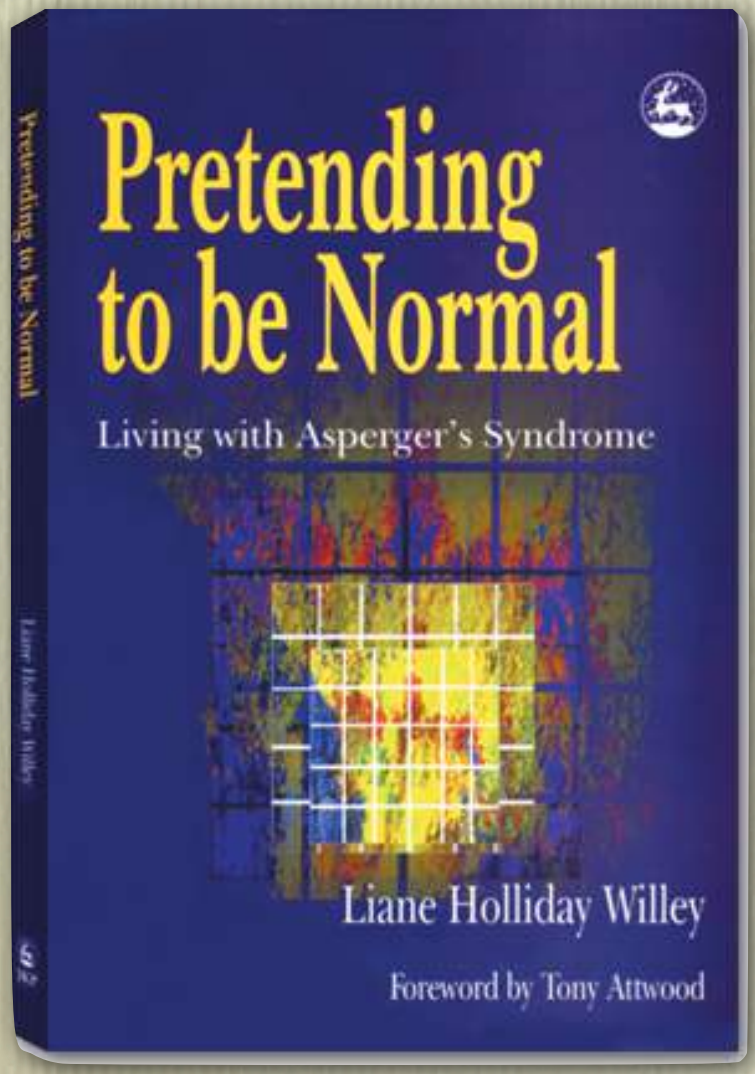
- https://simple.wikipedia.org/wiki/Clever_Hans
- Can he calculate or can he pretend to calculate?

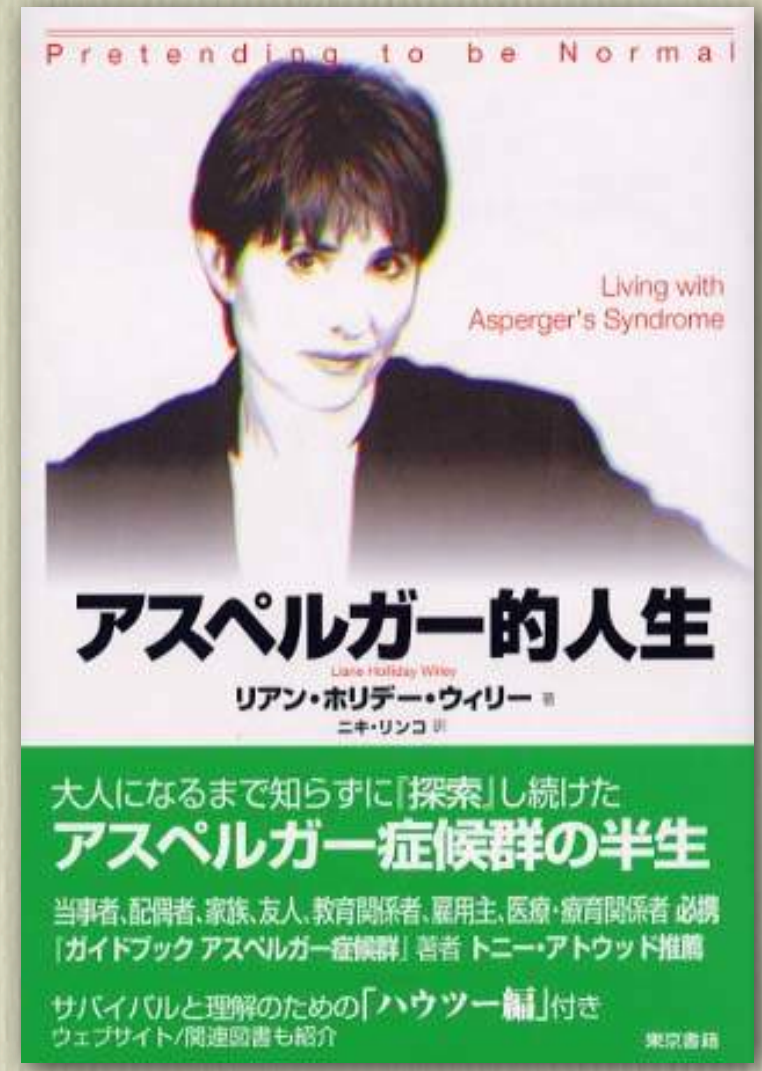


“Pretending to be normal”

A book written by Liane Holliday Willey

- She is autistic (Asperger’s syndrome).





Your definition of human-like machine?

- **Necessary conditions**
- **Sufficient conditions**
- **Necessary and sufficient conditions**
- **What can researchers do?**
  - Different researchers may claim different "necessary" conditions.
  - What a researcher can do is just to satisfy his/her own "necessary" conditions to make his/her own human-like machines.

Final assignment

- **1. Find a paper which discusses human-like machines.**
  - Summarize the paper and make comments on the paper. (a few A4 pages)
  - Some researchers are discussing human-likeness of behaviors of ChatGPT.
- **2. Describe your *own* necessary conditions of human-like machines.**
  - Your unique definition is welcome. (one or a few A4 pages)
- **3. Comment on the content of this class.**
  - Just one paragraph is OK.

- **Submission**
  - Your report should be submitted via. ITC-LMS.
  - Your files should have filenames as
    - [student_id]_paper.pdf and [student_id]_[your name].pdf
    - Ex. 36-302439_paper.pdf and 36_302439_nobuaki_minematsu.pdf

- Deadline = 23:59:59 on Jan 30 (Tue). You have 2 weeks to go.