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 the definition of "human-like" robots?

A difference bet. machines and humans

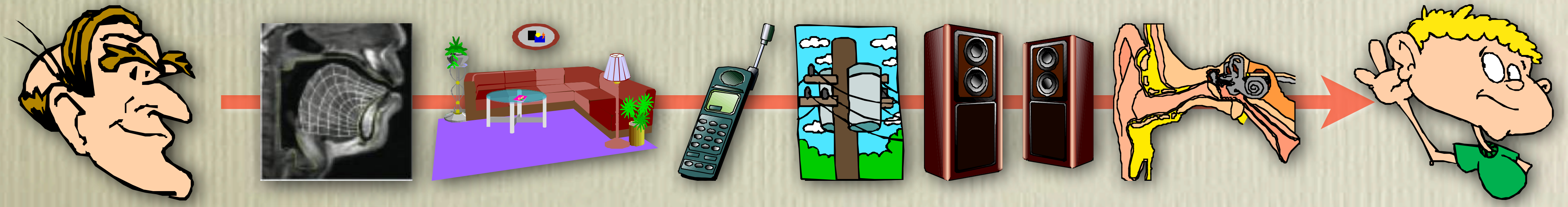
Machine strategy (engineers' strategy): ASR

- Collecting a huge amount of speaker-balanced data
  - Statistical training of acoustic models of individual phonemes (allophones)
- Adaptation of the models to new environments and speakers
  - Acoustic mismatch bet. training and testing conditions must be reduced.

?

Human strategy: HSR

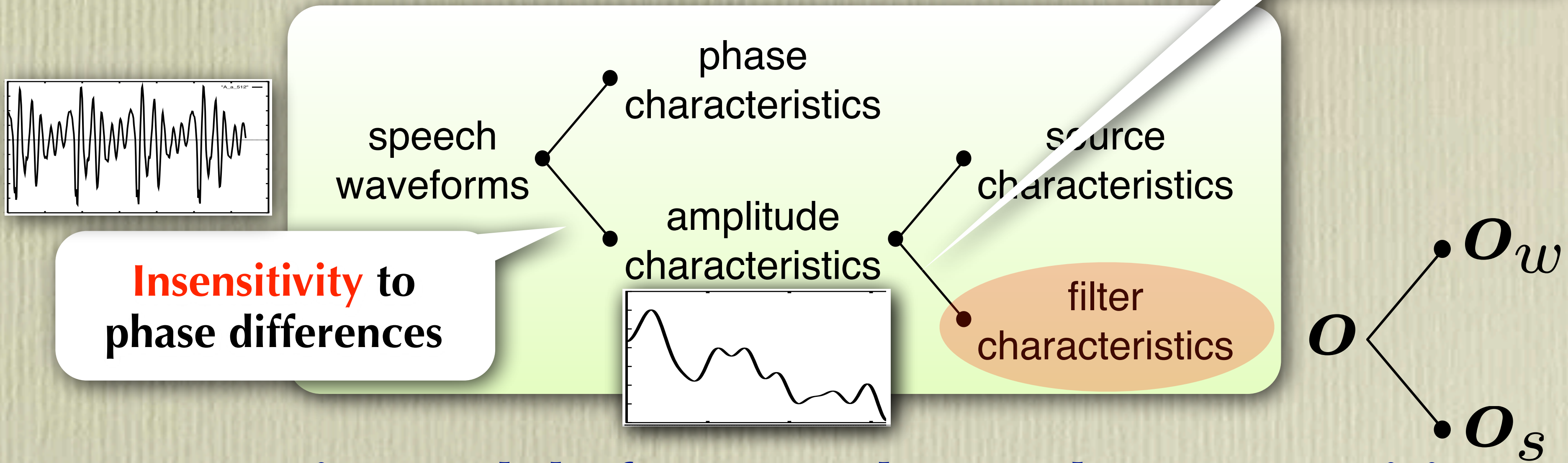
- A major part of the utterances an infant hears are from its parents.
  - The utterances one can hear are extremely speaker-biased.
- Infants don't care about the mismatch in lang. acquisition.
  - Their vocal imitation is not acoustic, it is not impersonation!!



Feature separation to find specific info.

- ## De facto standard acoustic analysis of s

Insensitivity to pitch differences

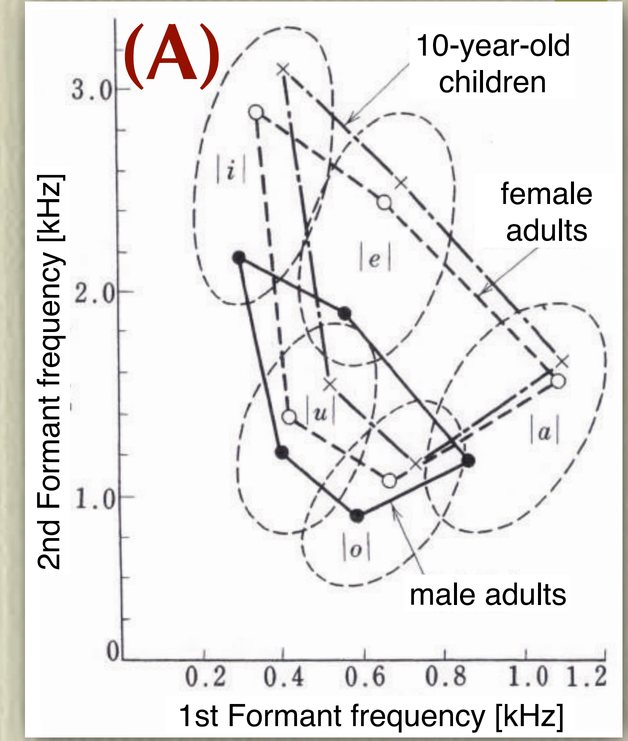


Insensitivity to phase differences

- ## 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?

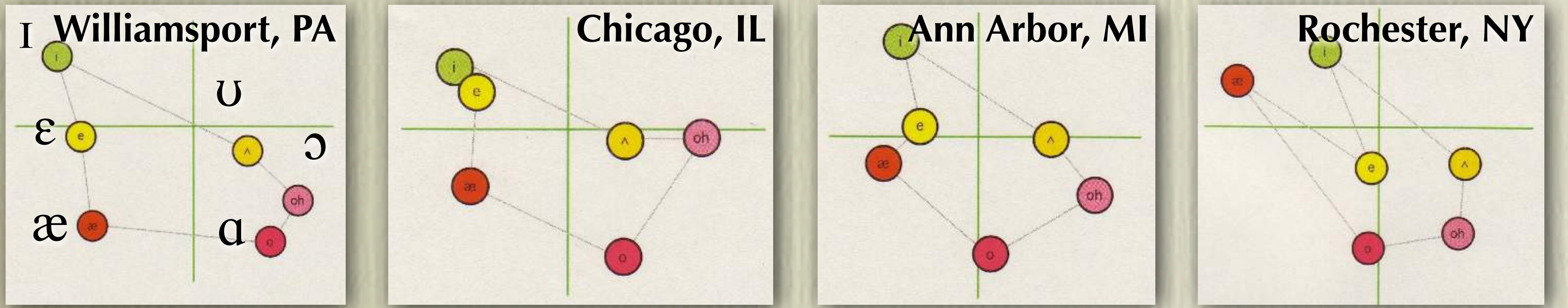
Insensitivity and sensitivity

- **Infants' vocal learning is**
  - insensitive to age and gender differences. **(A)**
  - sensitive to accent differences. **(B)**
- **Infants' vocal learning seems to be**
  - insensitive to feature instances and sensitive to feature relations.
    - **(A)** = instances and **(B)** = relations.
  - Relations, i.e., shape of distribution can be represented geometrically as distance matrix.



formant frequencies of adults and children

(B)

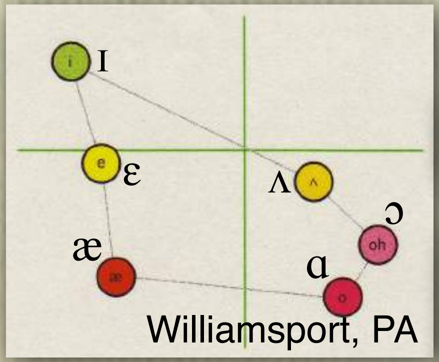


Distribution of normalized formants among AE dialects [Labov et al.'05]

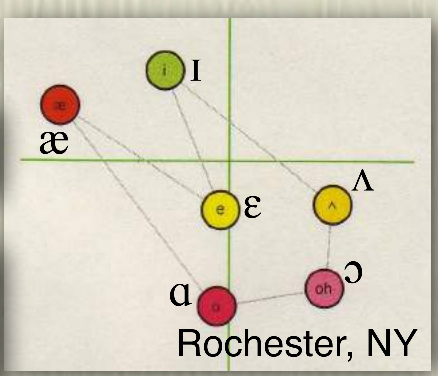
“Separately brought up identical twins”

- **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

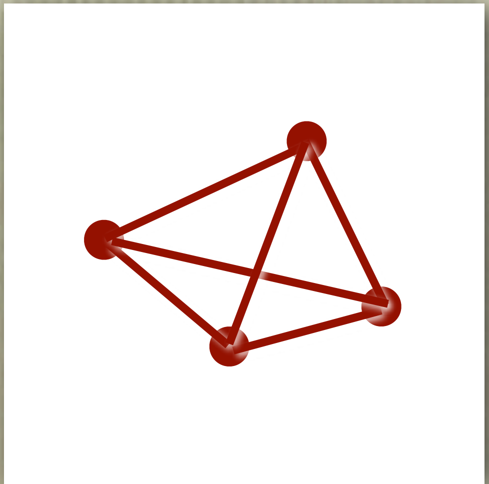


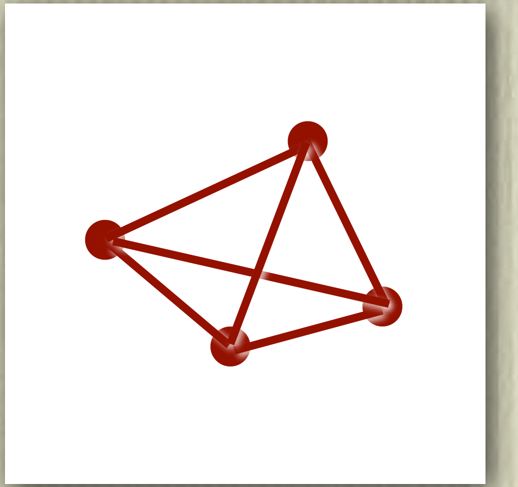
Machines that don’t learn what infants don’t learn.

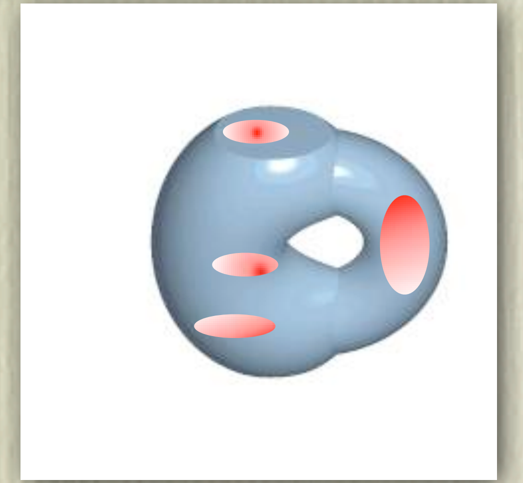
Invariance in variability

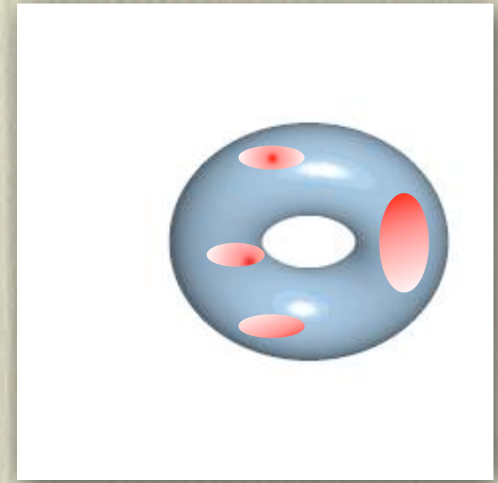
Topological invariance [Minematsu'09]

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





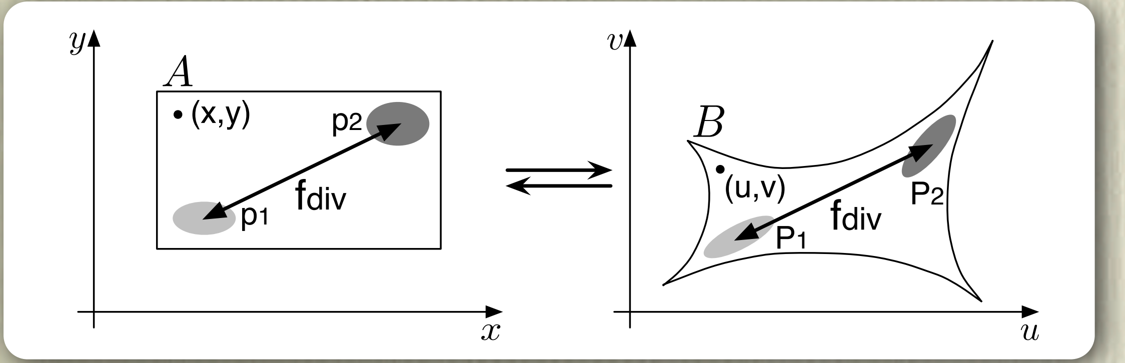




Complete transform-invariance

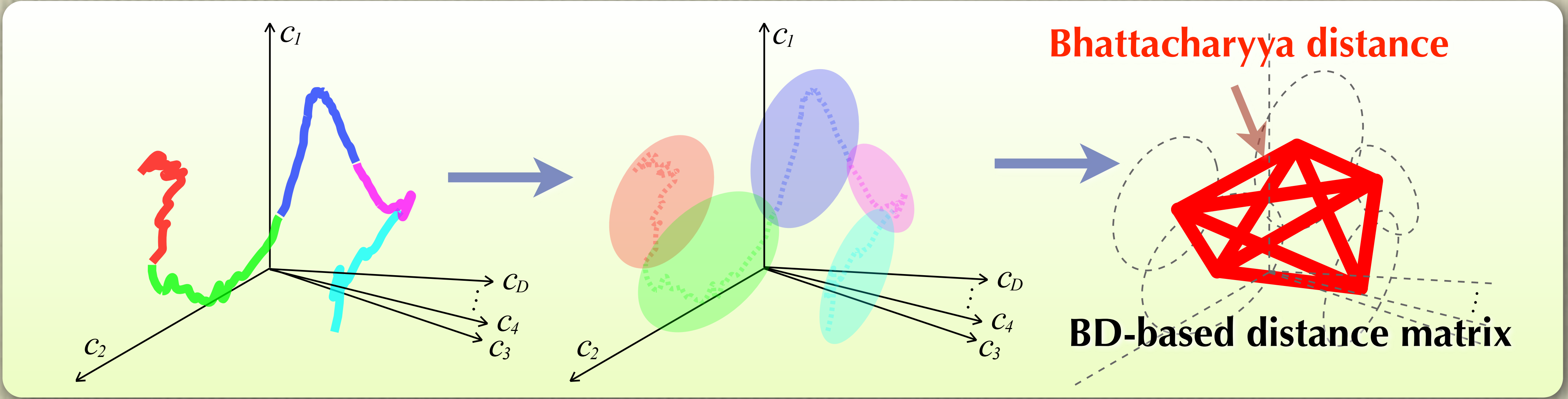
Any general expression for invariance?[Qiao'10]

- BD is just one example of invariant contrasts.
- f-divergence is invariant with any kind of transformation.
  - $f_{div}(p_1, p_2) = \int p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}$
  - $g(t) = t\log(t) \to f_{div} = \mathrm{KL} - \mathrm{div.}$ $\qquad g(t) = \sqrt{t} \to -\log(f_{div}) = \mathrm{BD}$
  - $f_{div}(p_1, p_2) = f_{div}(P_1, P_2)$
- 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. $M = p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right)$



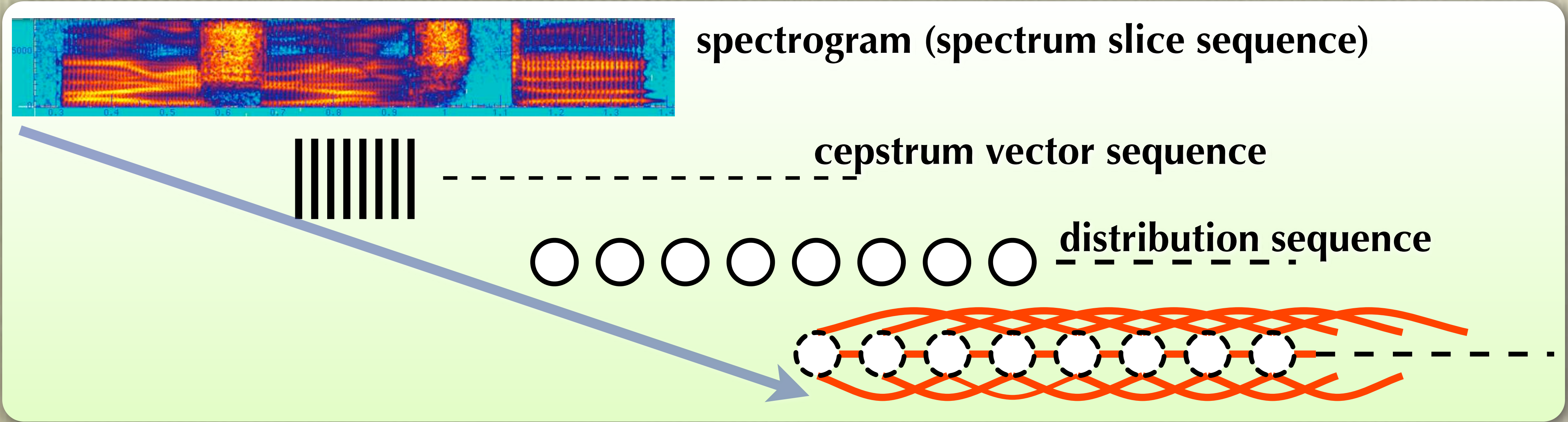
Invariant speech structure

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



Bhattacharyya distance

BD-based distance matrix



spectrogram (spectrum slice sequence)

cepstrum vector sequence

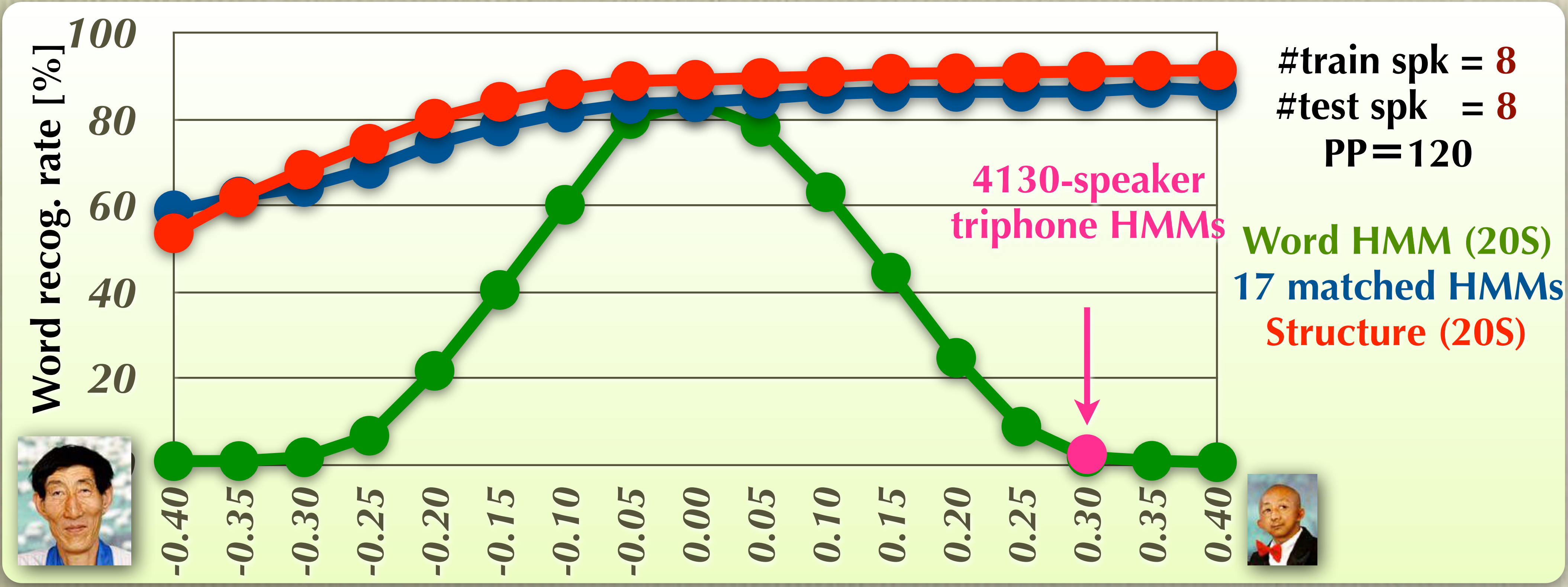
distribution sequence

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

Application of structures to ASR

Isolated word recognition using warped utterances

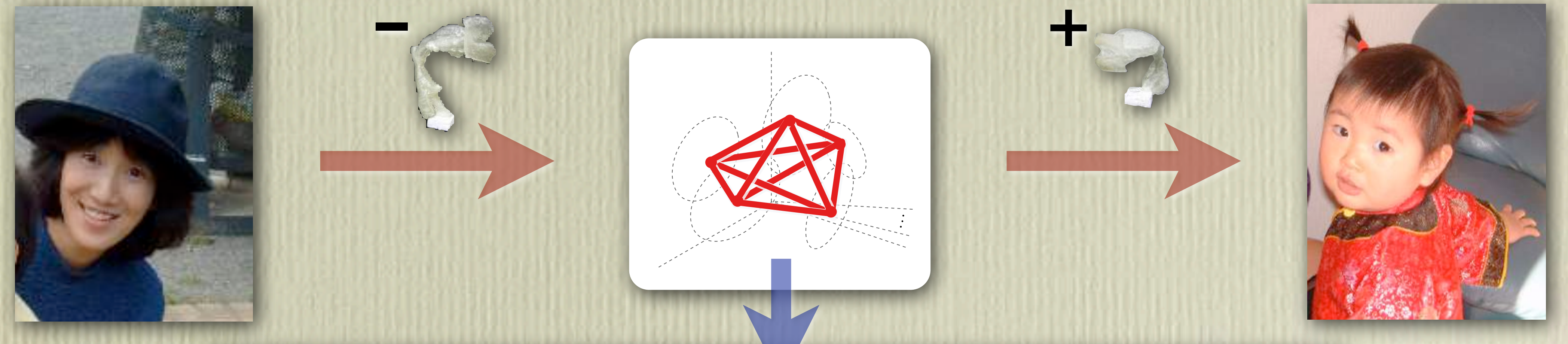
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used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

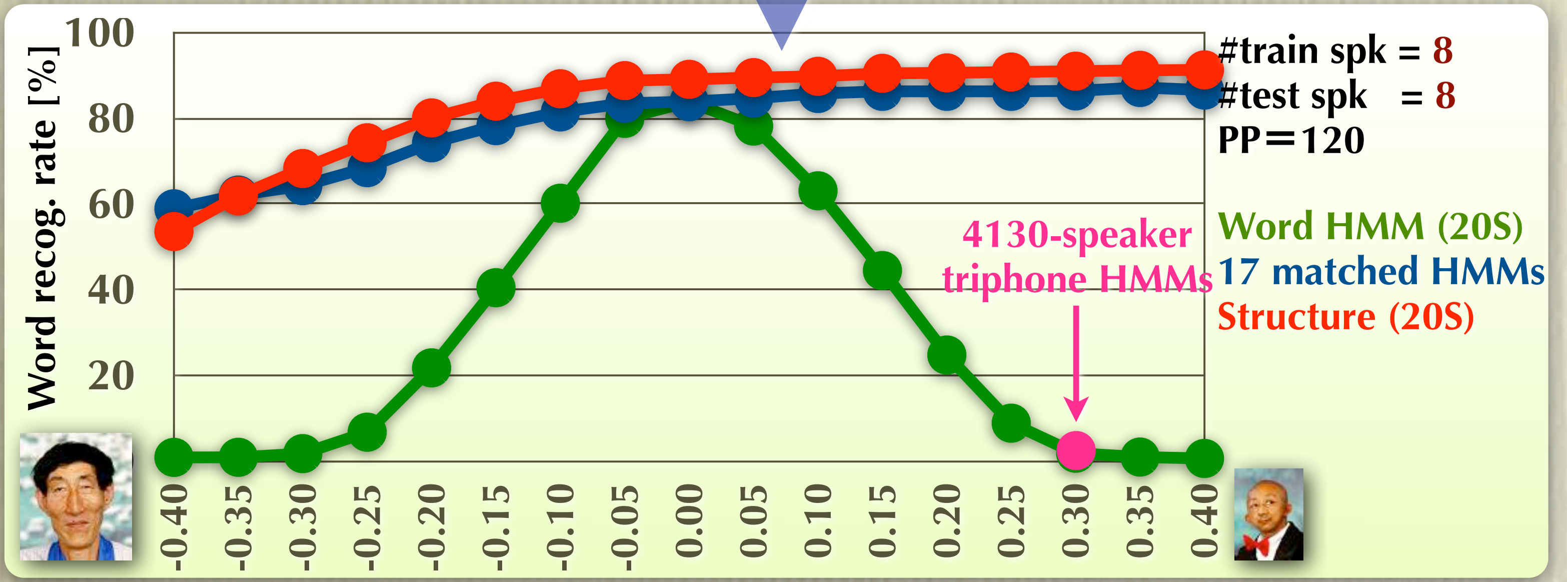


An experiment with real vocal imitation

Demonstration with my wife and daughter

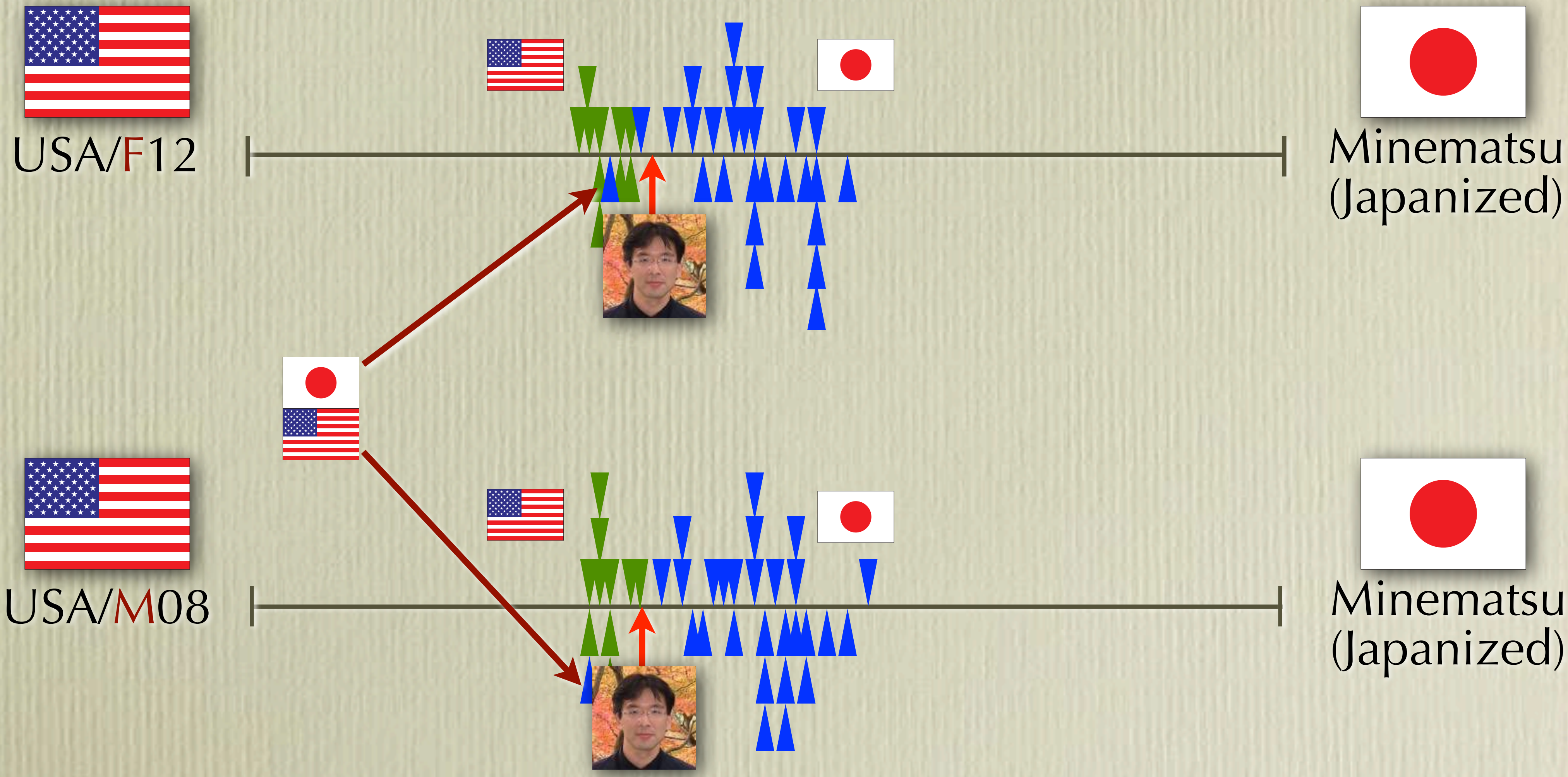
- Constraint conditions are given by my wife.
- Initial conditions are given by my daughter.





A big solution for CALL development

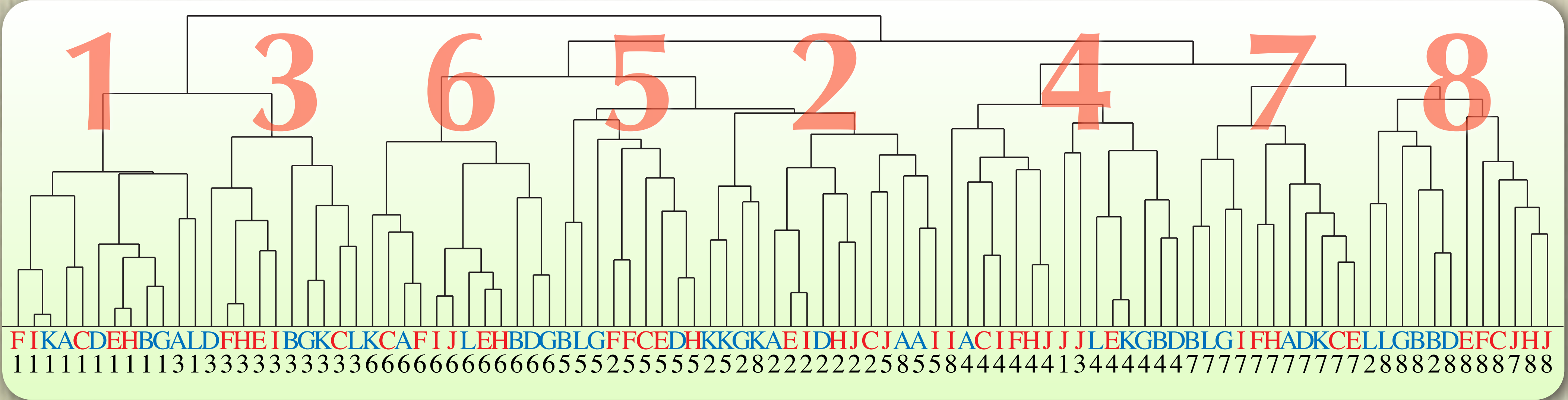
Proficiency estimation based on structural distance



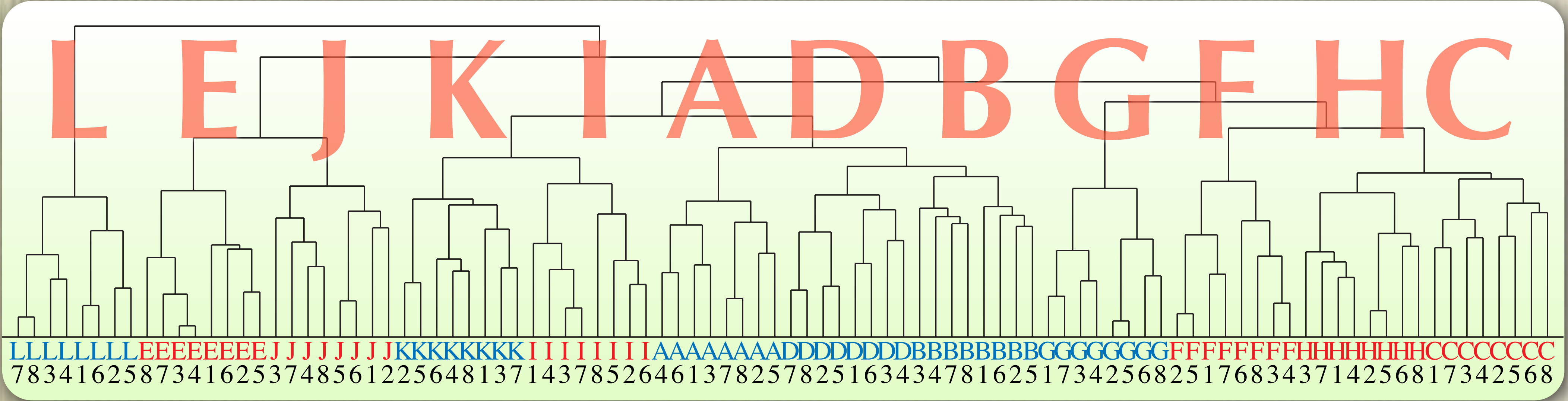
(Minematsu@ICSLP 2004)

Clustering of learners

Contrast-based comparison



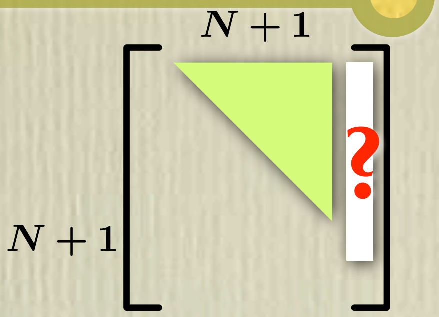
Substance-based comparison



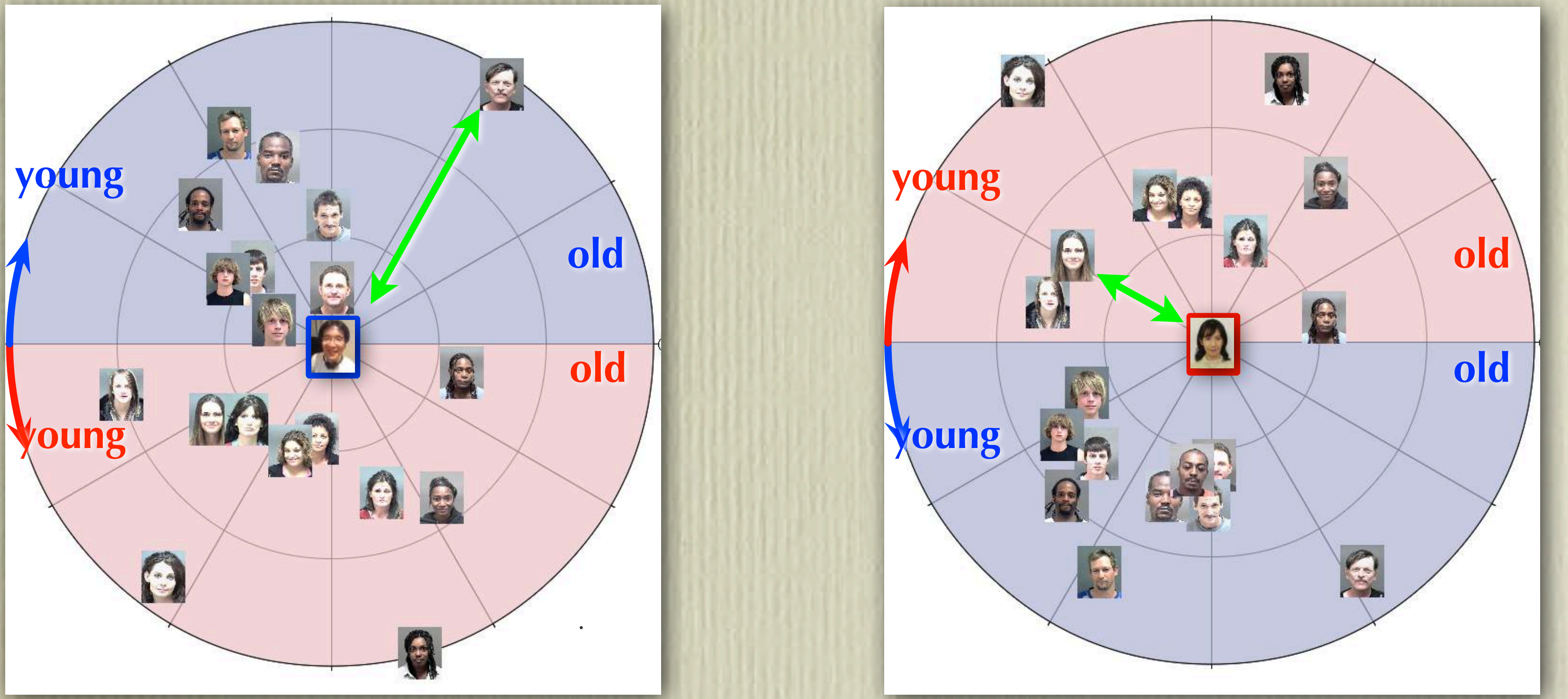
Application of speaker-pair-open prediction

TED talks browser from your viewpoint

- If TED talkers provide their SAA readings....
- If these readings are transcribed by phoneticians....



Visualization of pronunciation diversity [Kawase *et al.*,'14]



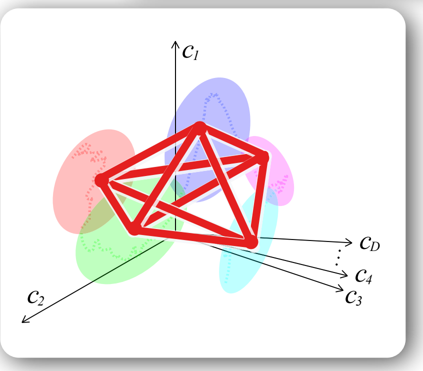
Y. Kawase, et al., "Visualization of pronunciation diversity of World Englishes from a speaker's self-centered viewpoint"

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

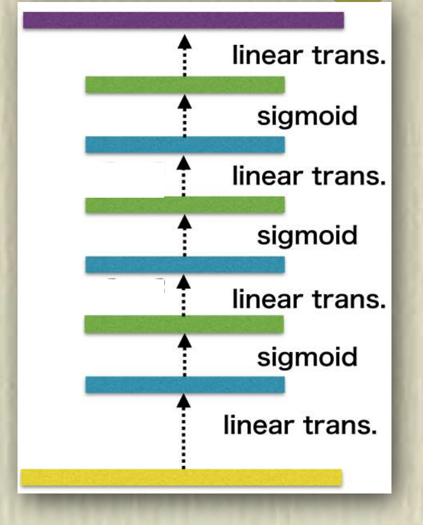


Menu of the last four lectures

- **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 the definition of "human-like" robots?

DNN and speech structure

- **Deep neural network** [Hinton+'06, '12]
  - Deeply stacked artificial neural networks
  - Results in a huge number of weights
  - Unsupervised pre-training and supervised fine-tuning

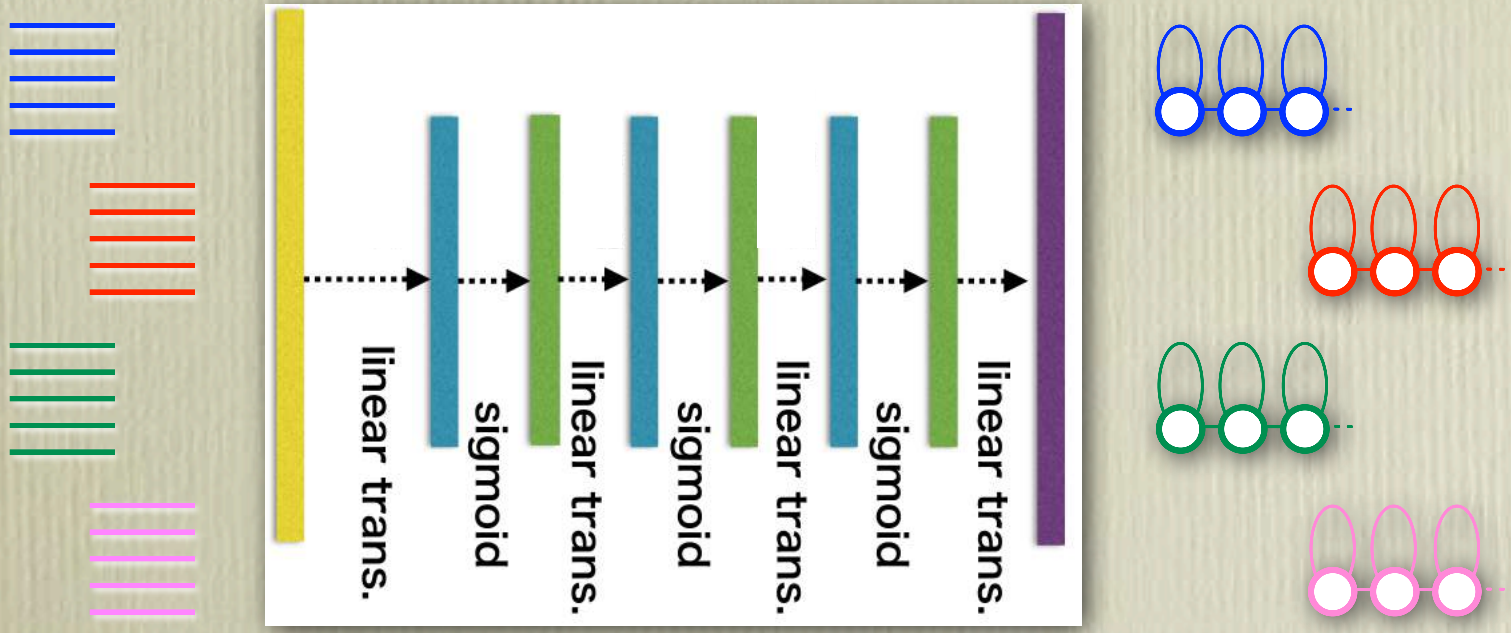


- **Findings in DNN-based ASR** [Mohamed+'12]
  - First several layers seem to work as extractor of invariant features or speaker-normalized features.
  - Still difficult to interpret structure and weights of DNN physically.
    - Interpretable DNNs are becoming one of the hot topics [Sim'15].
- **A simple question asked in tutorial talks of DNN**
  - "What are *really* speaker-independent features?"
    - Asked by N. Morgan at APSIPA2013 and ASRU2013
- **Some similarities between DNN and speech structure?**

DNN as posterior estimator

General framework for training DNN

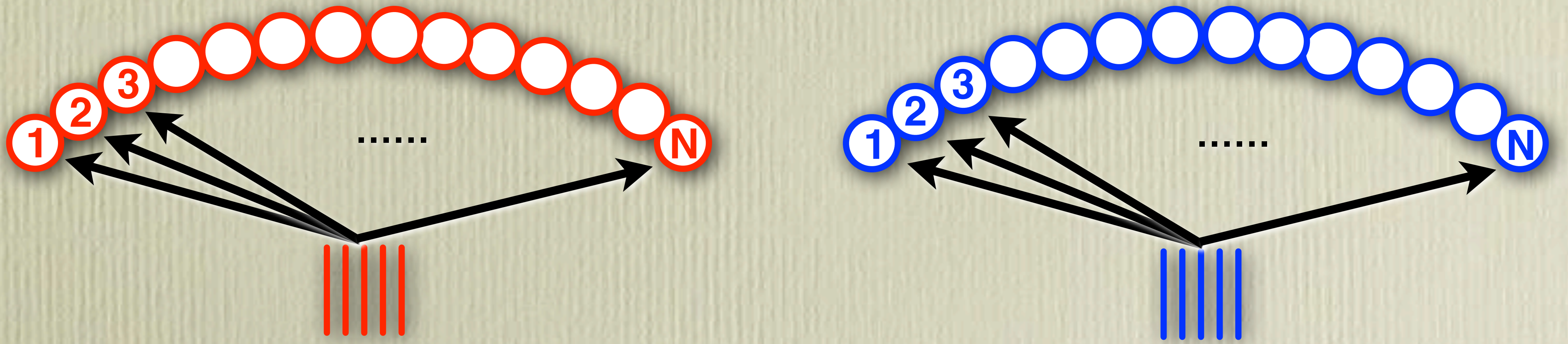
- Unsupervised pre-training and supervised training
  - In the latter training, speaker-adapted HMMs are used to prepare posteriors (=labels) for each frame of the training data.
- DNN is trained so that it can extract speaker-invariant features and estimate posteriors in a speaker-independent way.
- Output of DNN = posteriors (phoneme state posteriors in ASR)



Posteriors = normalized similarities

Posteriors of { $P(c_i|o)$ }

- $P(c_i|o) \propto P(o|c_i)P(c_i)$
- $\sum_i P(c_i|o) = 1.0$
- Can be interpreted as normalized similarity scores biased by priors.
- Output of DNN = normalized similarity scores to a definite set of speaker-adapted acoustic “anchors” of { c_i }.

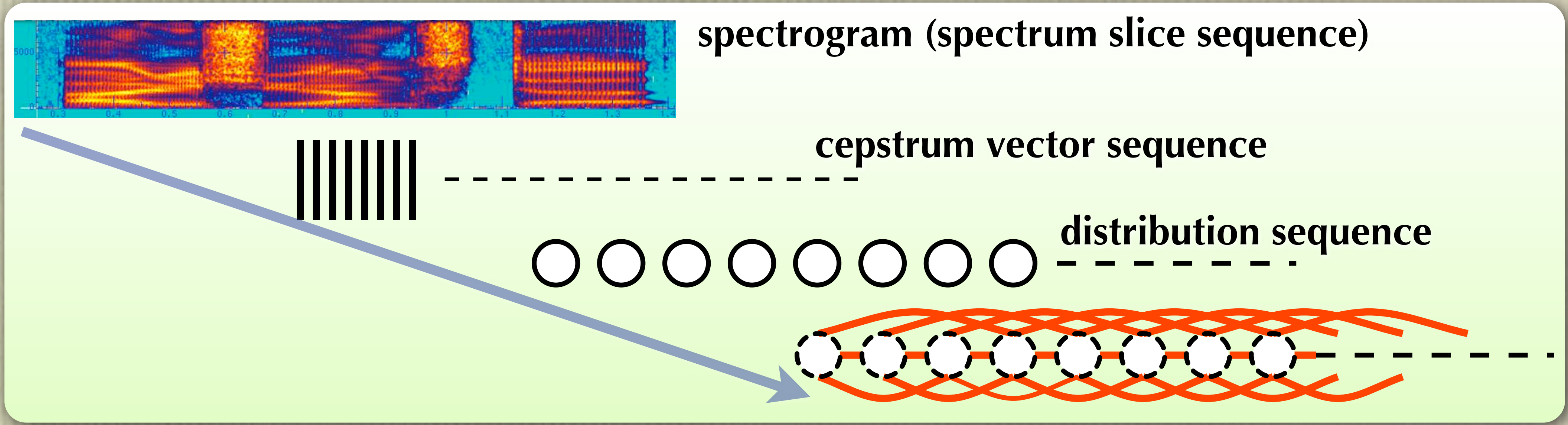


🟥 🟦 : speaker-dependent ⬛ : speaker-independent(invariant)

- Similarities scores can be converted to **distances to “anchors”**.
  - Either of similarity matrix or distance matrix is used for clustering.

Distances to anchors

Speech structure extracted from an utterance

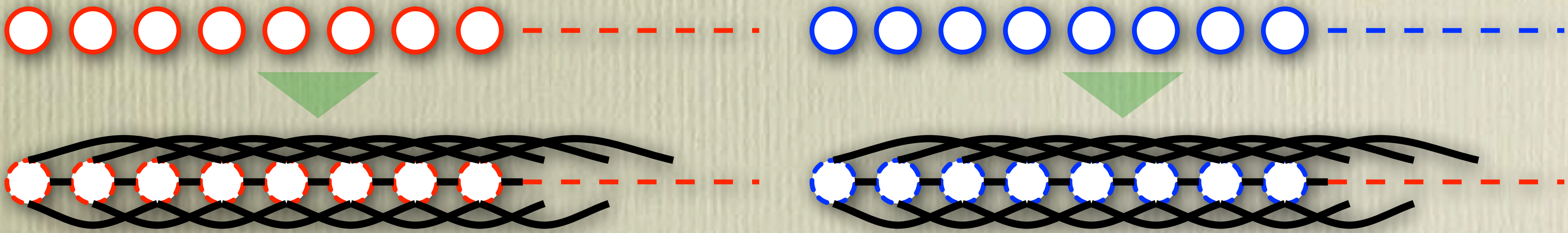


spectrogram (spectrum slice sequence)

cepstrum vector sequence

distribution sequence

Structure extraction for speakers 🟥 and 🟦



🟥 🟦 : speaker-dependent

⬛ : speaker-independent(invariant)

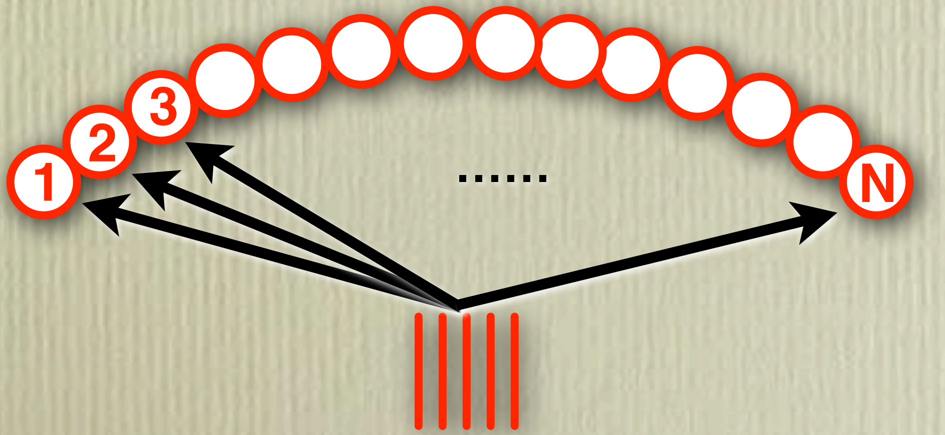
Invariant contrasts

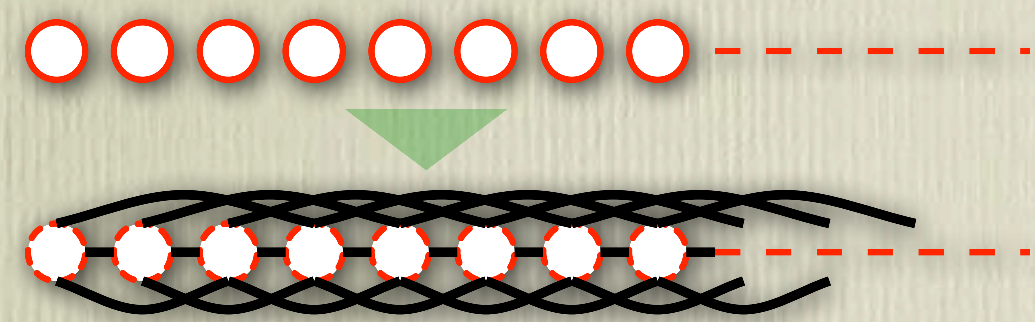
DNN as speaker-invariant contrast estimation

- Use of spk-dependent HMMs to prepare posterior labels
  - “Anchors” have to be given from researchers.
- A huge amount of data to guarantee spk-invariance of DNN

Str. extraction as speaker-invariant contrast detection

- Use of within-utterance acoustic events only
  - “Anchors” exist in a given utterance.
- Spk-invariance is guaranteed by invariant properties of f-div.





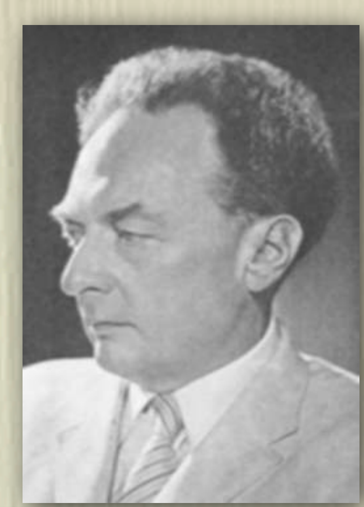
A claim found in classical linguistics

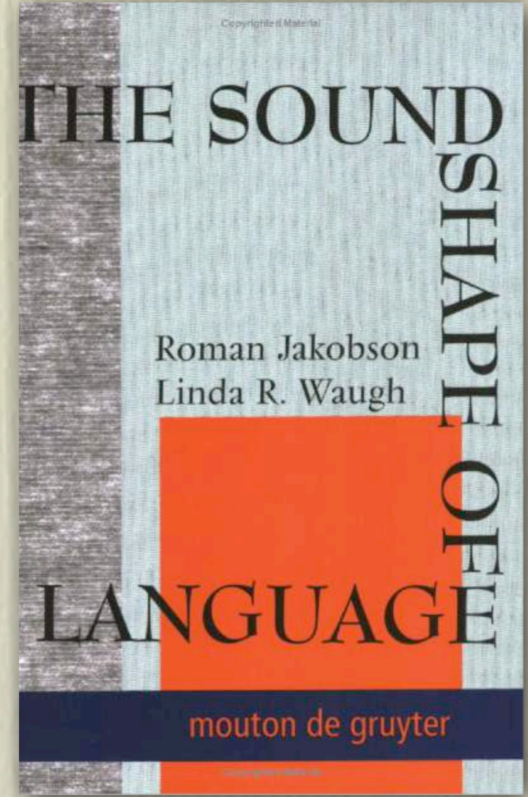
Theory of relational invariance [Jakobson+'79]

- Also known as theory of distinctive features
- Proposed by R. Jakobson

> We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables.

> Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. in their relations to the other sounds.



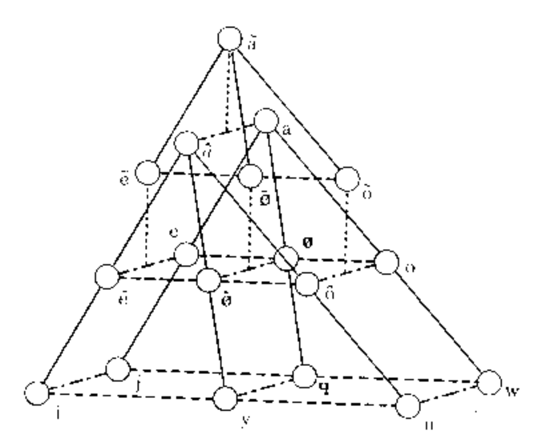


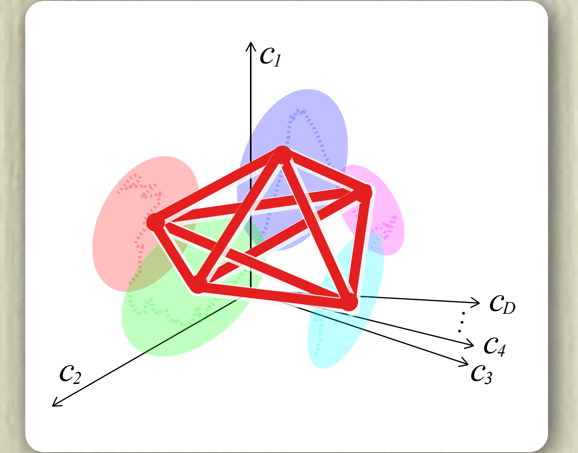
Copyrighted Material

THE SOUND SHAPE OF LANGUAGE

Roman Jakobson
Linda R. Waugh

mouton de gruyter

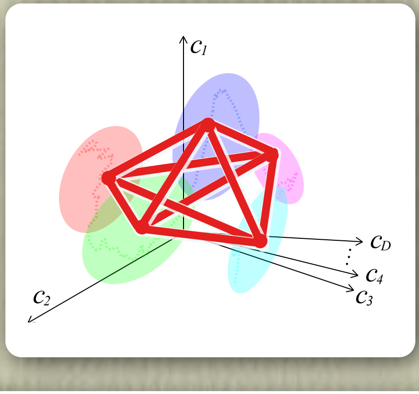


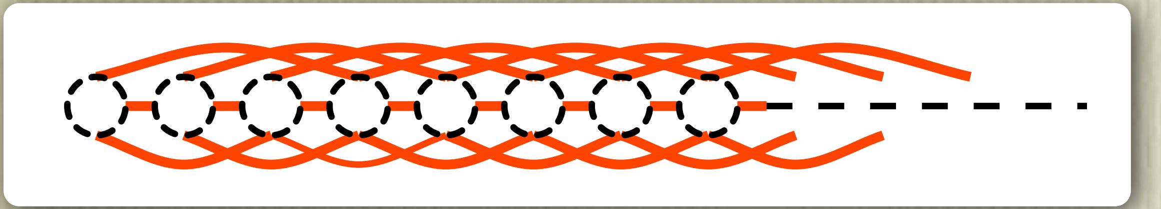


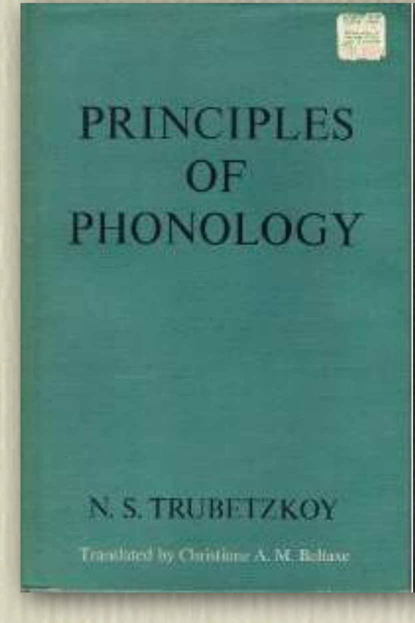
More classical claims in linguistics

Nikolay Sergeevich Trubetskoy (1890-1938)

- "The Principles of Phonology" (1939)
- The phonemes should not be considered as building blocks out of which individual words are assembled. Each word is a phonic entity, a Gestalt, and is also recognized as such by the hearer.
- As a Gestalt, each word contains something more than sum of its constituents (phonemes), namely, the principle of unity holds the phoneme sequence together and lends individuality to a word.






PRINCIPLES
OF
PHONOLOGY

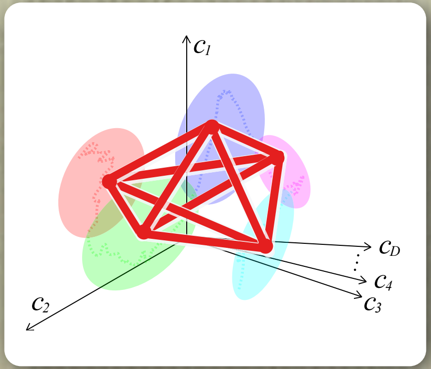
N. S. TRUBETZKOY

Translated by Christiane A. M. Baltaxe

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.



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

McGraw-Hill Paperbacks

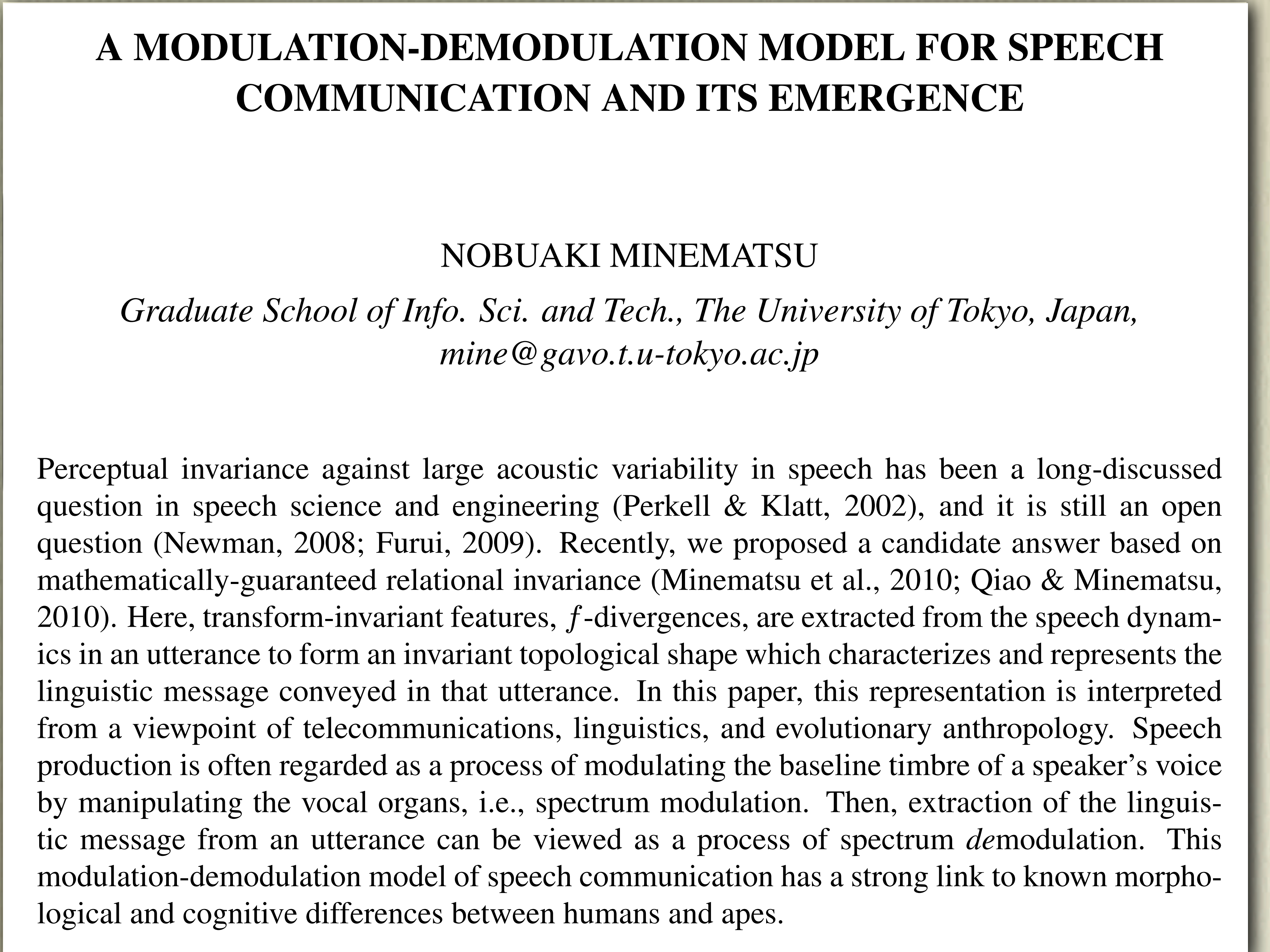
Course in General Linguistics

Ferdinand de Saussure

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 the definition of "human-like" robots?

Origin and evolution of language



A MODULATION-DEMODULATION MODEL FOR SPEECH COMMUNICATION AND ITS EMERGENCE

NOBUAKI MINEMATSU

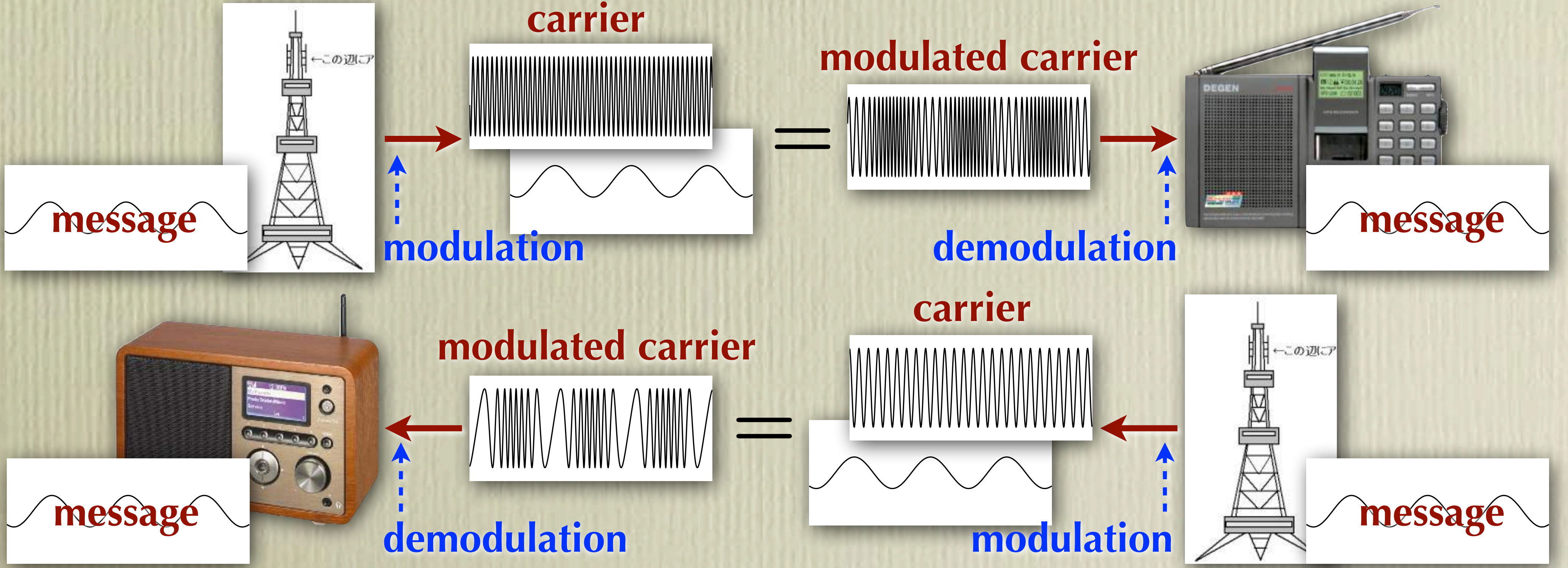
Graduate School of Info. Sci. and Tech., The University of Tokyo, Japan,
mine@gavo.t.u-tokyo.ac.jp

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.



carrier

message

modulation

=

modulated carrier

demodulation

message

carrier

modulated carrier

=

message

demodulation

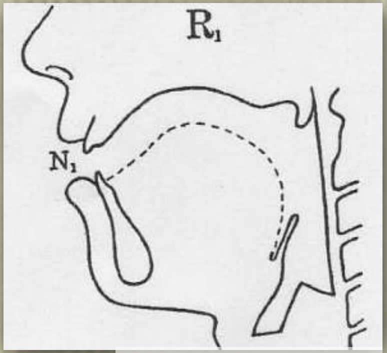
modulation

message

A way of characterizing speech production

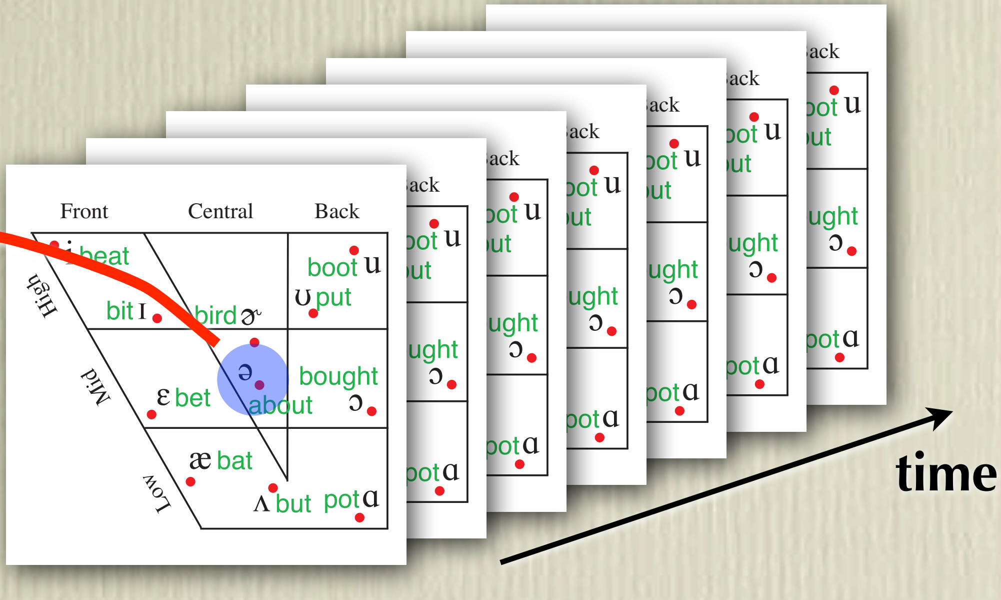
Speech production as spectrum modulation

- 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

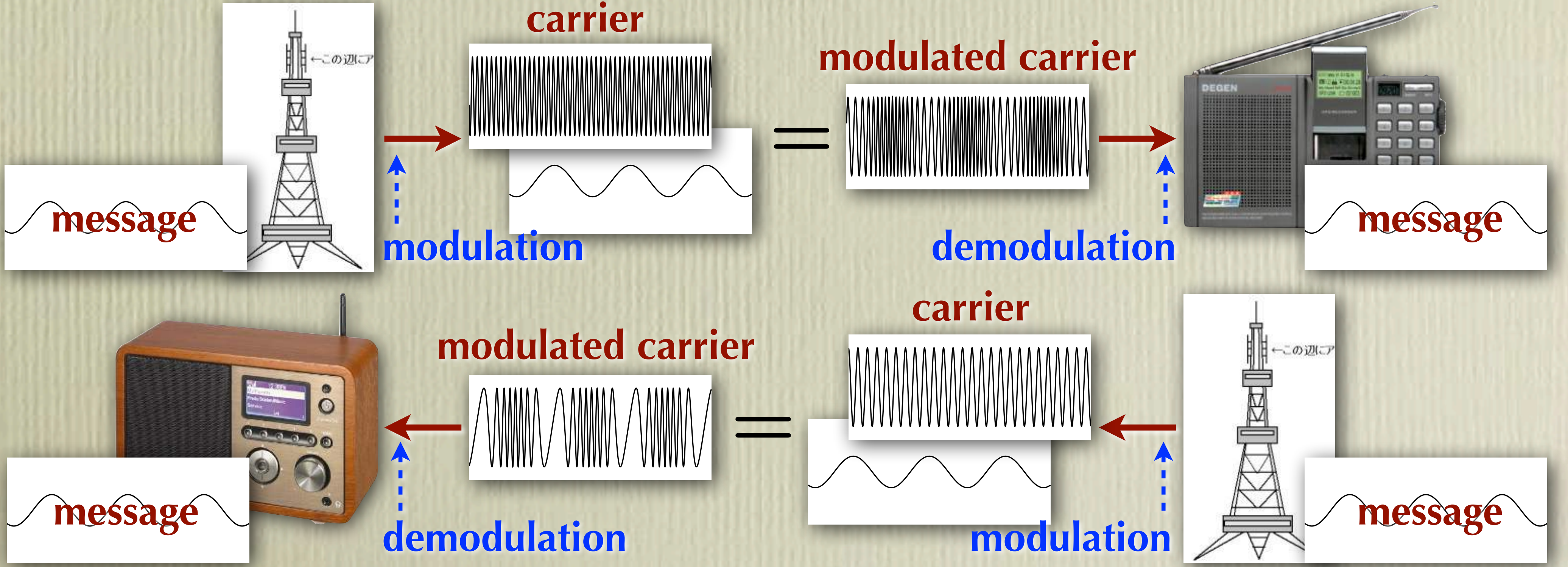
Schwa
= most lax
= most frequent
= home position
= spk.-specific baseline timbre



Demodulation used in telecommunication

Demodulation in frequency, amplitude, and phase

- Demodulation = a process of extracting a 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)

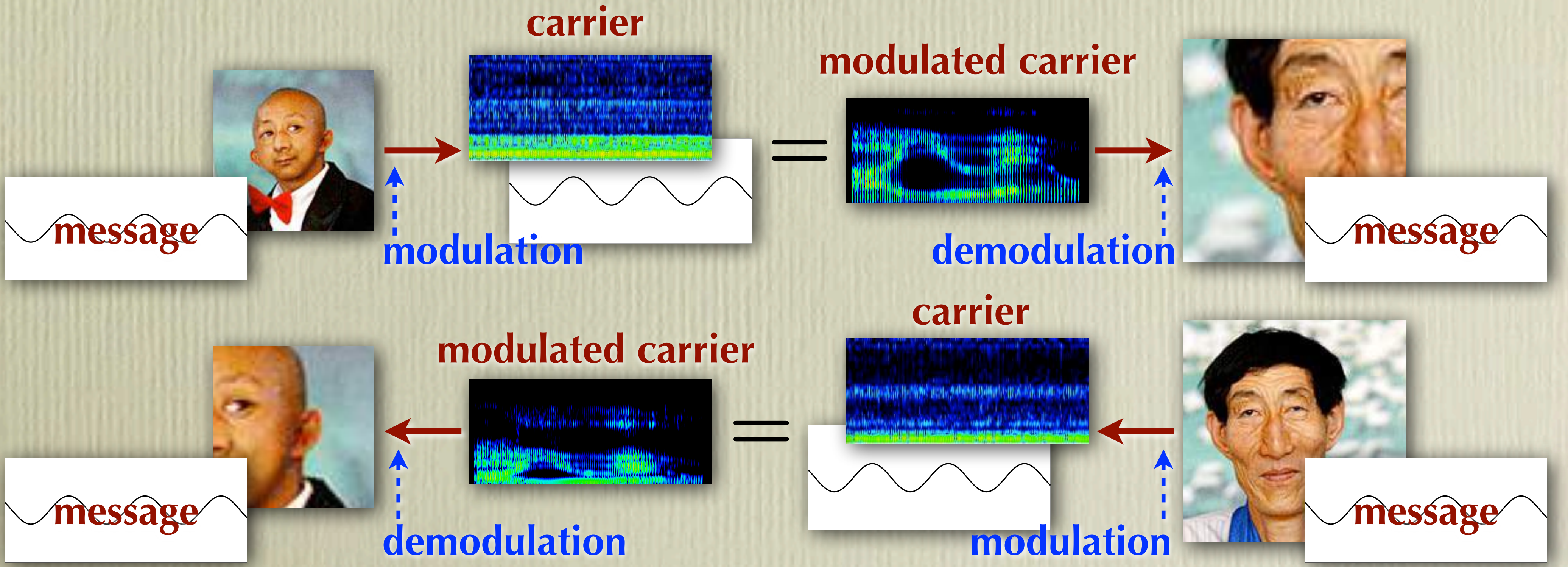


Spectrum demodulation

© AP

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)


carrier

modulated carrier

message

modulation

=

demodulation

message

carrier

modulated carrier

message

demodulation

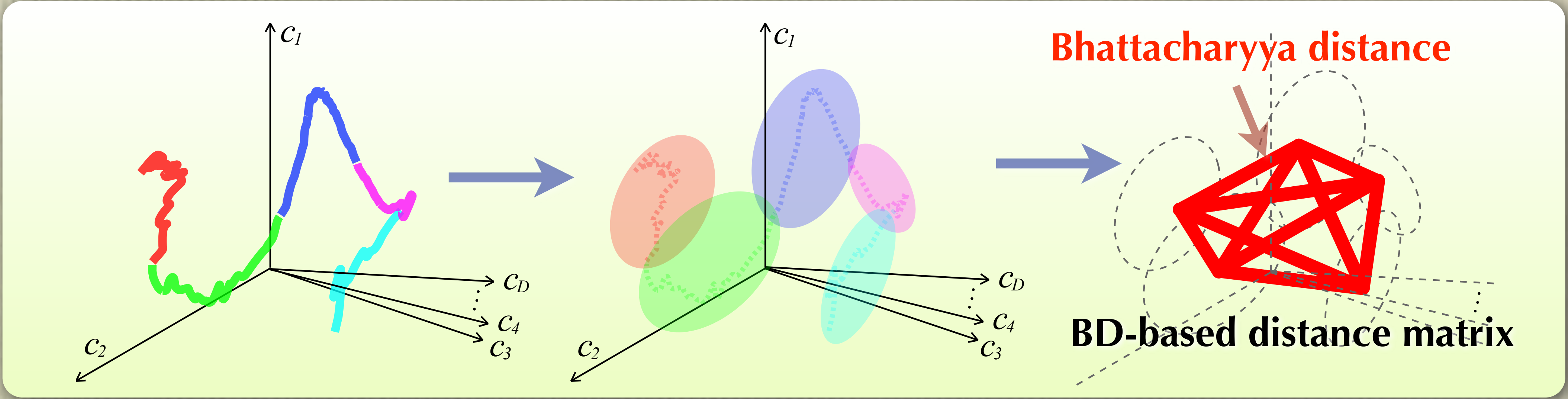
=

modulation

message

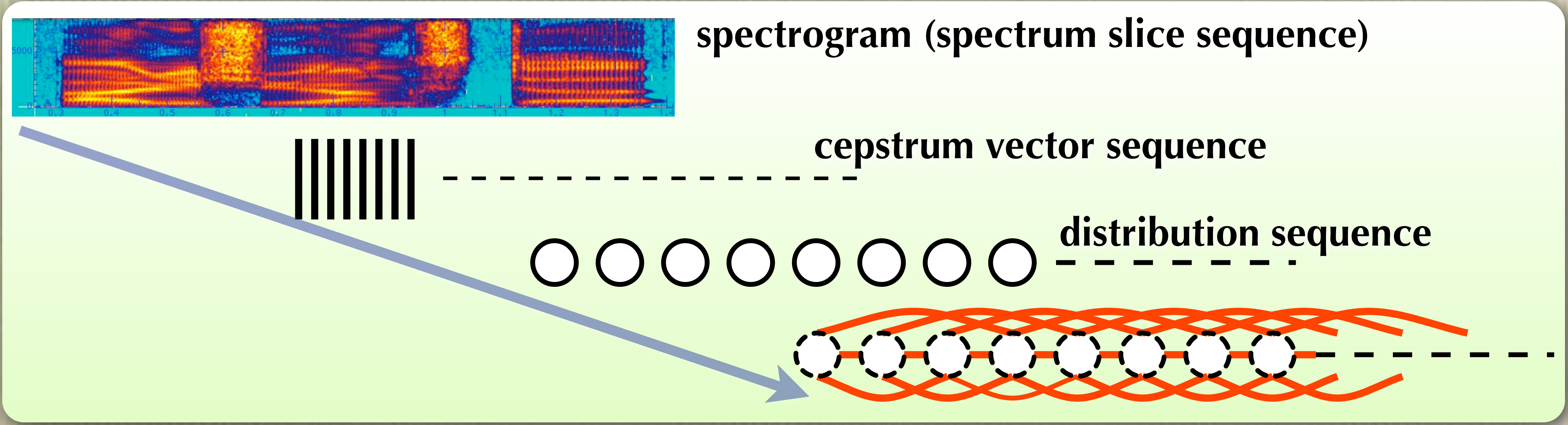
Invariant speech structure

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



Bhattacharyya distance

BD-based distance matrix



spectrogram (spectrum slice sequence)

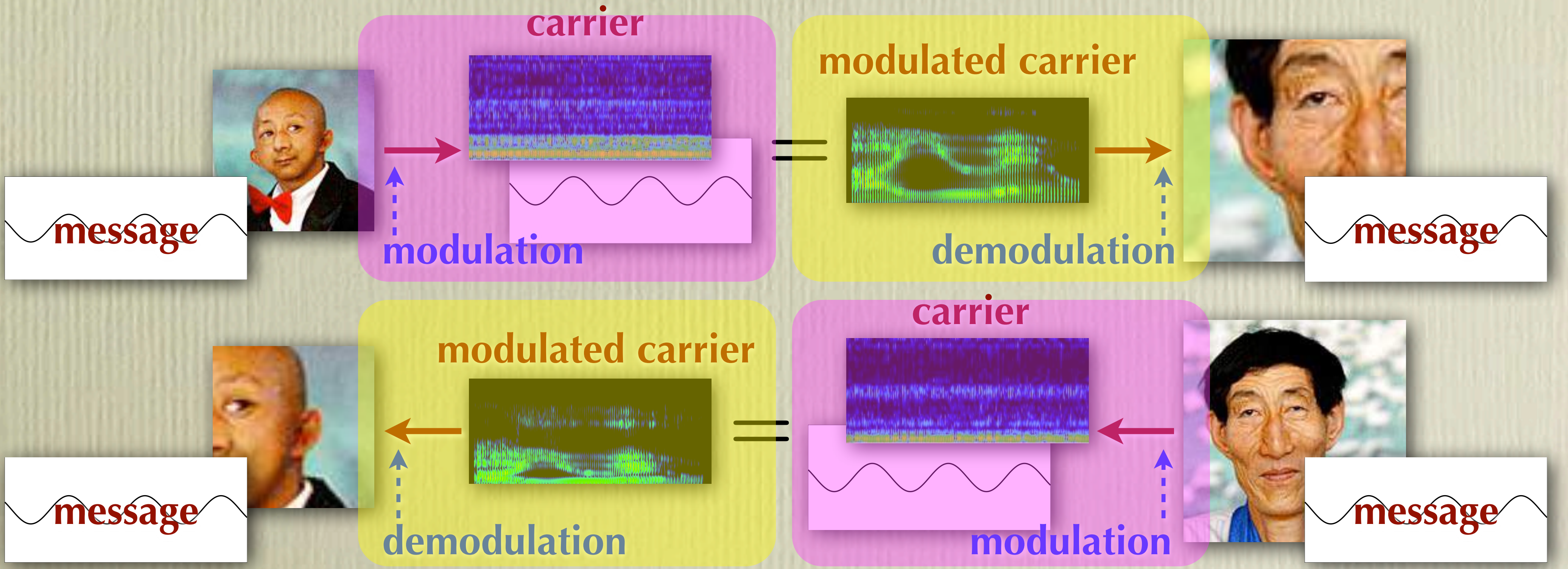
cepstrum vector sequence

distribution sequence

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

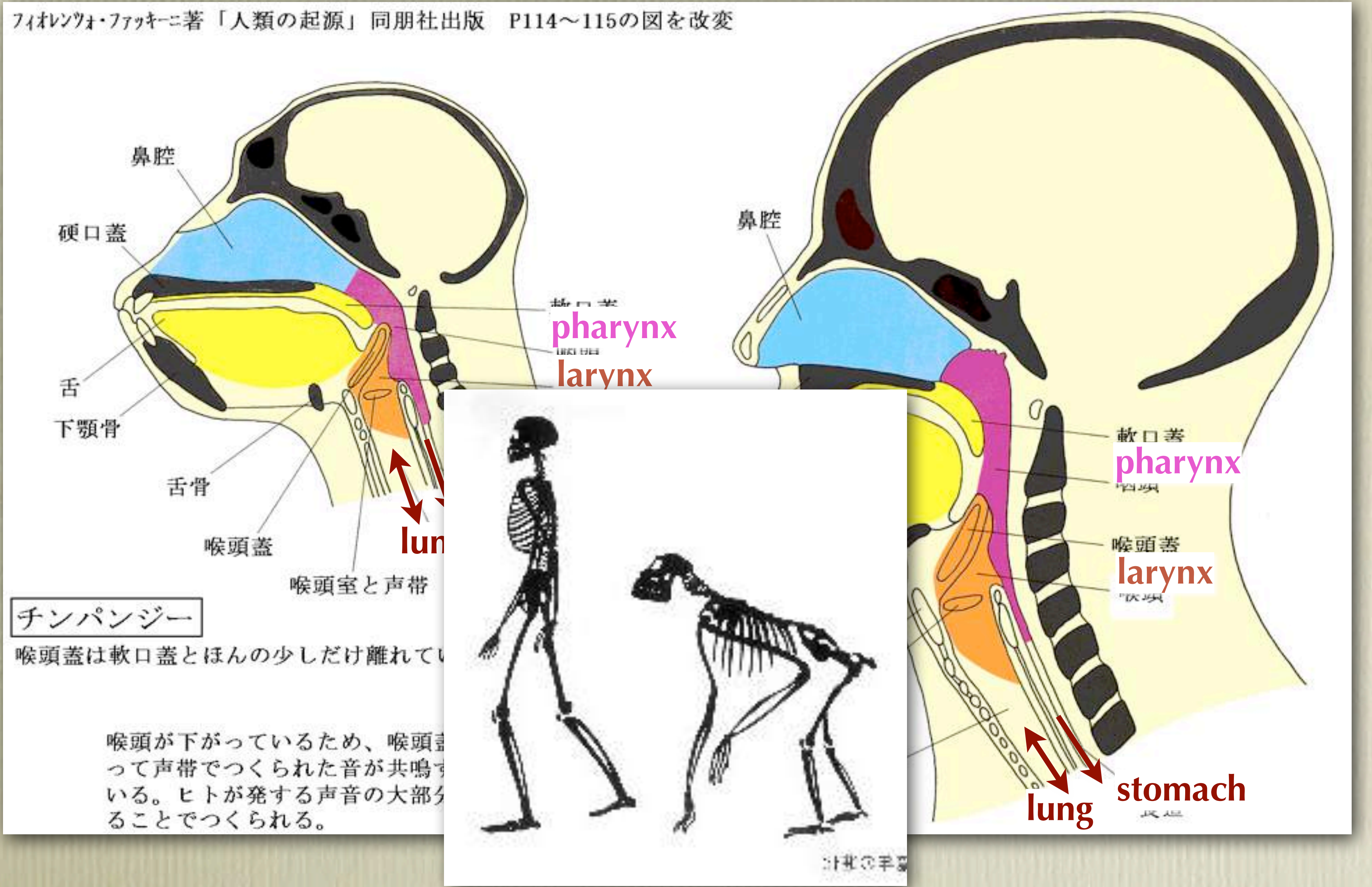
Two questions

- **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



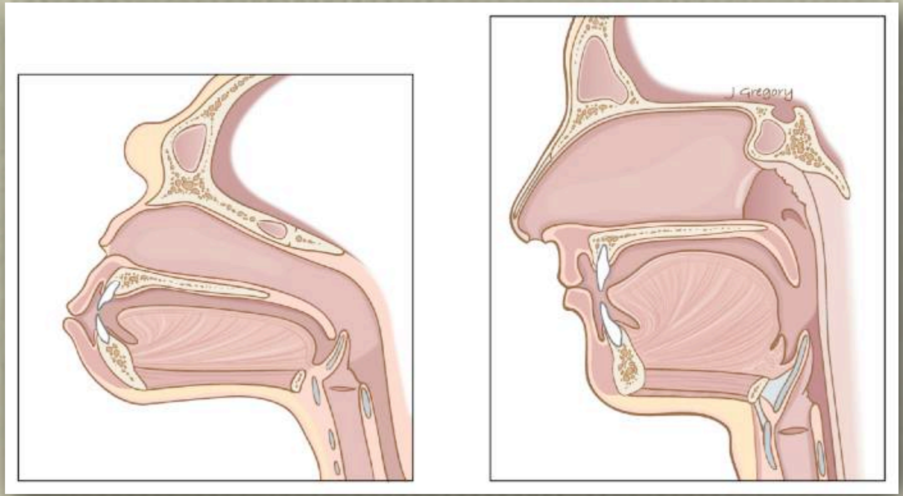
チンパンジー

喉頭蓋は軟口蓋とほんの少しだけ離れてい

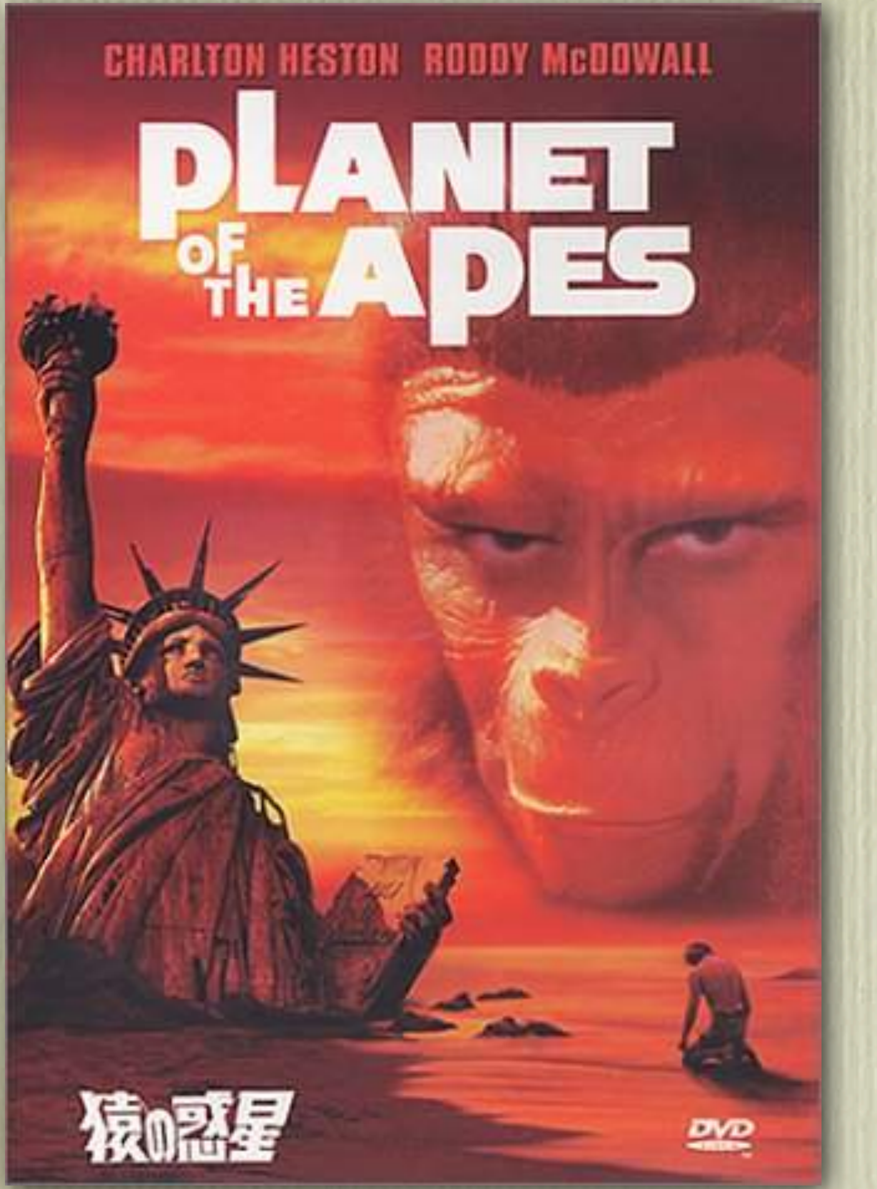
喉頭が下がっているため、喉頭蓋
って声帯でつくられた音が共鳴す
いる。ヒトが発する声音の大部分
ることでつくられる。

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.



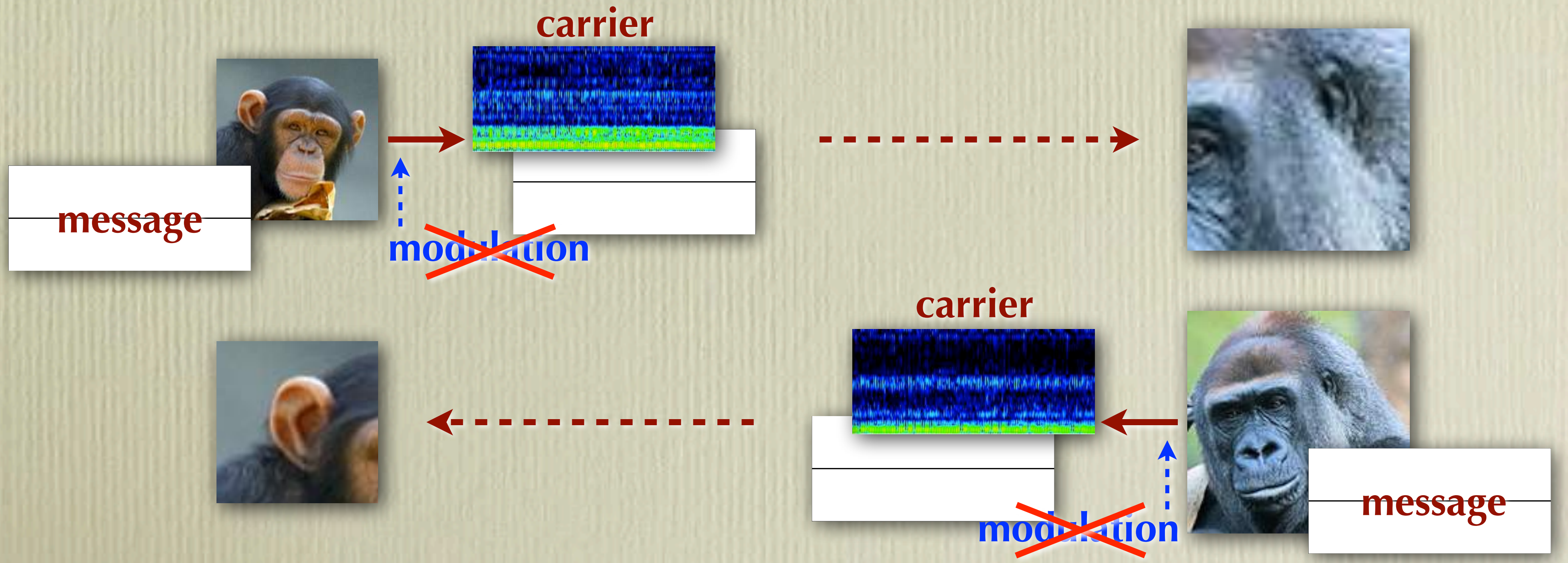
Old and new "Planet of the Apes"



Q1: Does the ape have a good modulator?

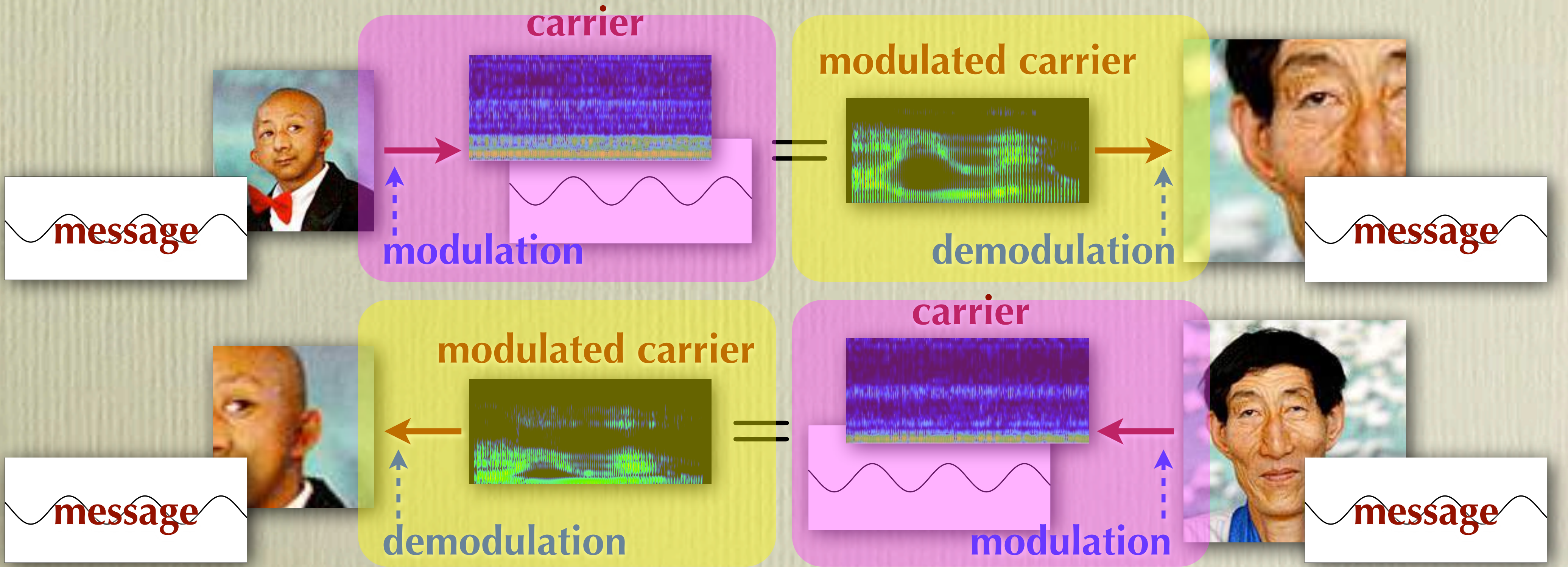
Morphological characteristics of the ape's tongue

- 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.



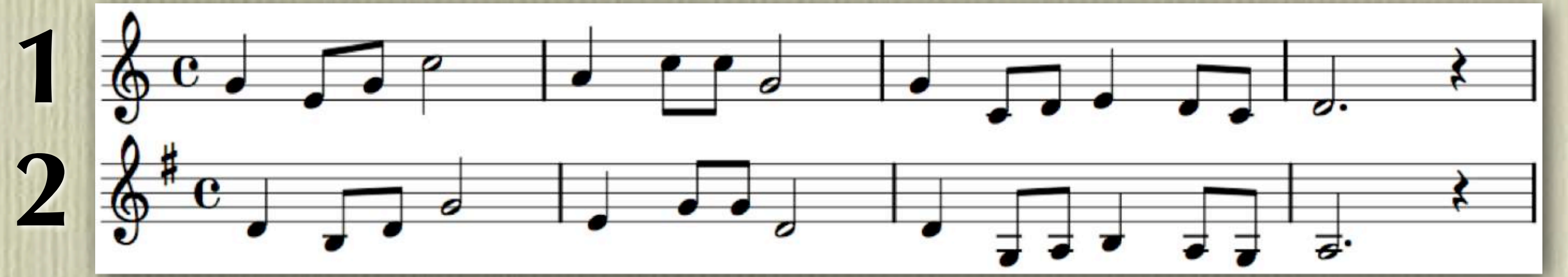
Two questions

- **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

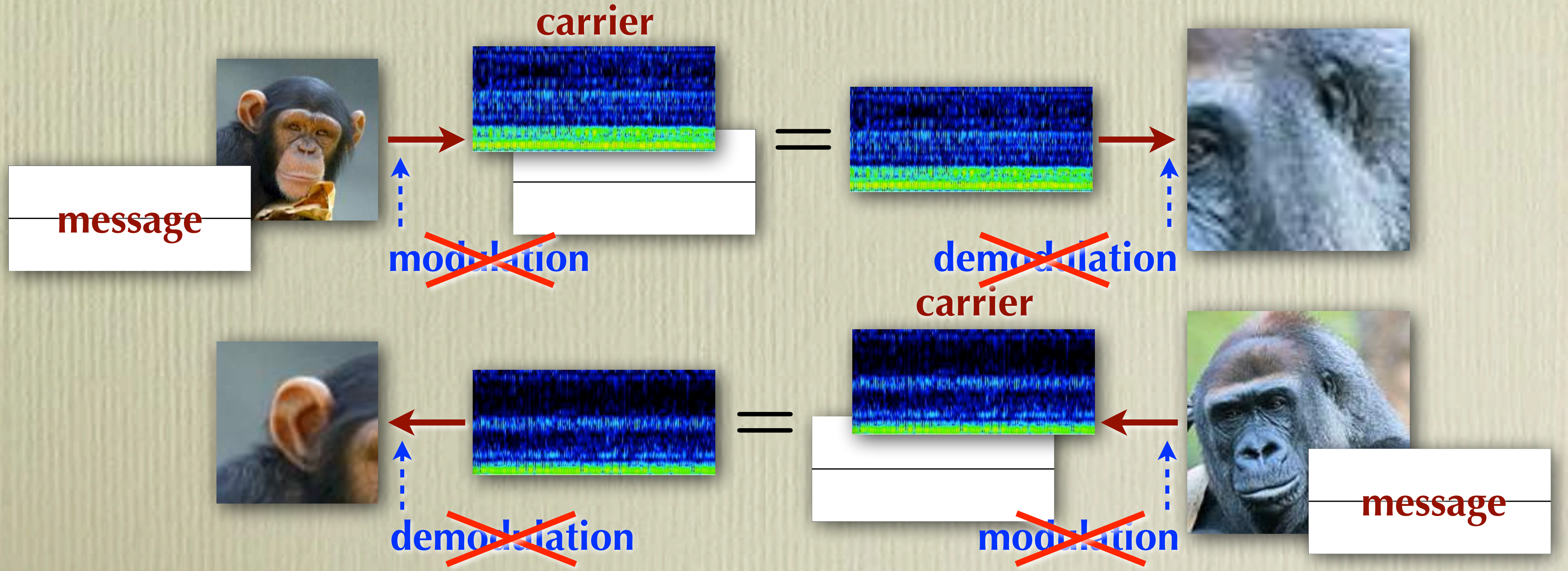
At least, frequency (pitch) demodulation seems difficult.

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].

Q2: Does the ape have a good demodulator?

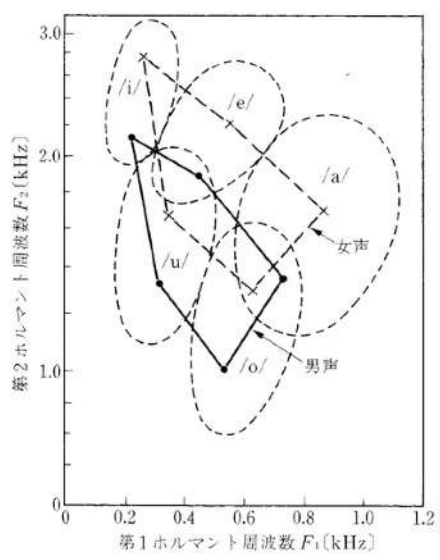
- **Cognitive difference bet. the ape and the human**
  - Humans can extract embedded messages in the modulated carrier.
  - It seems that animals treat the (modulated) carrier as it is.
- **From the (modulated) carrier, what can they know?**
  - The apes can identify individuals by hearing their voices.
    - Lower/higher formant frequencies = larger/smaller apes

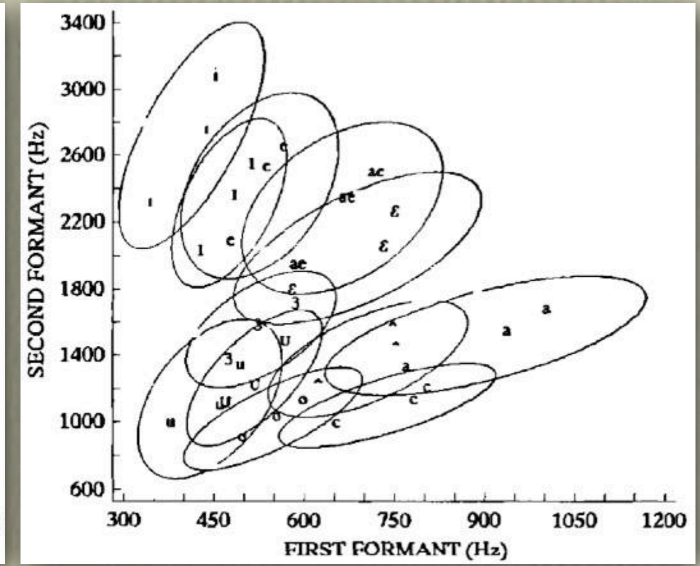


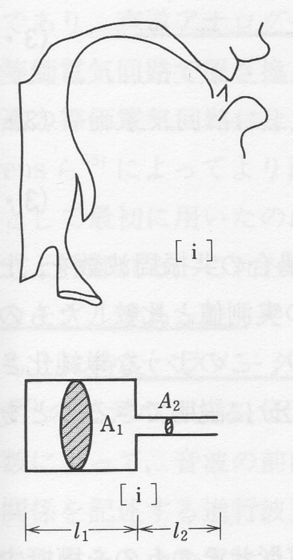
Function of the voice timbre

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.







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

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

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

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.

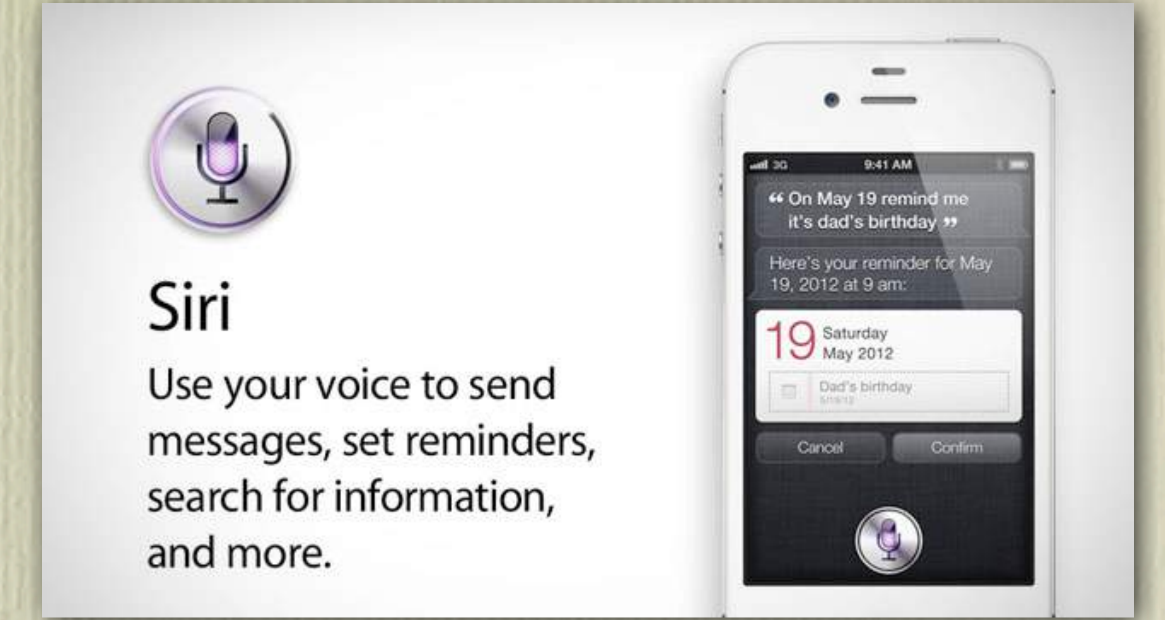
But the voice timbre can be changed easily.

- 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 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 the definition of "human-like" robots?

What is the goal of speech engineering?



Siri

Use your voice to send messages, set reminders, search for information, and more.

“On May 19 remind me it's dad's birthday”

Here's your reminder for May 19, 2012 at 9 am:

19 Saturday May 2012

Dad's birthday

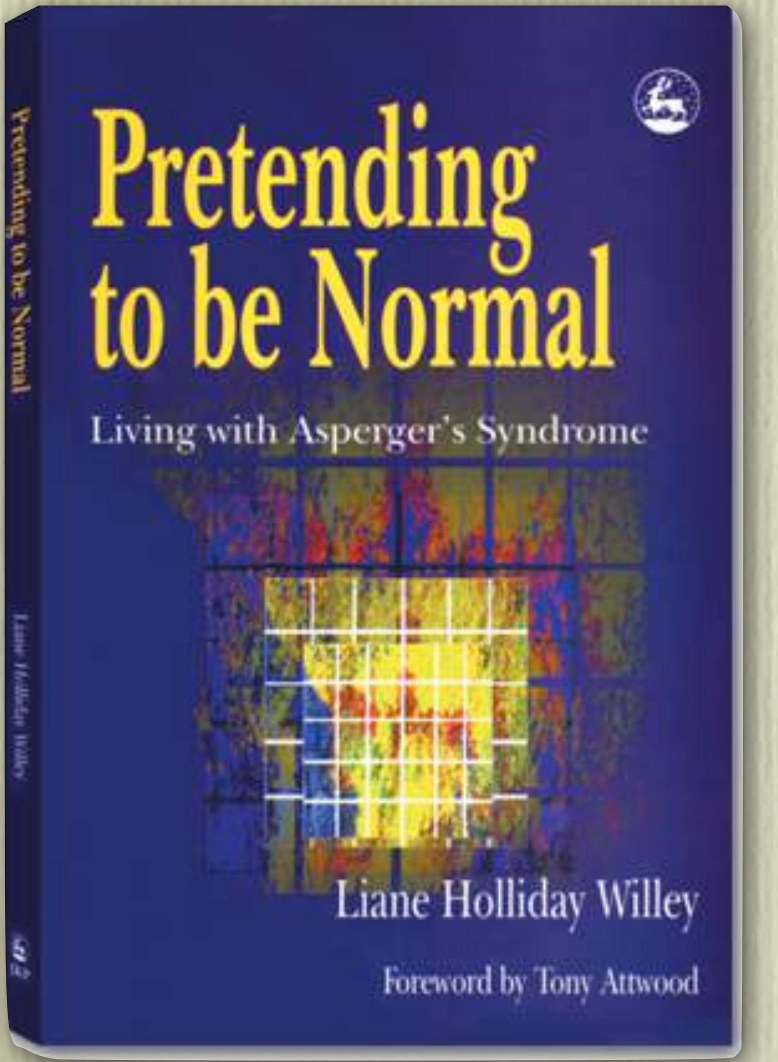
Clever Hans

A horse who can "calculate"

- 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).



Pretending to be Normal
Living with Asperger's Syndrome
アスペルガー的人生
Liane Holliday Willey
リアン・ホリデー・ウィリー 著
ニキ・リンコ 訳
大人になるまで知らずに「探索」し続けた
アスペルガー症候群の半生
当事者、配偶者、家族、友人、教育関係者、雇用主、医療・療育関係者 必携
「ガイドブック アスペルガー症候群」著者 トニー・アトウッド推薦
サバイバルと理解のための「ハウツー編」付き
ウェブサイト/関連図書も紹介
東京書籍

Definition of “human-likeness”

- 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 robot.

Final assignment

- **1. Read the following two papers and give your own comments.**
  - Both papers are available at the lecture’s site.
    - http://www.gavo.t.u-tokyo.ac.jp/~mine/japanese/media2017/class.html
  - **A: “Speech structure and its application to robust speech processing”**
  - (A’: “音声に含まれる言語的情報を非言語情報から音響的に分離して抽出する方法の提案 〜人間らしい音声情報処理の実現に向けた一検討〜”)
  - **B: “A modulation and demodulation model for speech communication and its emergence”**
- **2. Show your own necessary conditions of “human-likeness”.**
- **3. Comment on the content of this class. Your comments will be reflected on this class in the future.**
- Submission
  - PDF should be sent to mine@gavo.t.u-tokyo.ac.jp
  - The file name should be [student_id]_[your name].pdf
- Deadline = Feb. 6 (Tue) 23:59