

# Cognitive Media Processing #10

---

**Nobuaki Minematsu**



# Title of each lecture



- Theme-1

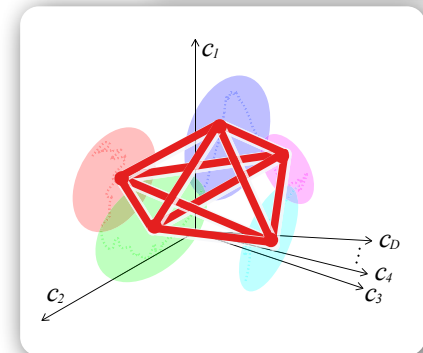
- ~~Multimedia information and humans~~
- ~~Multimedia information and interaction between humans and machines~~
- ~~Multimedia information used in expressive and emotional processing~~
- ~~A wonder of sensation - synesthesia -~~

- 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



# Speech is extremely variable.

Various factors change speech acoustics easily.



The world's tiniest high school girl





# 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!!

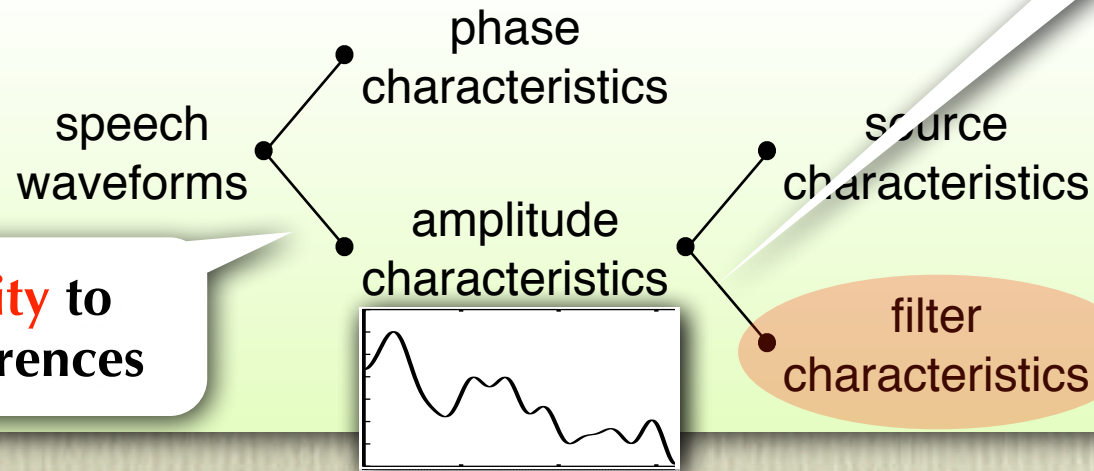
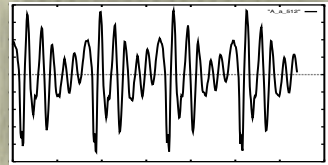




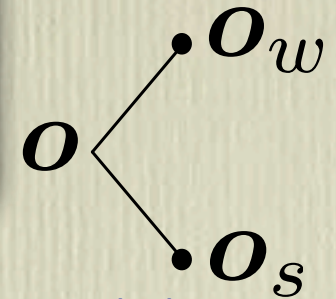
# Feature separation to find specific info.

## 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?



# Inensitivity in our language learning

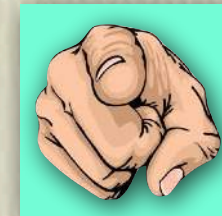
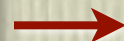
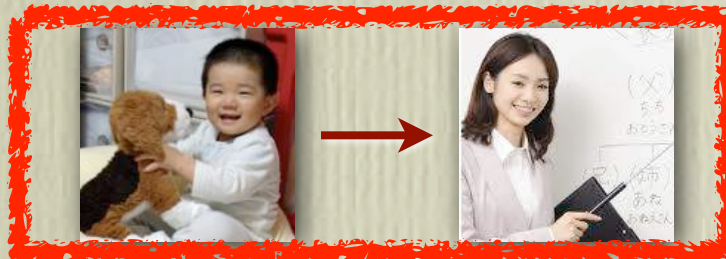
## Vocal learning (including vocal imitation)

- A imitate(s) B vocally.
  - A: students and B: teachers
  - A: infants and B: parents (caretakers)
  - A: you and B: professional singer (Karaoke)
  - But A do not impersonate B.
    - Acoustically *mismatched* imitation.
- We're very insensitive to speaker identity transmitted via speech.



## Acoustically **matched** imitation is often found in

- Autistics (自閉症), who have language disorder [Grandin'96]
- Animals' vocal imitation (birds, dolphins, whales, etc) [Okanoya'08]





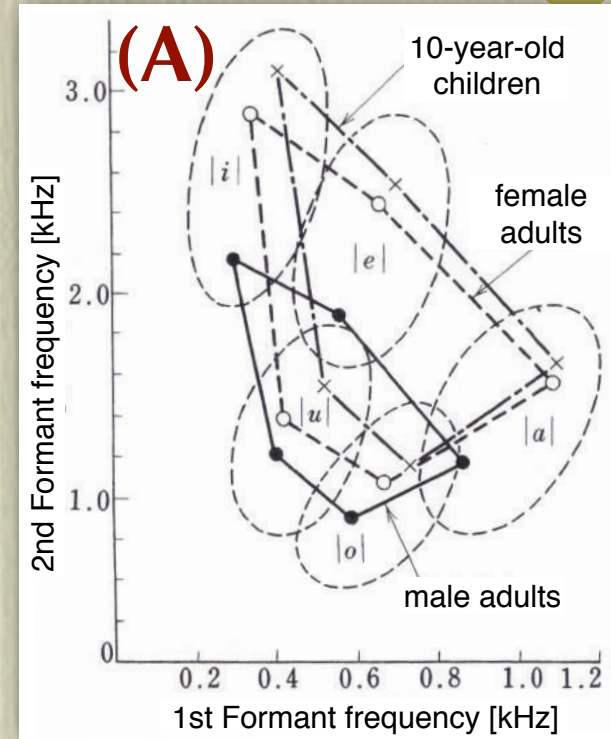
# Inensitivity and sensitivity

## Infants' vocal learning is

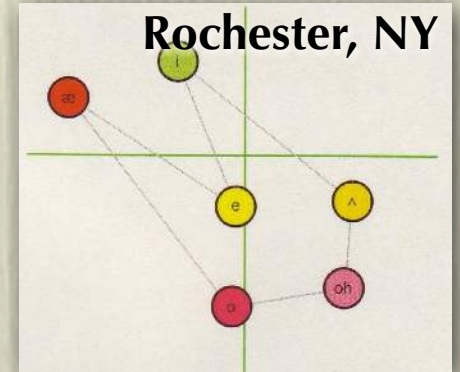
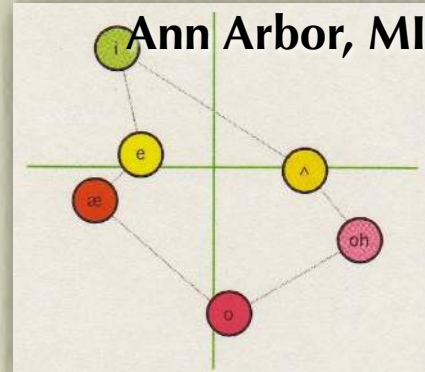
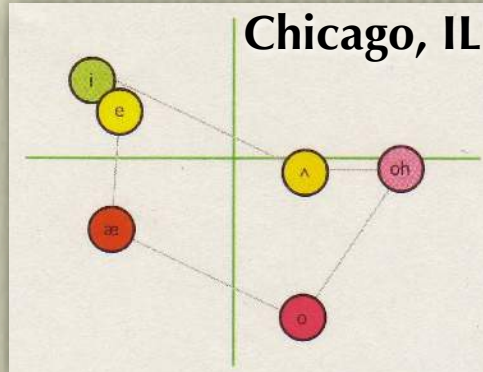
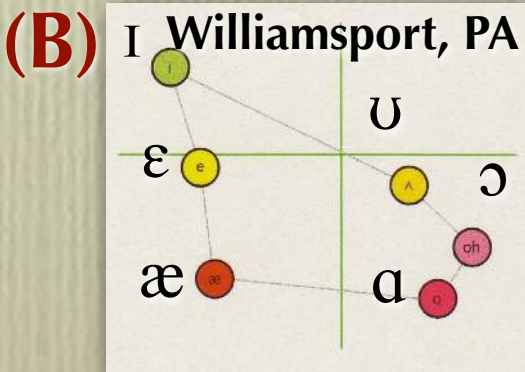
- insensitive to age and gender differences. **(A)**
- sensitive to accent differences. **(B)**

## 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



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



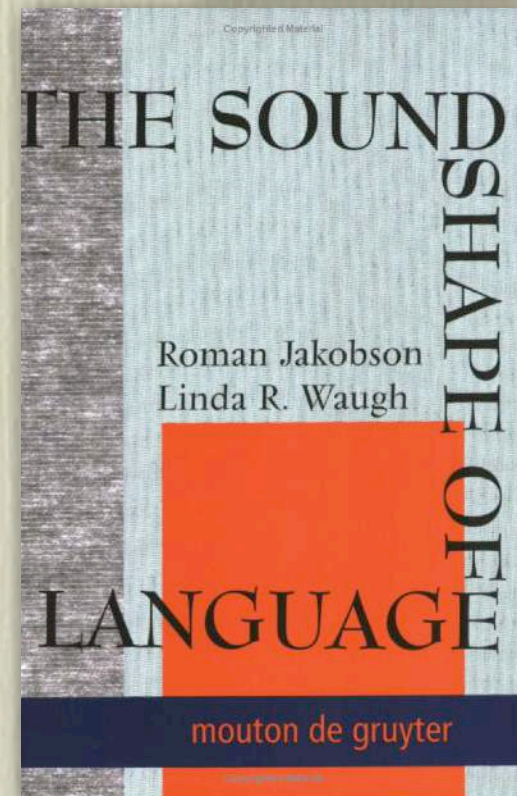
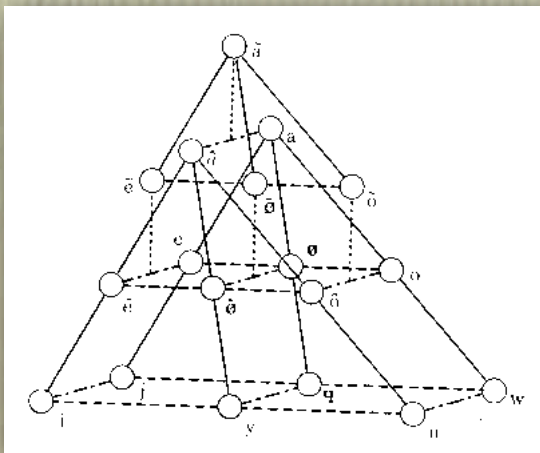
# A claim found in classical linguistics

## 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.

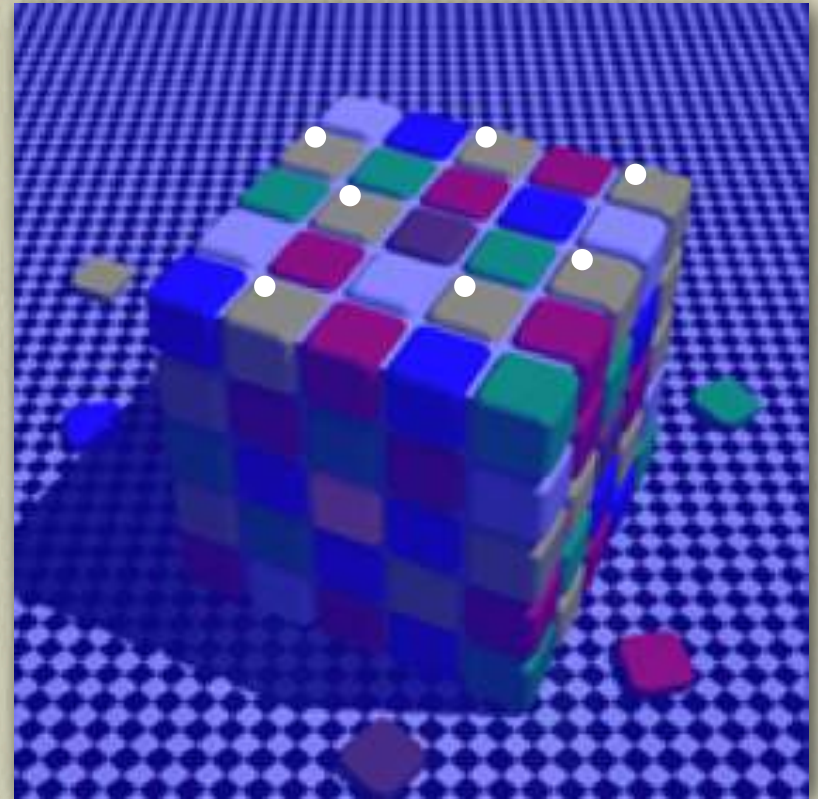
Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. in their relations to the other sounds.





# Invariant **color** perception against its bias

The Rubik's cube seen through colored glasses [Lotto'99]



- We perceive that the two cubes are identical.
- Different / identical colors are claimed to be identical / different.
- Not only wavelength (absolute property) of each patch, but also it matters **what contrast each patch has to its surrounding patches.**



# Invariant **pitch** perception against its bias

## Key change (transposition) of a melody [Higashikawa'05]

1 

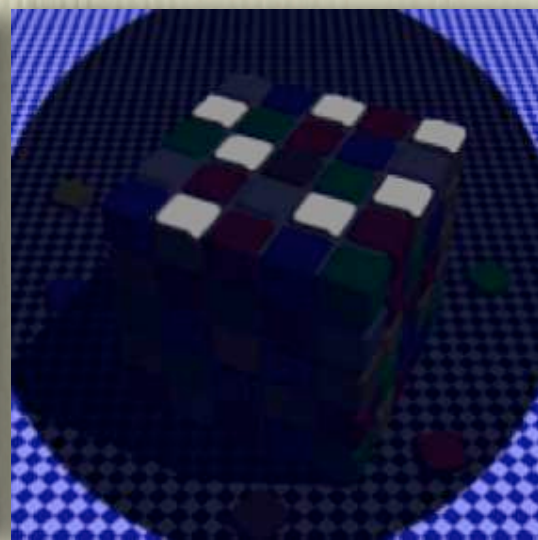
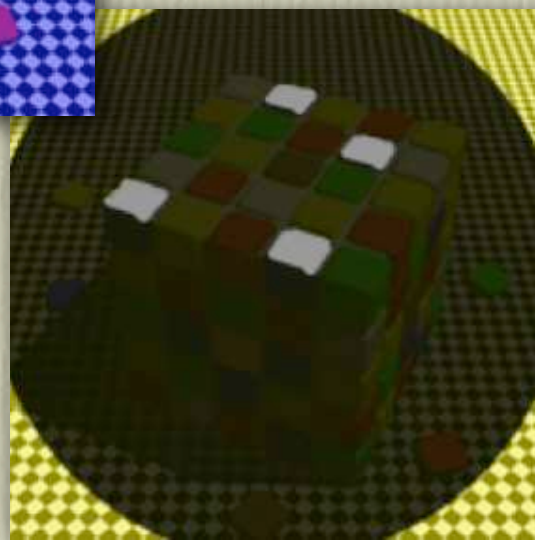
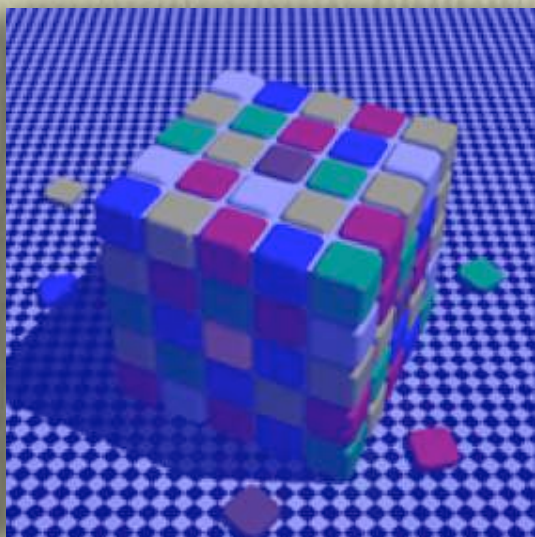
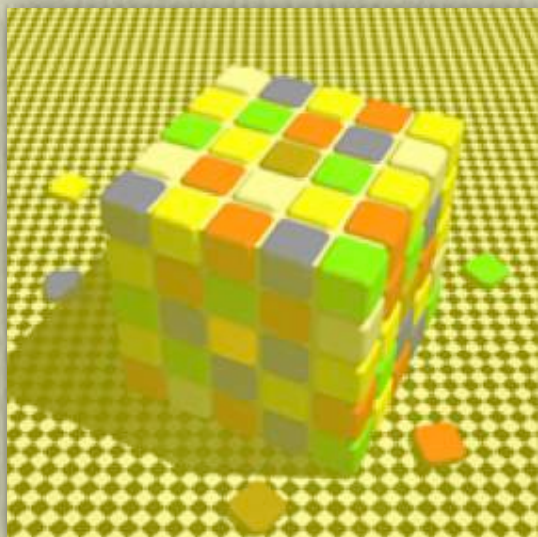
2 

- Absolute (perfect) pitch (Do, Re, Mi... = **pitch names**) (音名)
  - 1 = So, Mi, So, Do, La, Do, Do, So. 2 = Re, Ti, Re, So, Mi, So, So, Re.
- Relative pitch **with transcription ability** (Do, Re... = **syllable names**)
  - 1 = **So**, Mi, So, Do, **La**, Do, Do, So. 2 = So, Mi, So, **Do**, **La**, Do, Do, So. (階名)
- Relative pitch **without transcription ability**
  - 1 = La, La, La, La, La, La, La, La. 2 = La, La, La, La, La, La, La, La
- **Different / identical** tones are claimed to be **identical / different**.
- Not fundamental frequency (absolute property) of each tone, but it only matters **what contrast each tone has to its surrounding tones**.



# An evolutionary point of view

How old is the relative perception in evolution? [Briscoe'01]





# An evolutionary point of view

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



1 = 2





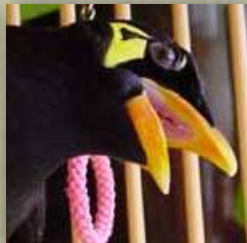
# Language acquisition through **vocal imitation**

## VI = children's active imitation of parents' utterances

- Language acquisition is based on vocal imitation [Jusczyk'00].
- VI is very rare 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 $\neq$ acoustic = ??

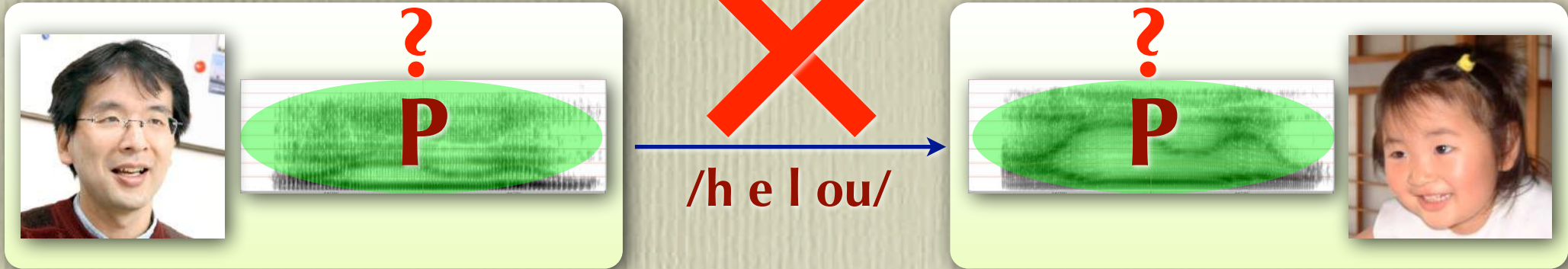
- Acoustic imitation performed by myna birds [Miyamoto'95]
  - They imitate the sounds of cars, doors, dogs, cats as well as human voices.
  - Hearing a very good myna bird say something, one can guess its owner.
- **Beyond-scale** imitation of utterances performed by children
  - No one can guess a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





# Language acquisition through vocal imitation

Utterance → symbol sequence → production of each sym.



- Phonemic awareness is too poor to decompose an utterance.

## Several answers from developmental psychology

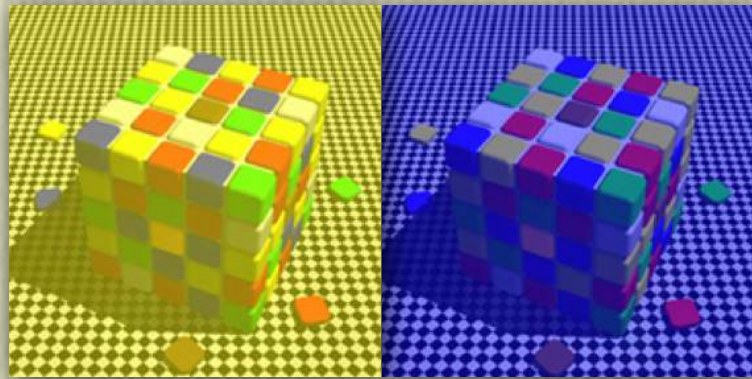
- Holistic/related sound patterns embedded in utterances
  - **Holistic wordform** [Kato'03]
  - **Word Gestalt** [Hayakawa'06]
  - **Related spectrum pattern** [Lieberman'80]
- The patterns have to include **no** speaker information in themselves.
  - If they do it, children have to try to impersonate their fathers.
  - **What is the speaker-invariant and holistic pattern in an utterance?**



# Invariant **timbre** perception against its bias

## Invariant and constant perception wrt. **color and pitch**

- Contrast-based information processing is important.
- Holistic & relational processing enables **element** identification.



## Invariant and constant perception wrt. **timbre**

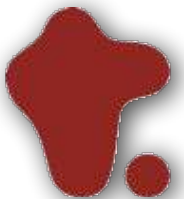
- Contrast-based information processing is important.
- Holistic & relational processing enables **element** identification.



# A new framework for “human-like” speech machines #2

---

**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

- An interesting link to some behaviors found in language disorder
- An interesting thought experiment



# Impersonation vs. non-impersonation

A very talented impersonator of Seiko Matsuda



Seiko's impersonator



Seiko's daughter



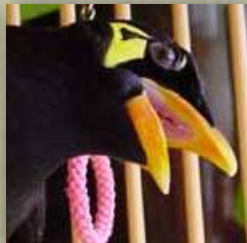
# Language acquisition through **vocal imitation**

## VI = children's active imitation of parents' utterances

- Language acquisition is based on vocal imitation [Jusczyk'00].
- VI is very rare 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 $\neq$ 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].





# “I impersonate a teacher.”

## Some comments from an autistic women

- Q: “How do you do vocal imitation in a Karaoke box or in a class of foreign language?”
- A: “I impersonate a professional singer or a teacher.”
  - B: “Acoustic imitation seems to be her default strategy of vocal imitation.”
- A: “Spoken language is difficult to use.”
  - A: “Written language and sign language are much easier.”





# TV program with talented impersonators

Can you enjoy such a TV program?

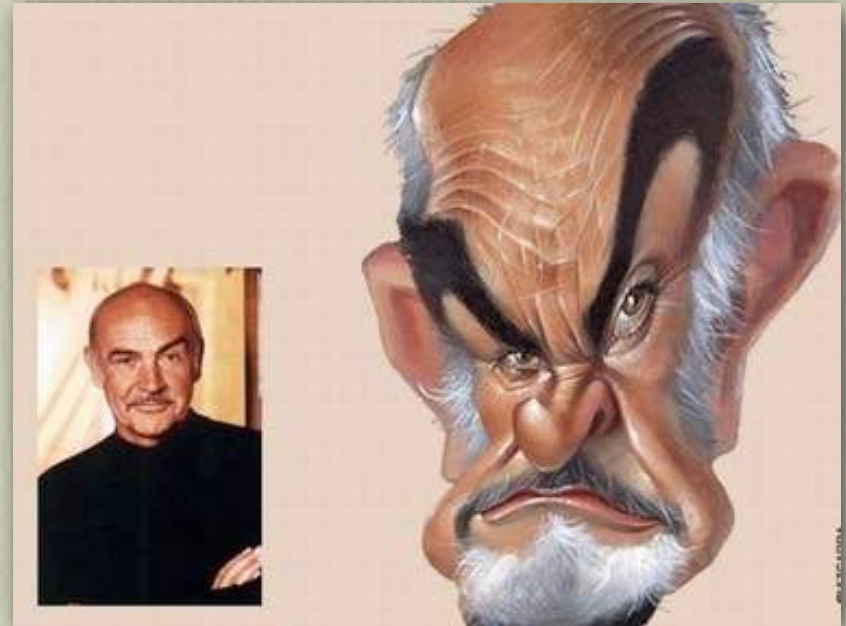
I cannot understand what is amusing.

Can you perceive any similarity between these pictures?

No. I believe that this is much similar to this picture.

**Robust perception of equivalence against deformation**

Our perception is very robust with a certain kind of deformation.





# TV program with talented impersonators

Can you enjoy such a TV program?

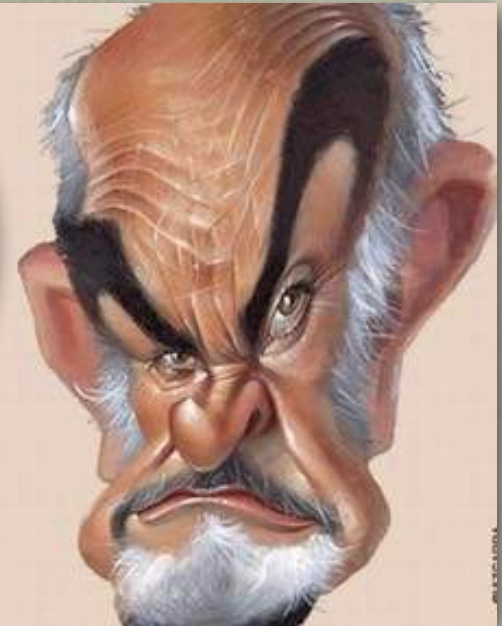
I cannot understand what is amusing.

Can you perceive any similarity between these pictures?

No. I believe that this is much similar to this picture.

**Robust perception of equivalence against deformation**

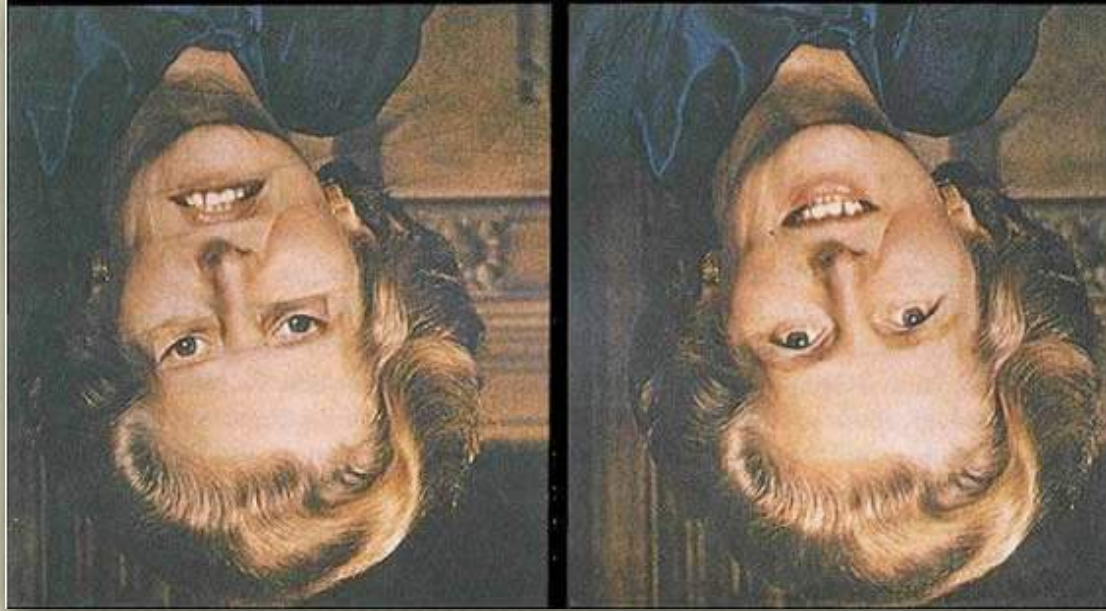
Our perception is very robust with a certain kind of deformation.





# Non-robustness with other deformation

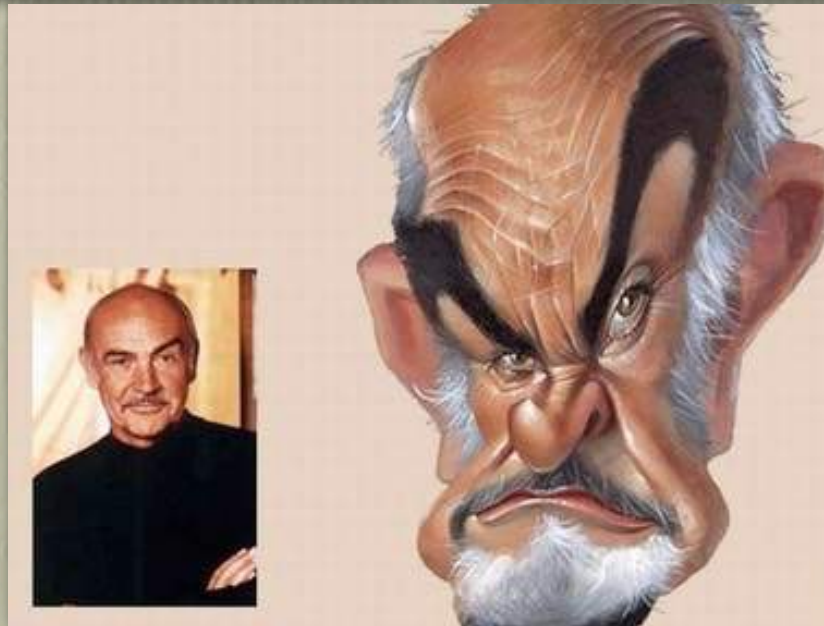
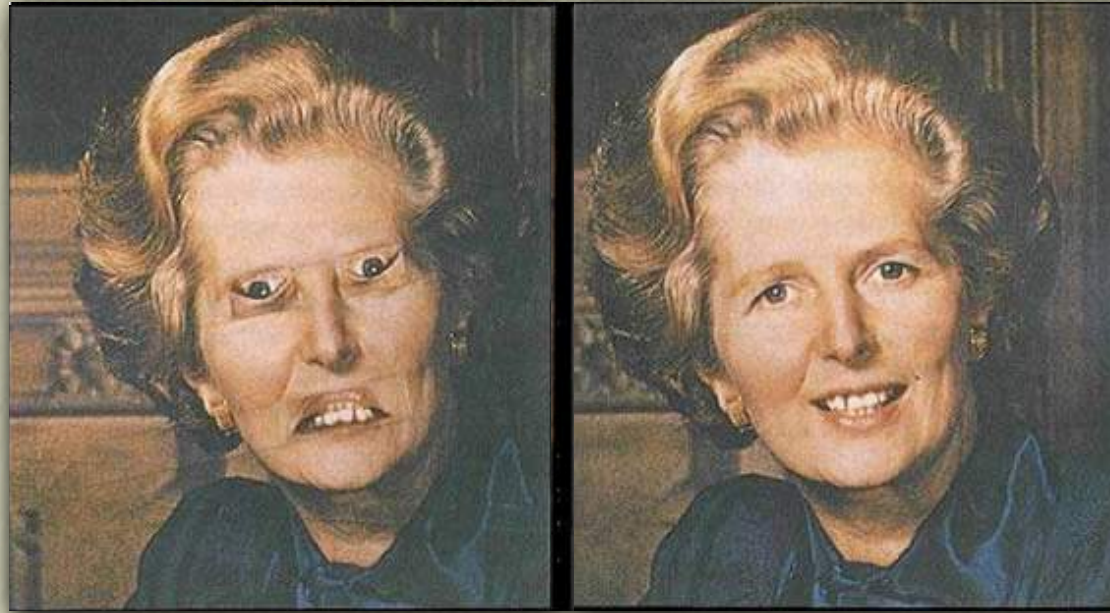
## Thatcher illusion





# Non-robustness with other deformation

## Thatcher illusion





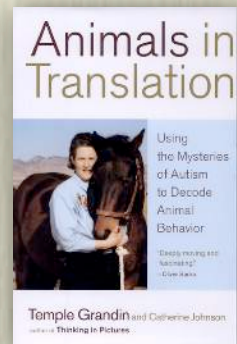
# Claims from a professor of animal sciences

## Dr. Temple Grandin @ Colorado State University

- She is herself autistic (Asperger syndrome).
- Autistics often imitate the utterances of TV/radio commercials.
  - TV/radio often gives “acoustically” identical utterances.
  - The utterances from family members change “acoustically” time to time.
- They often imitate the sounds of objects such as cars, doors, etc.
  - These sounds, including human voices, are just acoustic sounds.

## Interesting claims from her

- Similarity of information processing between animals and autistics
- Storing the detailed aspects of input stimuli as they are in the brain
  - Animal : **local / detail / absolute**
  - Human : **holistic / abstract / relative**
    - Good ability to generalize





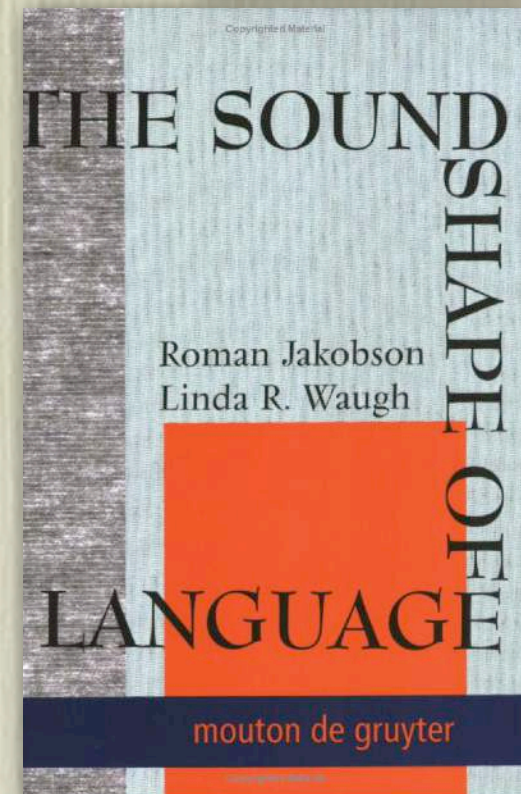
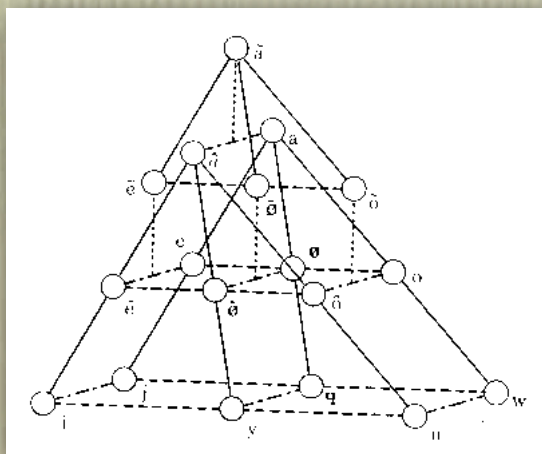
# A claim found in classical linguistics

## 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.





# Temple Grandin's TED talk

You can hear her talk at TED.

 Ideas worth spreading

Talks	TED Conferences	TED Conversations	About TED
Speakers	TEDx Events	TED Community	TED Blog
Playlists <b>NEW</b>	TED Prize	TED-Ed	TED Initiatives
Translations	TED Fellows		

We're creating a new TED.com experience.  
Want to try it out? [Request an invitation](#) today.

Follow TED    

Think you know what makes a safe driver?  TOYOTA Let's Go Places



# A book written by an autistic boy

“I can understand my mother’s utterances only”.



僕はお母さんの言うことならすべてわかります。それは、第1に安心感、第2に言葉のリズムや高低が良くわかっていて、第3に話の予測がつきやすいためでしょう。

どこにいてもどんなときでも、僕がわかる言葉は、お母さんだけです。  
僕は、どうしても今まで言葉が理解できないのか、わかりませんでした。他のみんなが指示されたことにすぐに反応できて、その通りに動けることが不思議でした。  
僕には聞こえないのです。  
音は聞こえているけれど、意味になって頭の中に入ってこないのです。話しているのが本人だとわかれば、慣れれば言っていることはわかります。でも、同じ人でも場所や状況が違えば、その人だということがわからないのです。

<http://www.nhk.or.jp/school-blog/300/195393.html>



# 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

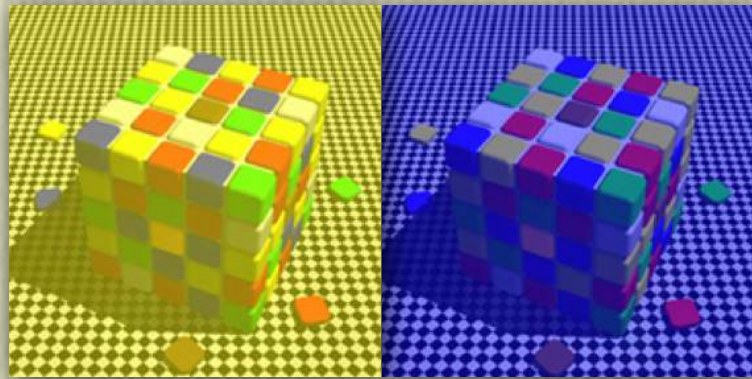
- An interesting link to some behaviors found in language disorder
- An interesting thought experiment



# Invariant **timbre** perception against its bias

## Invariant and constant perception wrt. **color and pitch**

- Contrast-based information processing is important.
- Holistic & relational processing enables **element** identification.



## Invariant and constant perception wrt. **timbre**

- Contrast-based information processing is important.
- Holistic & relational processing enables **element** identification.



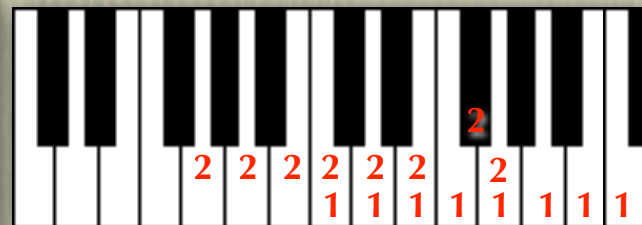


# Invariant **pitch** perception against its bias

## A melody and its transposed version [Higashikawa'05]

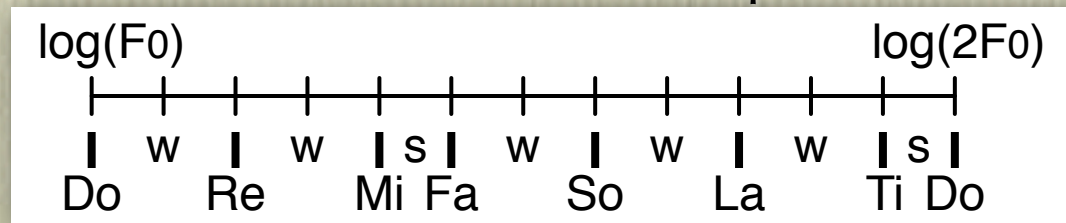
1) 

2) 

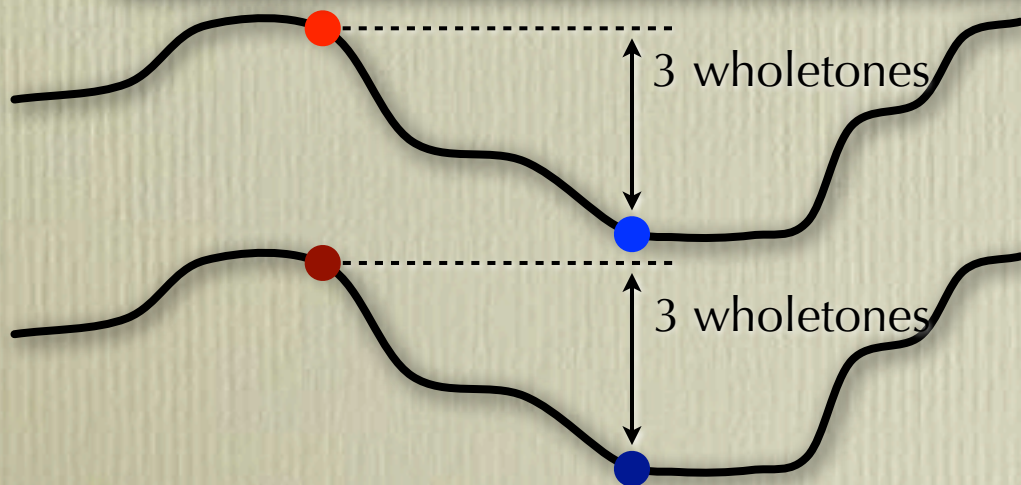


Listeners with RP can perceive the same sound name sequence.

- So Mi So Do / Ra Do Do So / So Do Re Mi Re Do / Re
- The same sound distribution pattern is found in **1)** and **2)**.



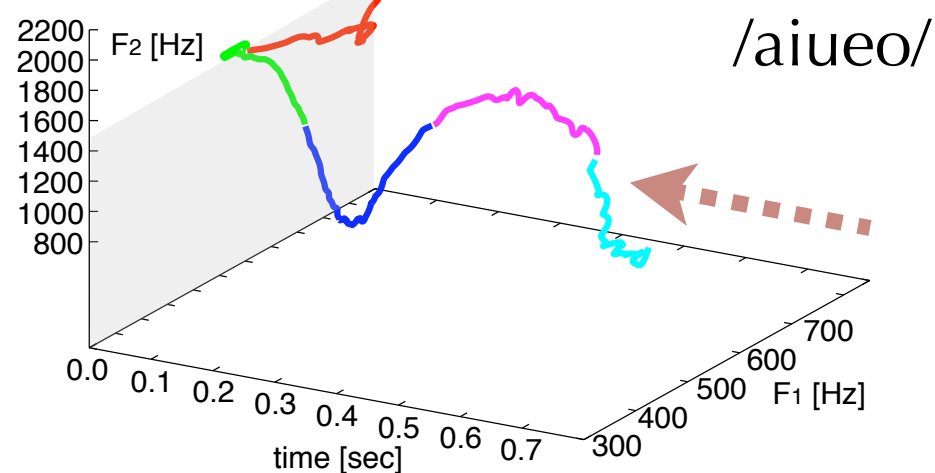
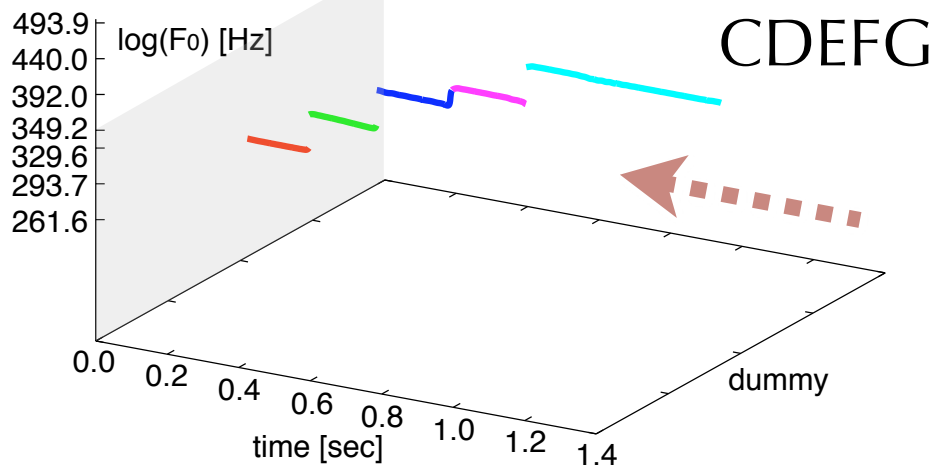
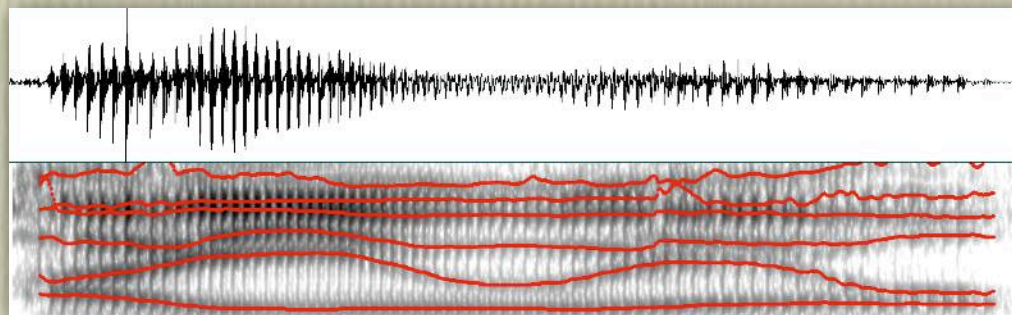
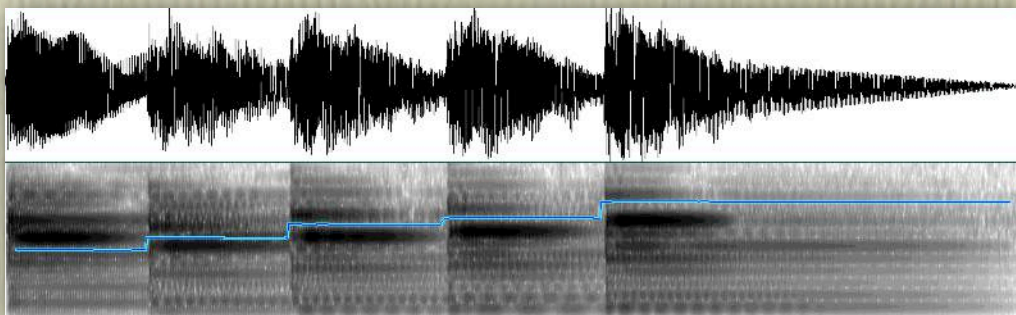
**Whole = 2 Semi**



➔ **● ● and ● ●** have to be fa & ti or ti & fa due to **contrastive constraints**.



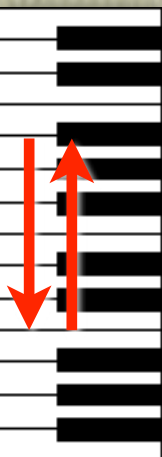
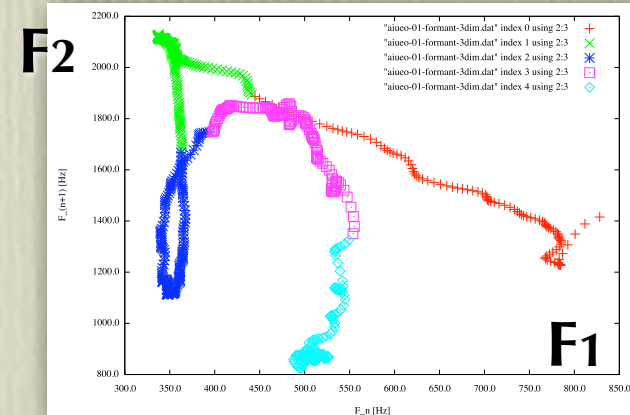
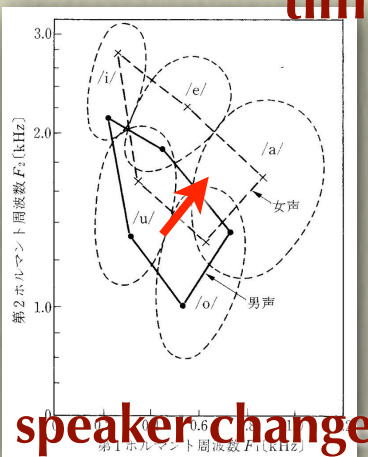
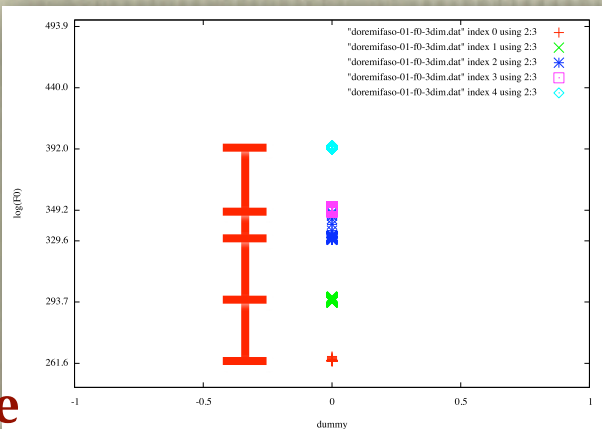
# Relative pitch vs. relative timbre



pitch modulation

timbre modulation

$\log(F_0)$

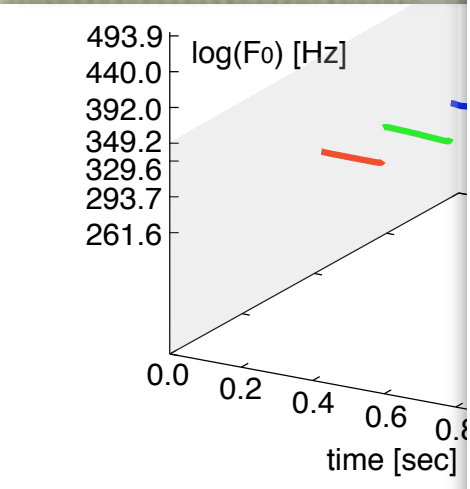
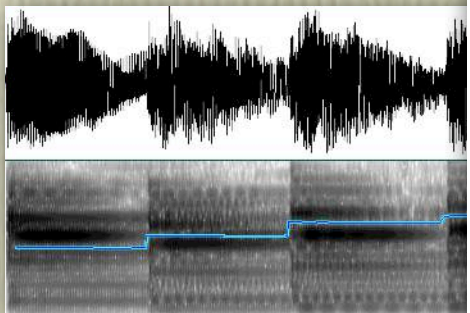


key change

speaker change

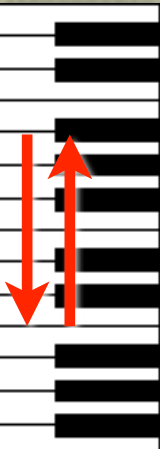


# Relative vowel space in the vocal tract

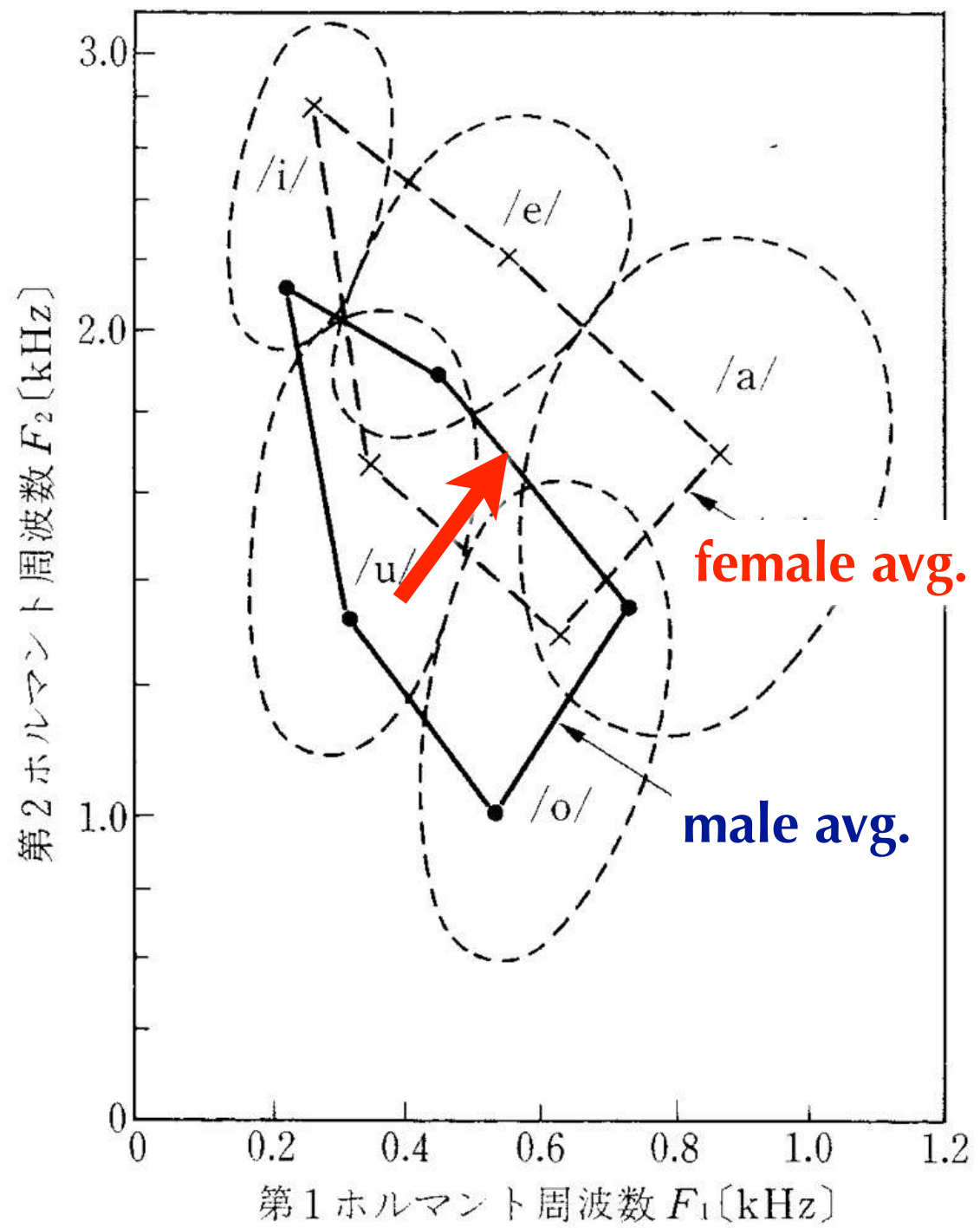


pitch n

$\log(F_0)$



key change



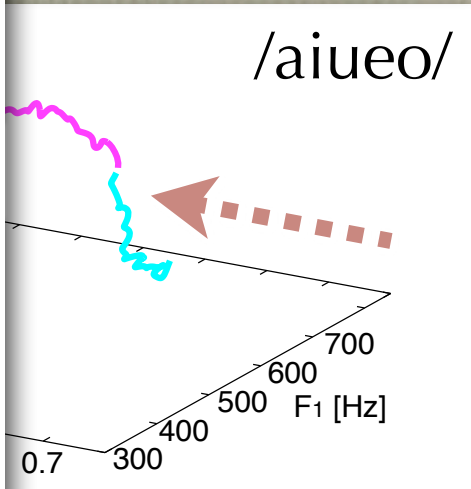
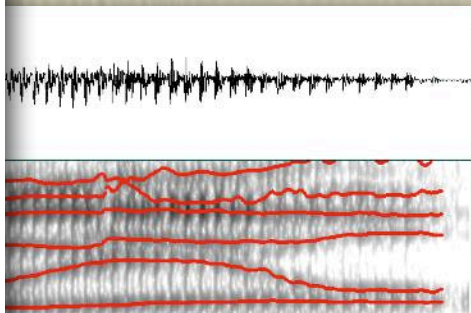
第2ホルマント周波数  $F_2$  [kHz]

第1ホルマント周波数  $F_1$  [kHz]

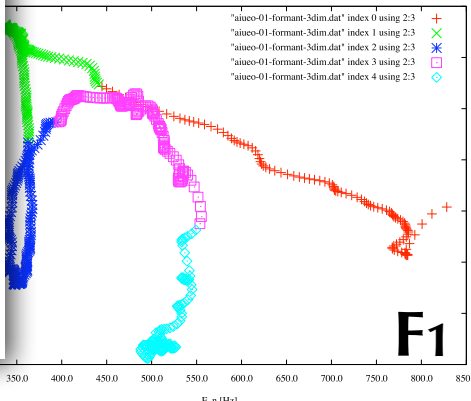
female avg.

male avg.

speaker change



dulation



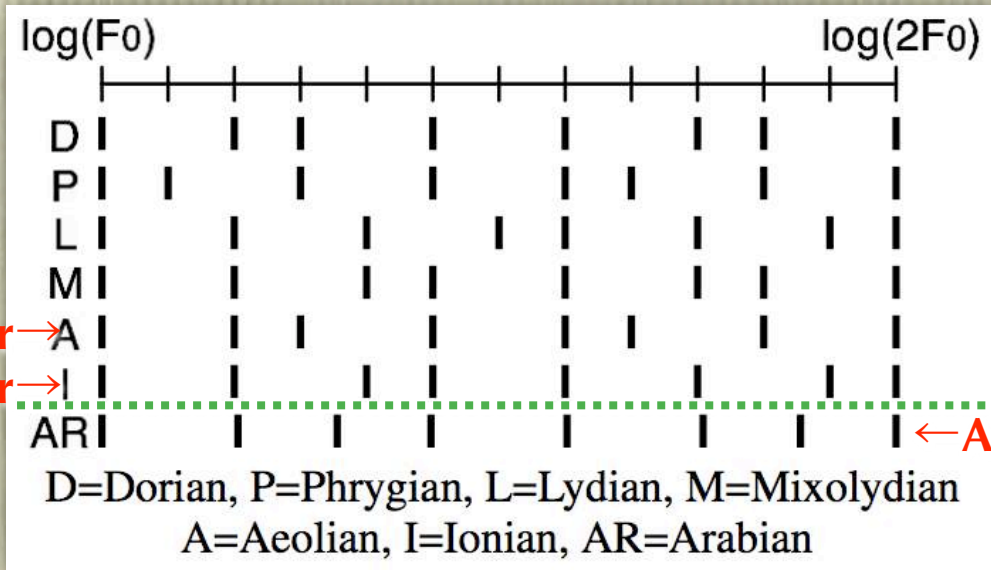
$F_1$

$F_2$  [Hz]



# Relative pitch vs. relative timbre

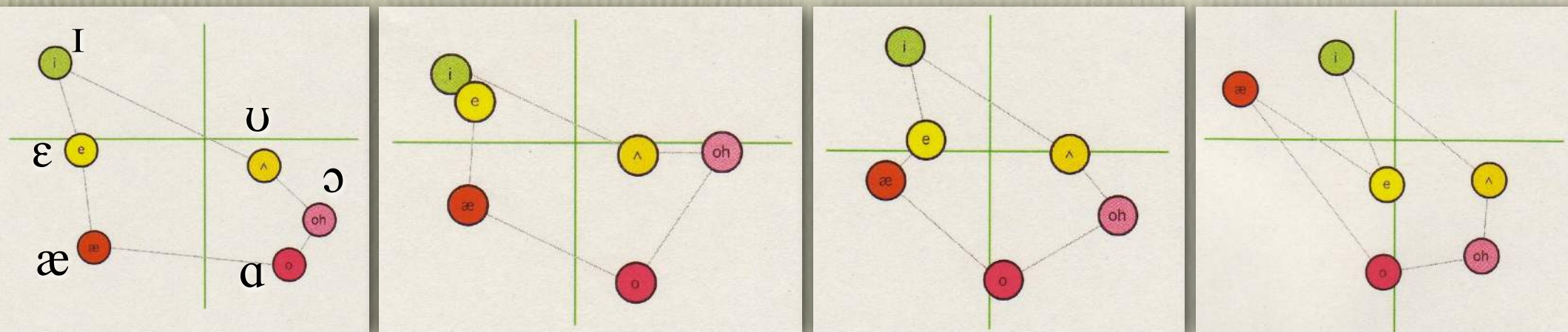
## Key-invariant arrangement of tones and its variants



- Western = 5 whole + 2 semi
- D to I = classical church music
- Arabic = with non-semi intervals
- Western music in Arabic scale

Minor → A  
Major → I  
← Arabic scale

## Spk-invariant arrangement of vowels and its variants



Williamsport, PA

Chicago, IL

Ann Arbor, MI

Rochester, NY



# Invariant **pitch** perception against its bias

## Key change (transposition) of a melody [Higashikawa'05]

1 

2 

- Absolute (perfect) pitch (Do, Re, Mi... = **pitch names**) (音名)
  - 1 = So, Mi, So, Do, La, Do, Do, So. 2 = Re, Ti, Re, So, Mi, So, So, Re.
- Relative pitch **with transcription ability** (Do, Re... = **syllable names**)
  - 1 = **So**, Mi, So, Do, **La**, Do, Do, So. 2 = So, Mi, So, **Do**, **La**, Do, Do, So. (階名)
- Relative pitch **without transcription ability**
  - 1 = La, La, La, La, La, La, La, La. 2 = La, La, La, La, La, La, La, La
- **Different / identical** tones are claimed to be **identical / different**.
- Not fundamental frequency (absolute property) of each tone, but it only matters **what contrast each tone has to its surrounding tones**.



# What's hard to do only with relative timbre?

People with RP w

- label a single tone

- Who cannot label

Identification of

- Difficult to label is

- Possible to transcr

melody cannot

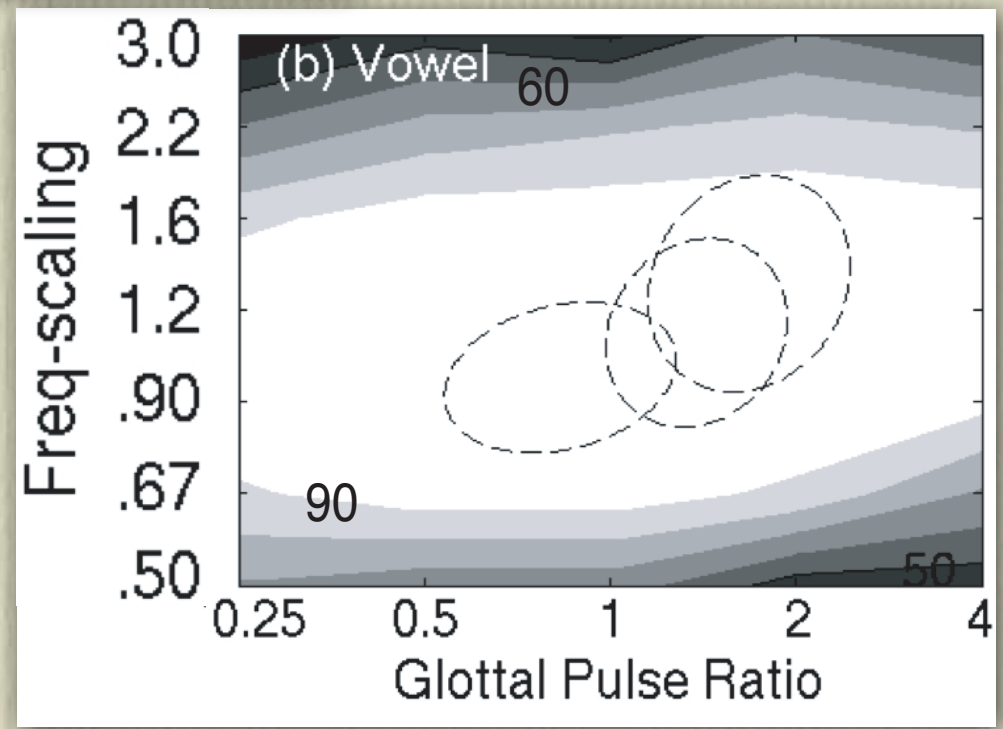
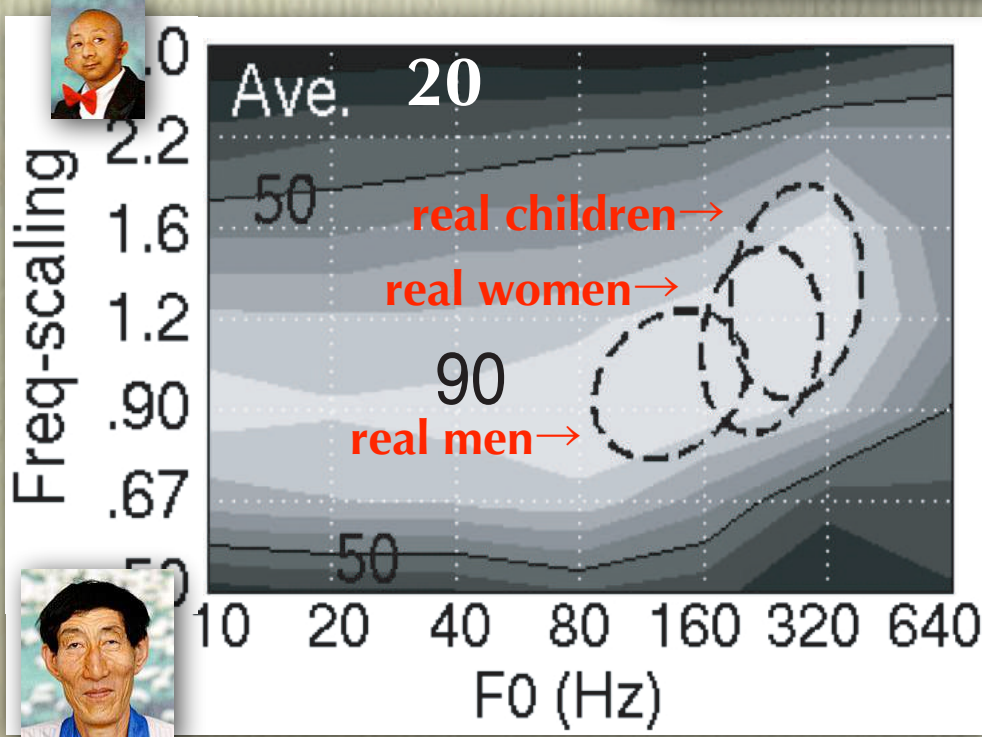
yllable name.

wel sound)?

giants and fairies

[04]

ce of morae [Hayashi'07]





# What's hard to

People with RP who

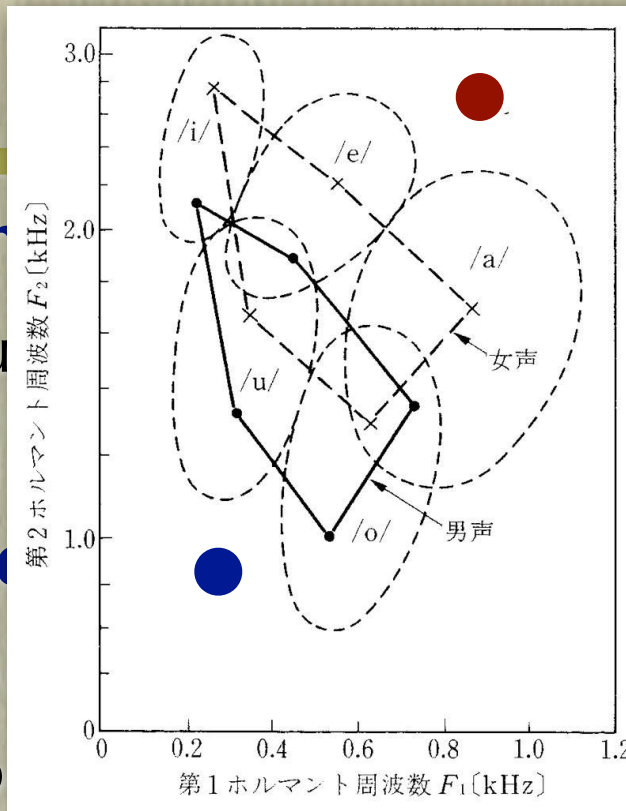
- label a single tone u

- Who cannot label a

## Identification of vowels

- Difficult to label iso

- Possible to transcrib



# relative timbre?

melody cannot

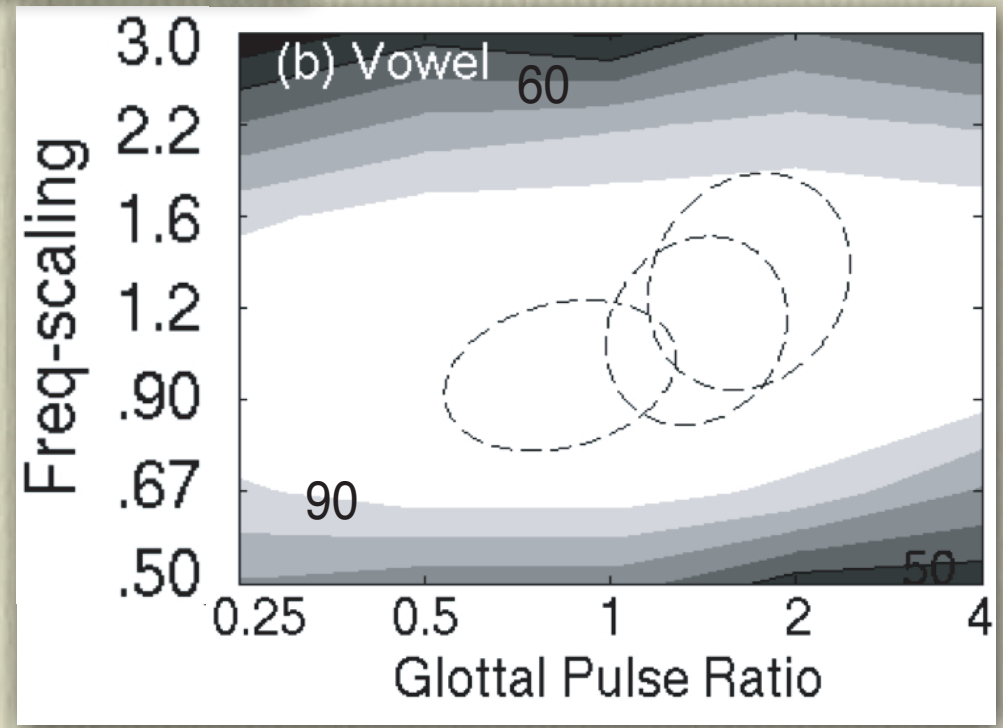
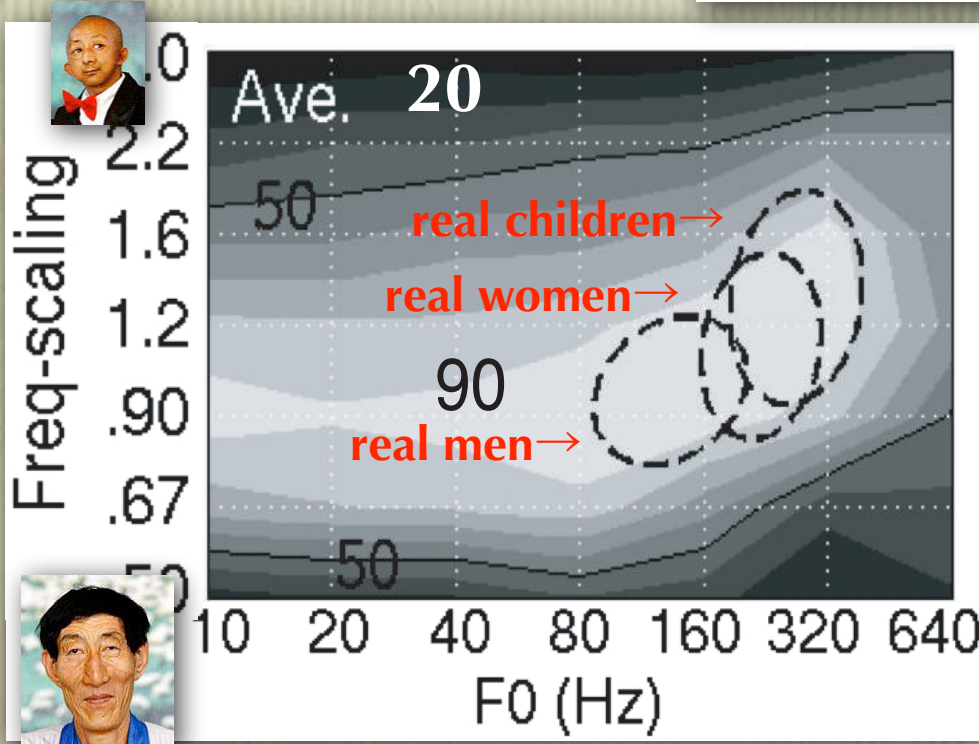
syllable name.

vowel sound)?

## giants and fairies

[oki'04]

presence of morae [Hayashi'07]





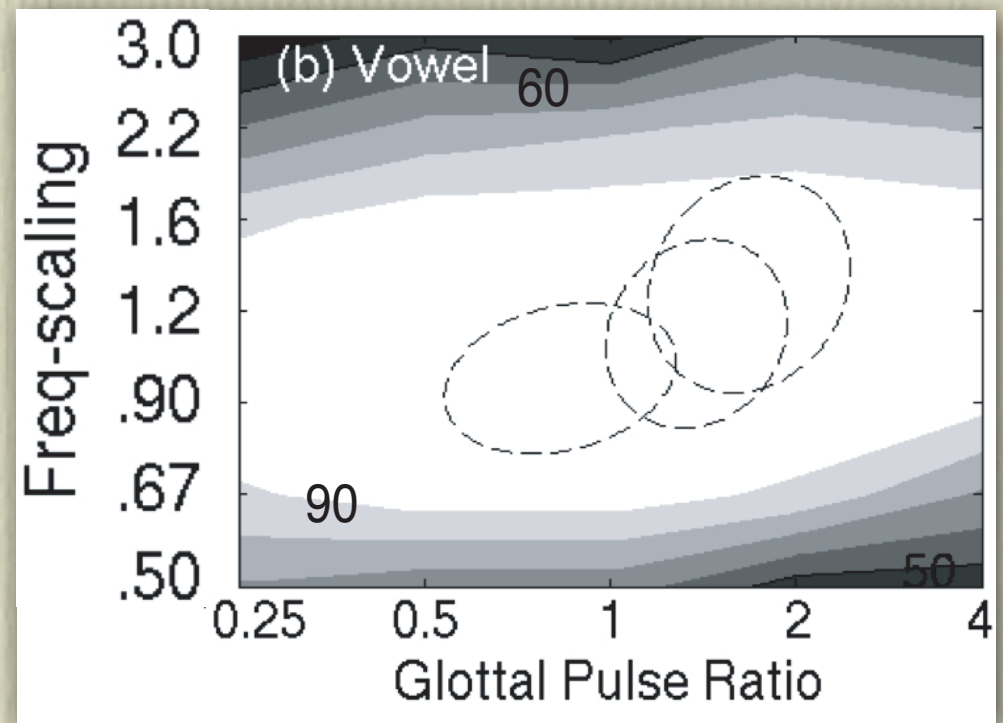
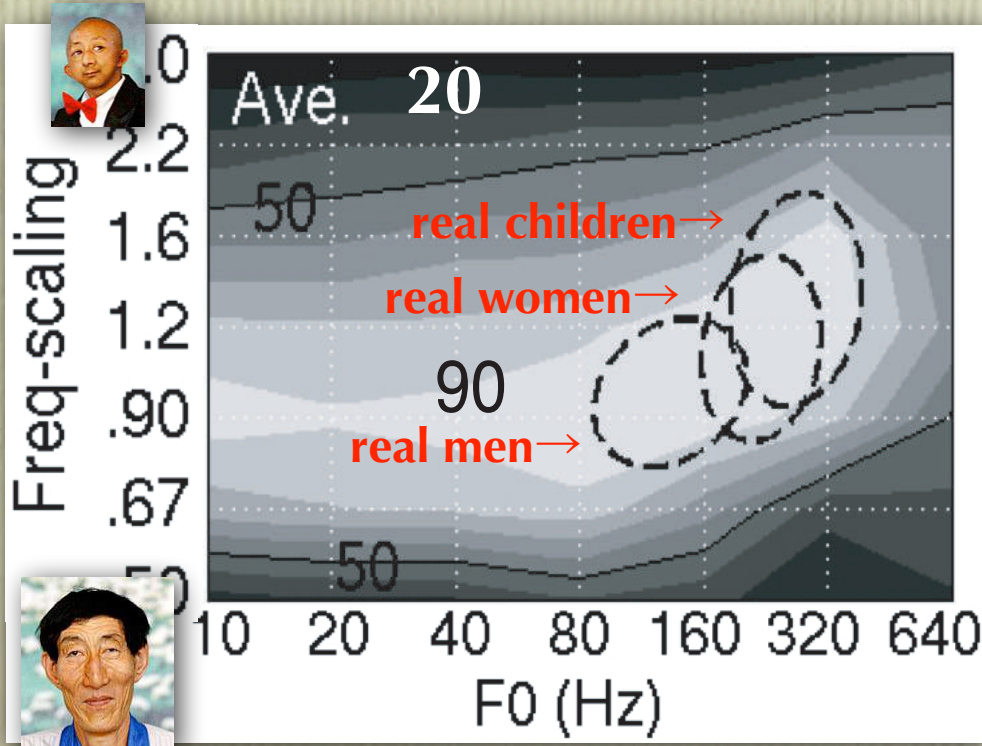
# What's hard to do only with relative timbre?

People with RP who can transcribe a melody **cannot**

- label a **single tone** using a pitch name or a syllable name.
- Who cannot label a single speech sound (vowel sound)?

**Identification of vowels produced by giants and fairies**

- Difficult to label isolated vowel sounds [Aoki'04]
- Possible to transcribe a meaningless sequence of morae [Hayashi'07]





# What's hard to do only with relative timbre?

People with RP v

label a single tone

Who cannot label

Identification of

Difficult to label is

Possible to transcr



melody cannot

yllable name.

wel sound)?

giants and fairies

[04]

ce of morae [Hayashi'07]

**Phonetic identification ability of isolated sounds may be unnecessary for oral communication?**

**Phoneme awareness is not needed for speech communication?**



# Invariant **pitch** perception against its bias

## Key change (transposition) of a melody [Higashikawa'05]

1 

2 

- Absolute (perfect) pitch (Do, Re, Mi... = **pitch names**) (音名)
  - 1 = So, Mi, So, Do, La, Do, Do, So. 2 = Re, Ti, Re, So, Mi, So, So, Re.
- Relative pitch **with transcription ability** (Do, Re... = **syllable names**)
  - 1 = **So**, Mi, So, Do, **La**, Do, Do, So. 2 = So, Mi, So, **Do**, **La**, Do, Do, So. (階名)
- Relative pitch **without transcription ability**
  - 1 = La, La, La, La, La, La, La, La. 2 = La, La, La, La, La, La, La, La
- **Different / identical** tones are claimed to be **identical / different**.
- Not fundamental frequency (absolute property) of each tone, but it only matters **what contrast each tone has to its surrounding tones**.

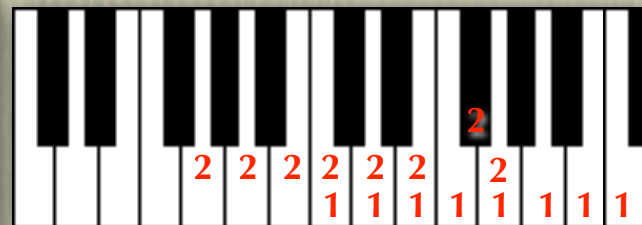


# Invariant **pitch** perception against its bias

## A melody and its transposed version [Higashikawa'05]

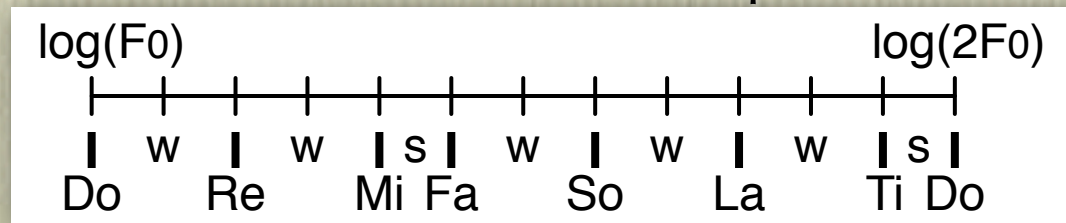
1) 

2) 

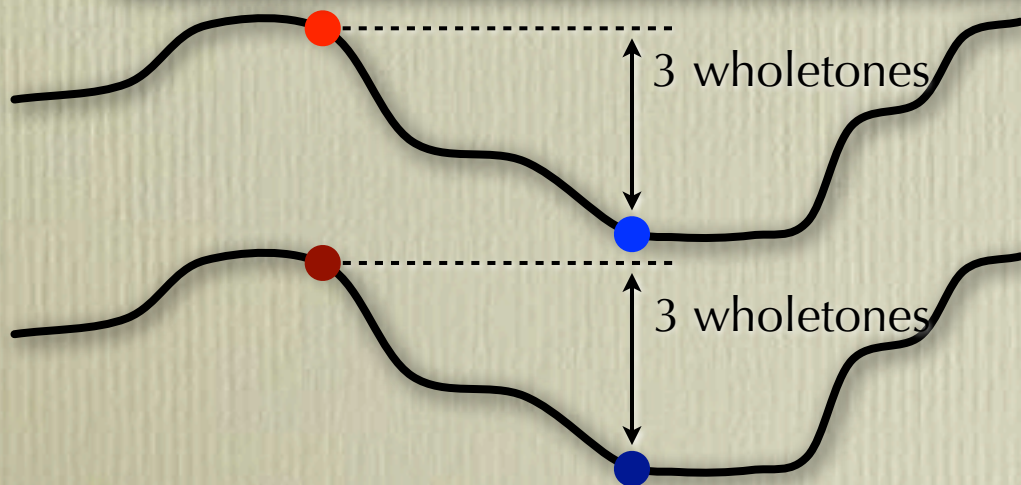


Listeners with RP can perceive the same sound name sequence.

- So Mi So Do / Ra Do Do So / So Do Re Mi Re Do / Re
- The same sound distribution pattern is found in **1)** and **2)**.



**Whole = 2 Semi**



➔ **● ● and ● ●** have to be fa & ti or ti & fa due to **contrastive constraints**.



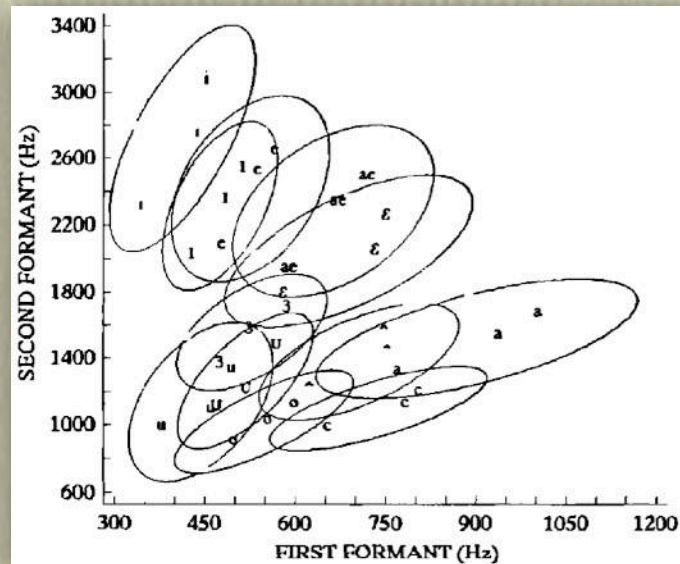
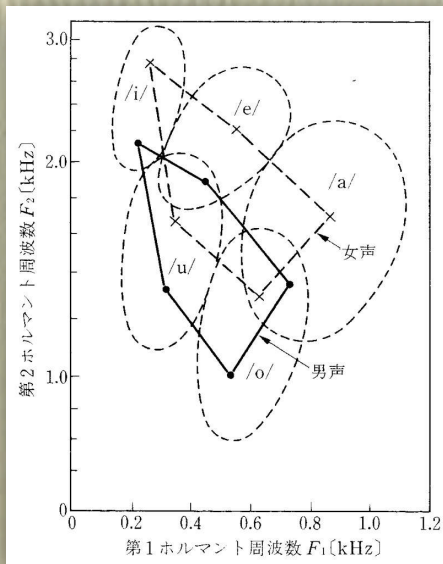
# Another hard thing to do for RP listeners

## Hard task for those who cannot transcribe a melody

- Keep the third *tone* in a given melody in mind. Then, raise your hand if you find the same *tone* in a new melody.
- If difficult to transcribe it using symbols, this request has to be hard.

## Hard task for the speech-version of these people

- Keep the third *sound* in a given utterance in mind. Then, raise your hand if you find the same *sound* in a new utterance.
- If difficult to transcribe it using symbols, this request has to be hard.



In US and UK, there have to be many people who have severe troubles in reading and writing?



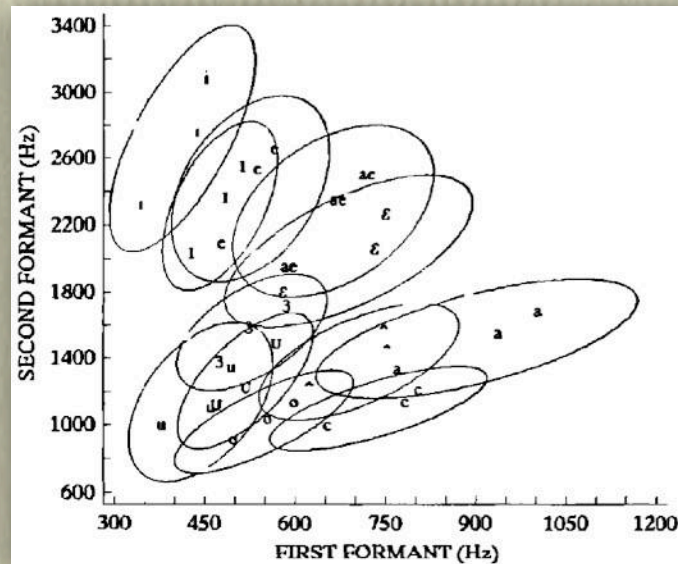
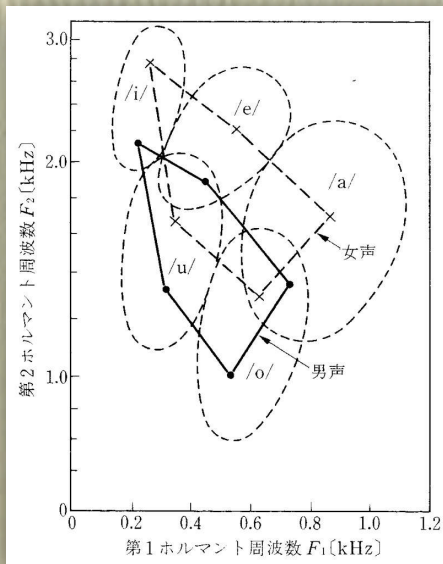
# Another hard thing to do for RP listeners

## Hard task for those who cannot transcribe a melody

- Keep the third *tone* in a given melody in mind. Then, raise your hand if you find the same *tone* in a new melody.
- If difficult to transcribe it using symbols, this request has to be hard.

## Hard task for the speech-version of these people

- Keep the third *sound* in a given utterance in mind. Then, raise your hand if you find the same *sound* in a new utterance.
- If difficult to transcribe it using symbols, this request has to be hard.



**Dyslexia**



And

g to

ners

Hard

cann

ody

Keep  
hand

en mel

e your

If

one in

difficult.

Hard

-versi

Keep  
hand

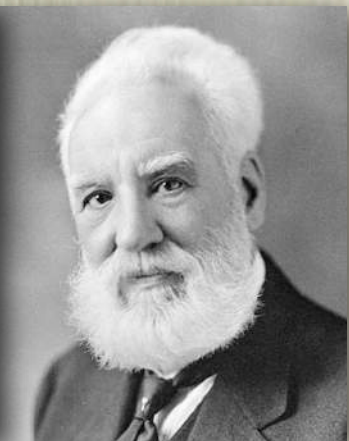
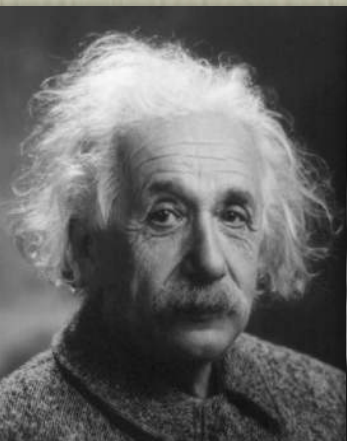
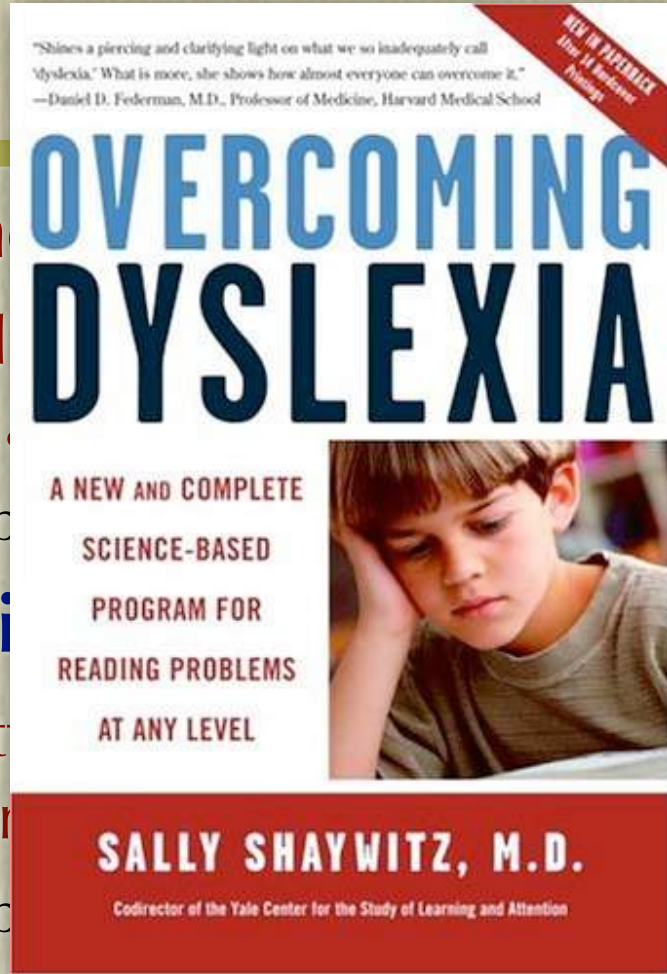
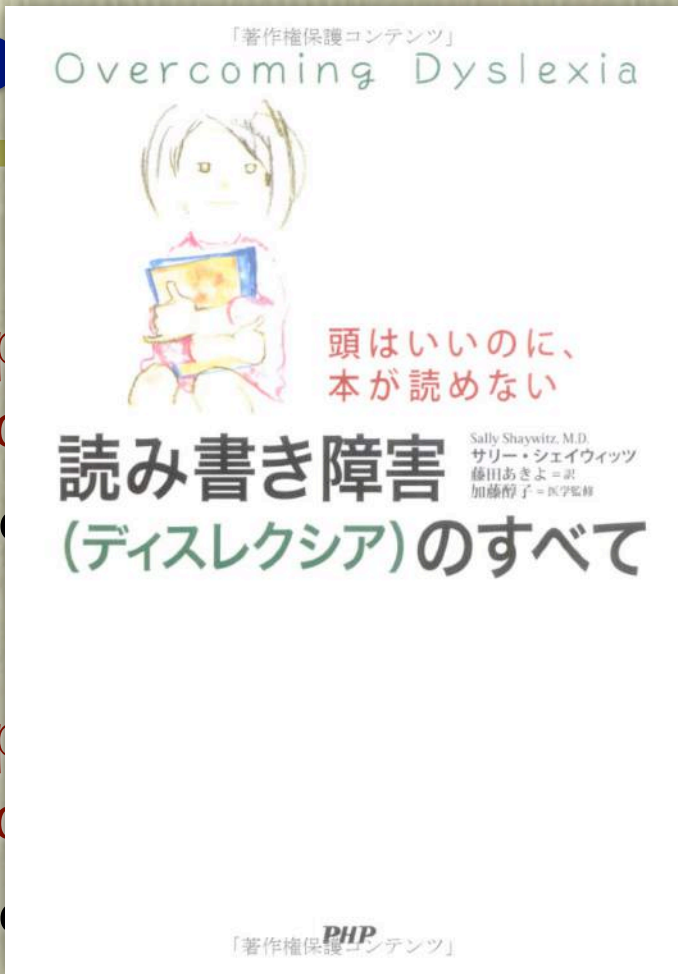
ven ut

raise your

If

ound in

difficult.



Dyslexia (phonological dyslexia)



# How I encountered dyslexia.

「あ」という声を聞いて母音「あ」と同定する能力は音声言語運用に必要なか？

はじめに、そんな馬鹿な、と思われるかもしれない。こんな実験を考えてみよう。身長300cmの巨人と50cmの小人に孤立母音を発声してもらう。通常音声学の教科書には、F<sub>1</sub>・F<sub>2</sub>

分かれる。しかし、その巨人、小人が無意味モーラ列を単

## 話し言葉の音声

### 第4章

音声認識研究からの一つの提言

峯松 信明

## 「あ」という声を聞いて母音「あ」と同定する能力は音声言語運用に必要なか？

はじめに、何、この変なタイトル？

タイトルを見て、多くの読者が首を傾げていることだろう。しかし、十一頁の本記事を読み終えた時に、ほぼ全ての読者に私の意図は通じるもの、と考えている。そう、「あ」という声を聞いて、それを有限個の音カテゴリーの一つとしての母音「あ」であると同定する能力は、音声言語運用の必要条件ではない。」との主張を本稿では展開する（文献1）（文献2）。

そんな馬鹿な、と思われるかもしれない。こんな実験

を考えてみよう。身長300cmの巨人と50cmの小人に孤立母音を発声してもらう。通常音声学の教科書には、F<sub>1</sub>・F<sub>2</sub>

「音声言語は流暢だし雄弁。頭は良いのかもしれない。でも何故か本が読めない、手紙が書けない。そういう成人が米国や英国に多かったりしませんか？ えーと、教育を受けていないとか、そういう事ではなく、彼らの認知特性として文字言語が何故か難しい……」

「先生、デイスレクシアってご存知なんですか？ 特に音韻性のやつ。」

「でいすれ……何ですかそれ？」

「変だな。先生、今、自分でデイスレクシアの説明してたじゃないですか。」

四一年間の人生の中で、あれほど口をあんどぐり開けたことは無い。顎が外れるかと思った。これは実話である。私は彼ら（文献15）の存在を、音声の物理学に基づいて予言していた。

日本語学4月号, p.187-197,  
明治書院(2008)



# “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?

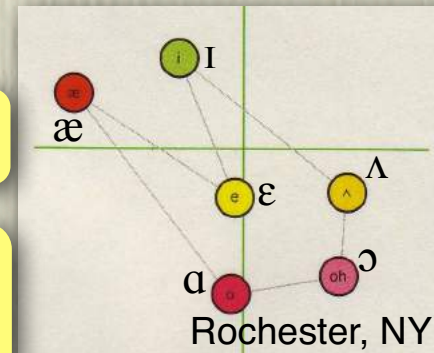
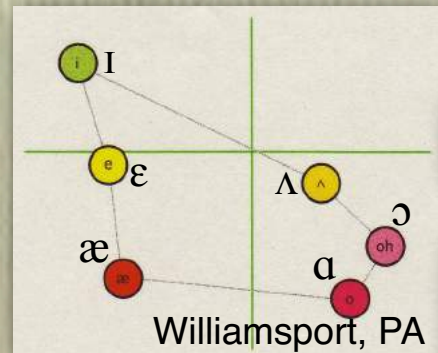


Diff. of VTL = Diff. of timbre



Diff. of regional accents = Diff. of timbre

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

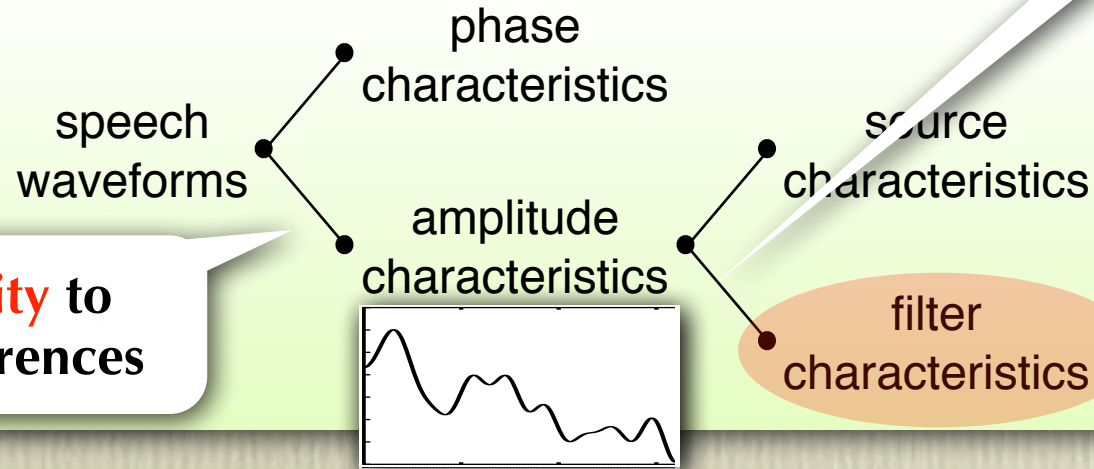
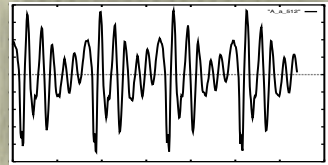




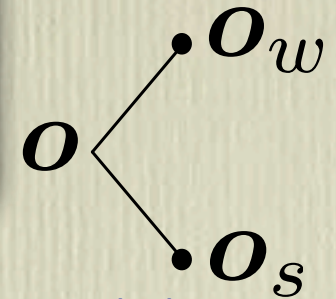
# Feature separation to find specific info.

## 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**

$$o \rightarrow o_w + o_s \text{ is possible or not?}$$



# Inensitivity in our language learning

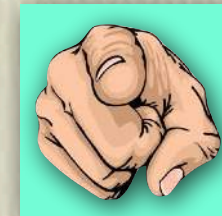
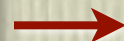
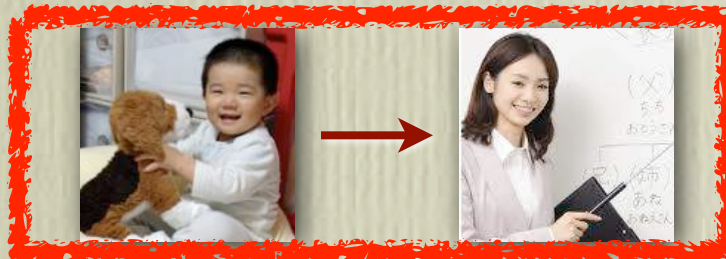
## Vocal learning (including vocal imitation)

- A imitate(s) B vocally.
  - A: students and B: teachers
  - A: infants and B: parents (caretakers)
  - A: you and B: professional singer (Karaoke)
  - But A do not impersonate B.
    - Acoustically *mismatched* imitation.
- We're very insensitive to speaker identity transmitted via speech.



## Acoustically **matched** imitation is often found in

- Autistics (自閉症), who have language disorder [Grandin'96]
- Animals' vocal imitation (birds, dolphins, whales, etc) [Okanoya'08]





# 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

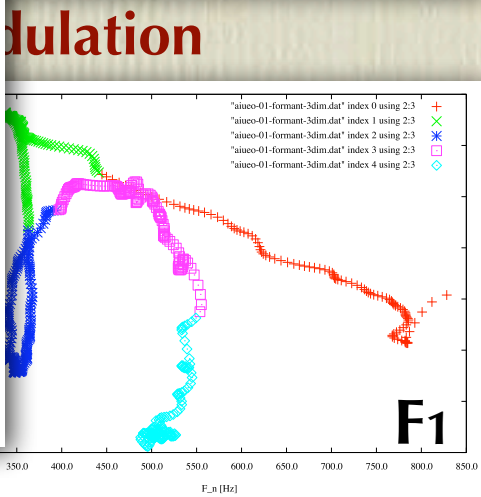
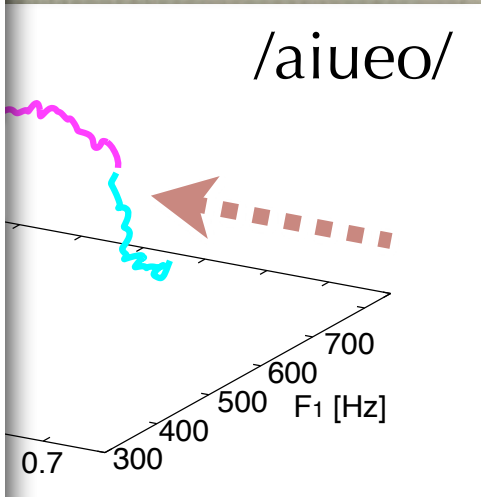
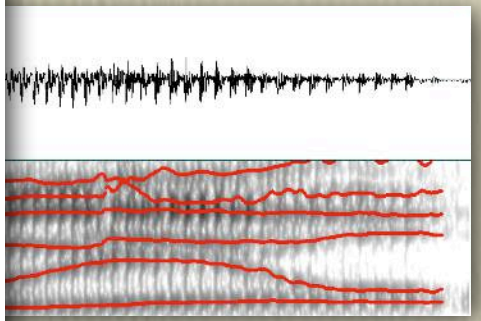
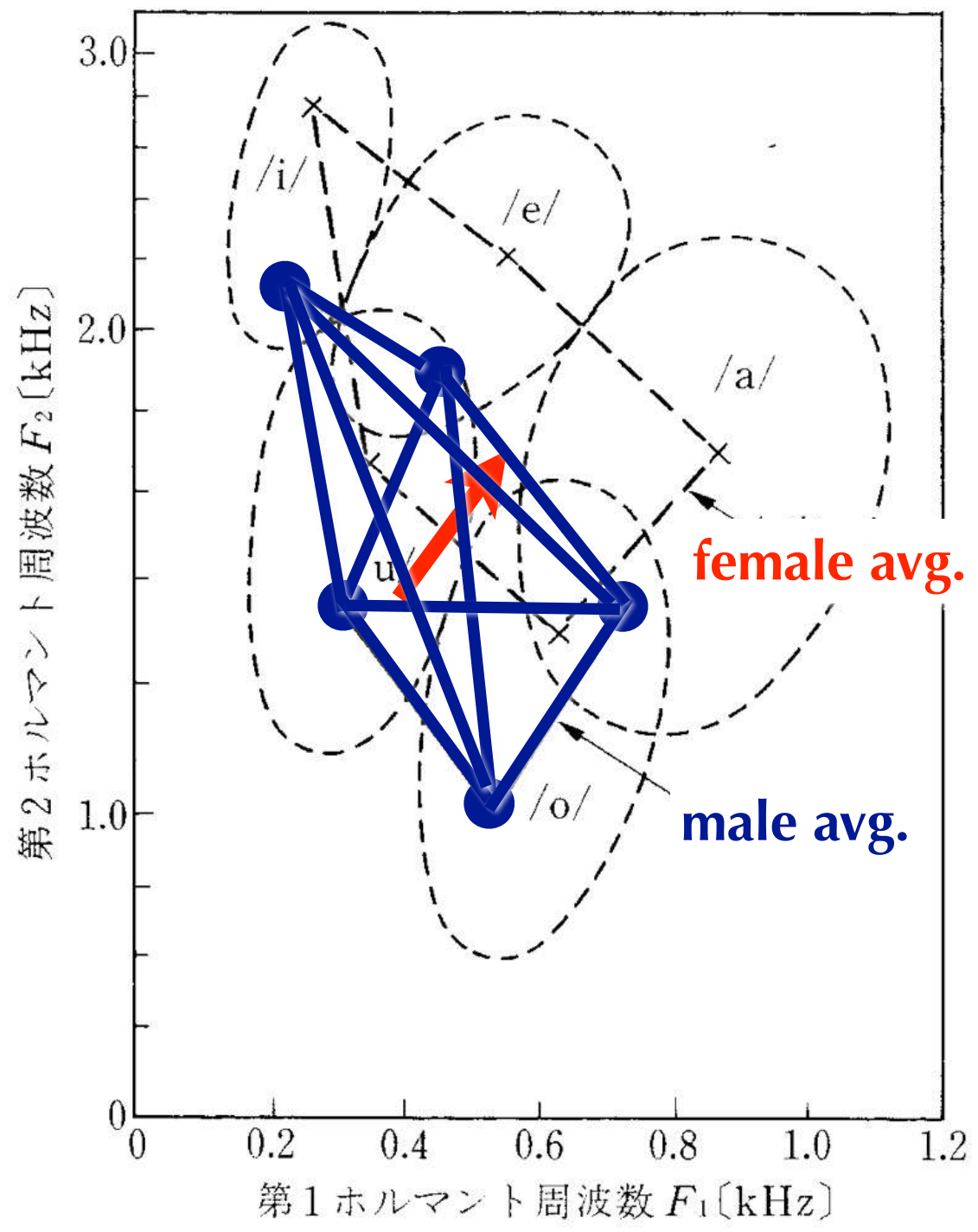
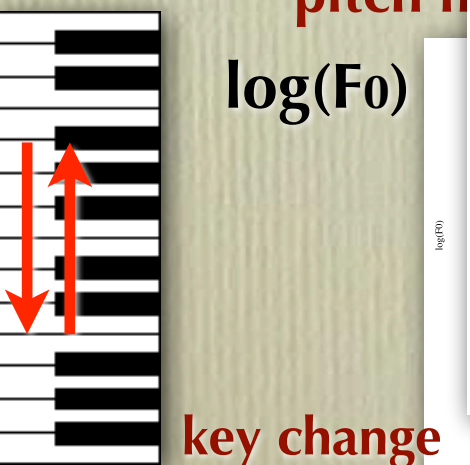
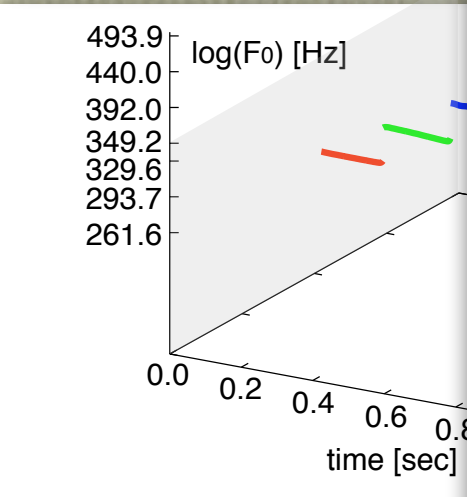
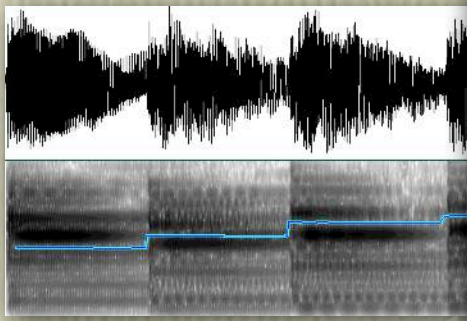
- Completely transform-invariant features --  $f$ -divergence --
- Implementation of word Gestalt as relative timbre perception
- Application of speech structure to robust speech processing

## Radical but interesting discussion

- An interesting link to some behaviors found in language disorder
- An interesting thought experiment



# Relative vowel space in the vocal tract



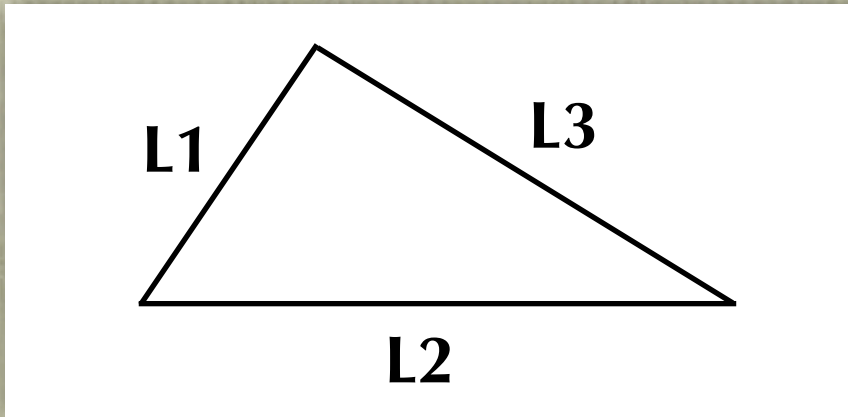
key change

speaker change



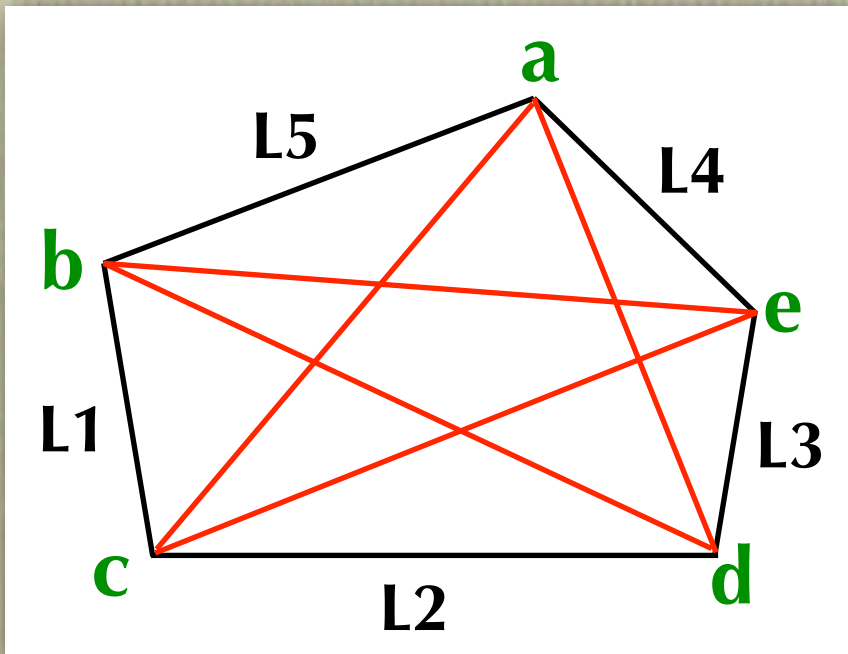
# Definition of the shape of a thing

## Triangle



$(L1, L2, L3)$

## N-point general geometrical structure



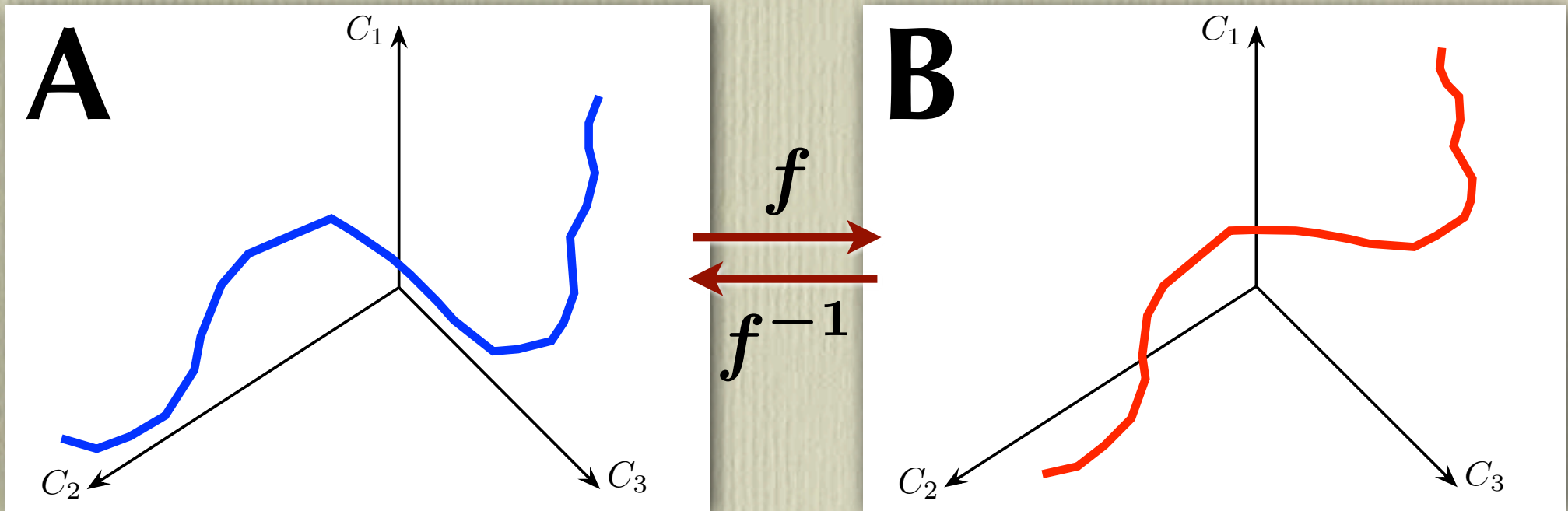
$$\begin{matrix} & \mathbf{a} & \mathbf{b} & & \mathbf{e} \\ \mathbf{a} & d_{11} & d_{12} & \dots & d_{1N} \\ \mathbf{b} & d_{21} & d_{22} & \dots & d_{2N} \\ \mathbf{c} & d_{31} & & & \\ \mathbf{d} & : & & & \\ \mathbf{e} & d_{N1} & d_{N2} & \dots & d_{NN} \end{matrix}$$



# Math. modeling of speaker variability

Speaker difference = mapping of voice spaces

Space of speaker A  $\leftrightarrow$  space of speaker B



Mapping of speaker A into any of 7 billion speakers

7 billion x 7 billion transformations are possible.

Truly speaker-independence = mapping-invariant contrasts

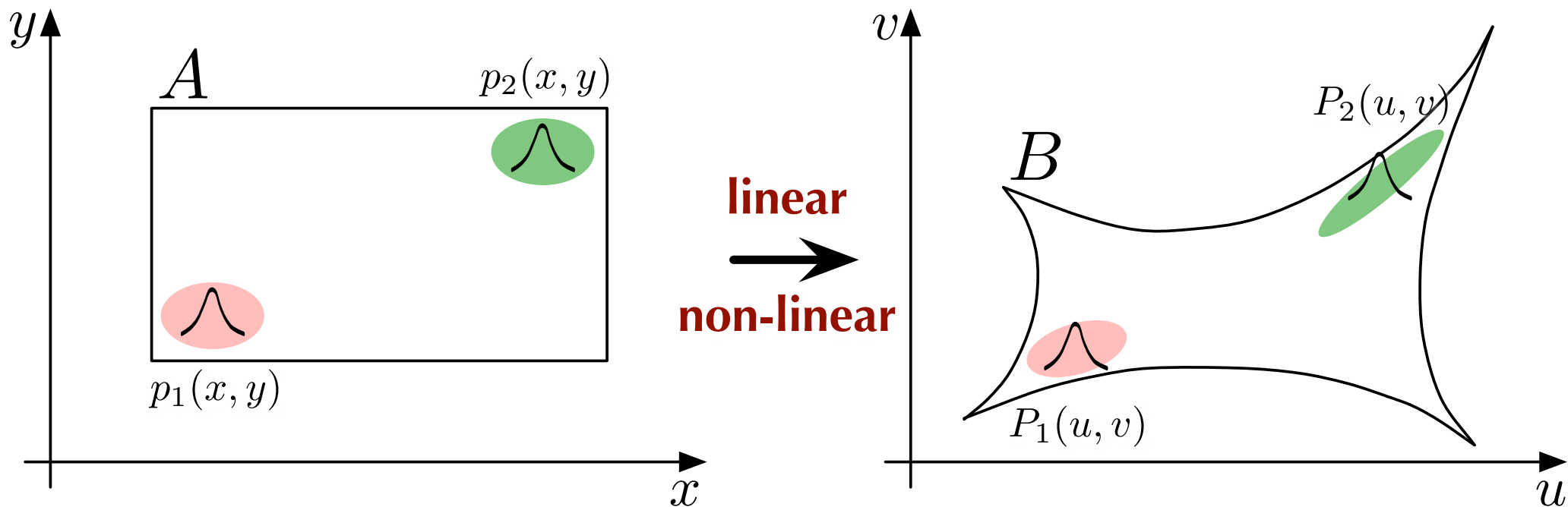
**Are there any contrastive features that are invariant with any mapping?**



# Complete transform-invariance

## Complete invariance between two spaces

- An assumption
  - The transform is convertible and differentiable anywhere.
- An event in a space should be represented as distribution.
  - Event  $p$  in space  $A$  is transformed into event  $P$  in space  $B$
  - $p$  and  $P$  are physically different (/a/ of speaker  $A$  and /a/ of speaker  $B$ )





# Complete transform-invariance

## Variable conversion and integral

- A single variable:  $x = x(t)$  ( $x_1 = x(t_1), x_2 = x(t_2)$ )

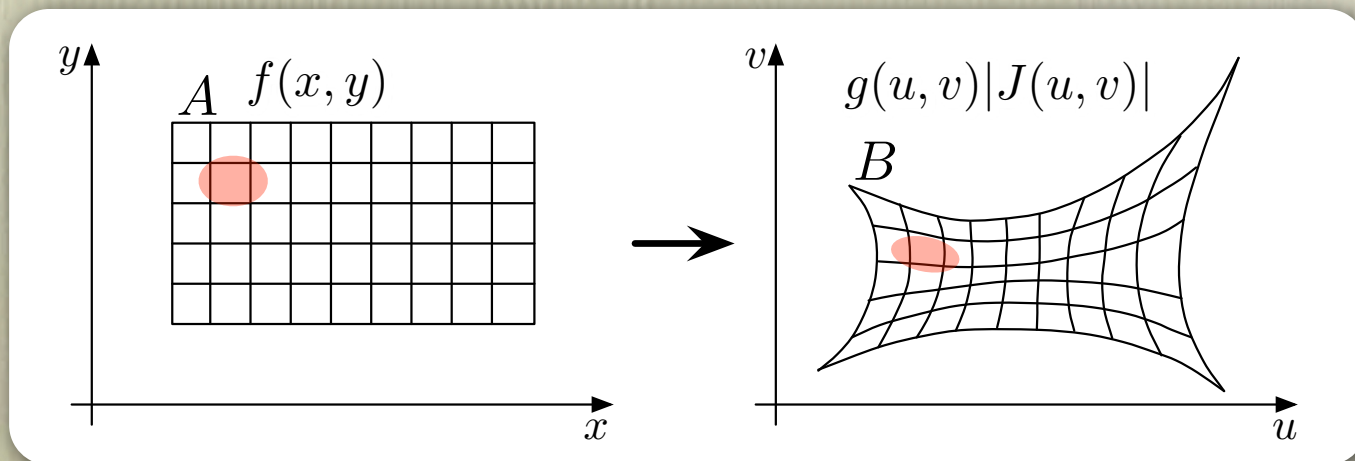
$$\int_{x_1}^{x_2} f(x) dx = \int_{t_1}^{t_2} f(x(t)) \frac{dx(t)}{dt} dt = \int_{t_1}^{t_2} g(t) x'(t) dt$$

- Two variables:  $x = x(u, v), y = y(u, v)$

$$\begin{aligned} x &= 3u + 2v - 5 \\ y &= 4u + 5v + 3 \end{aligned}$$

$$\iint_A f(x, y) dx dy = \iint_B f(x(u, v), y(u, v)) |J(u, v)| du dv$$

$$= \iint_B g(u, v) |J(u, v)| du dv \quad J(u, v) \equiv \frac{\partial(x, y)}{\partial(u, v)} \equiv \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$





# Complete transform-invariance

## Variable conversion and probability density function

- A single variable:  $x = x(t)$  ( $x_1 = x(t_1), x_2 = x(t_2)$ )

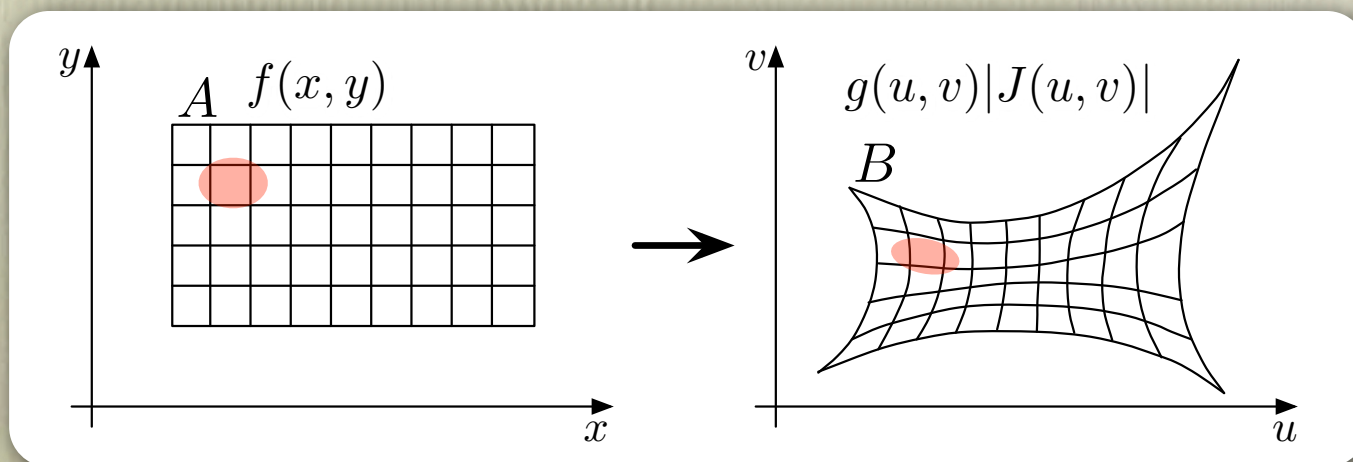
$$1.0 = \int_{x_1}^{x_2} p(x) dx = \int_{t_1}^{t_2} p(x(t)) \frac{dx(t)}{dt} dt = \int_{t_1}^{t_2} \underline{q(t)x'(t)} dt$$

- Two variables:  $x = x(u, v), y = y(u, v)$

$$\begin{aligned} x &= 3u + 2v - 5 \\ y &= 4u + 5v + 3 \end{aligned}$$

$$1.0 = \iint_A f(x, y) dx dy = \iint_B f(x(u, v), y(u, v)) |J(u, v)| du dv$$

$$= \iint_B \underline{g(u, v) |J(u, v)|} du dv \quad J(u, v) \equiv \frac{\partial(x, y)}{\partial(u, v)} \equiv \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$





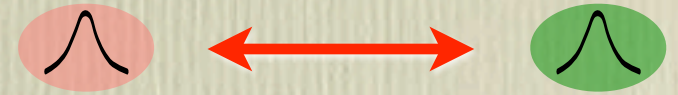
# Complete transform-invariance

## Bhattacharyya distance

One of the distance measures bet. two distributions

$x = x(u, v), y = y(u, v)$

$BD(p_1(x, y), p_2(x, y))$



$$= -\log \iint \sqrt{p_1(x, y)p_2(x, y)} dx dy$$

$$= -\log \iint \sqrt{q_1(u, v)q_2(u, v)} |J(u, v)| dx dy$$

$$= -\log \iint \sqrt{q_1(u, v) |J(u, v)| \cdot q_2(u, v) |J(u, v)|} du dv$$

$$= -\log \iint \sqrt{P_1(u, v)P_2(u, v)} du dv$$

$$= BD(P_1(u, v), P_2(u, v))$$

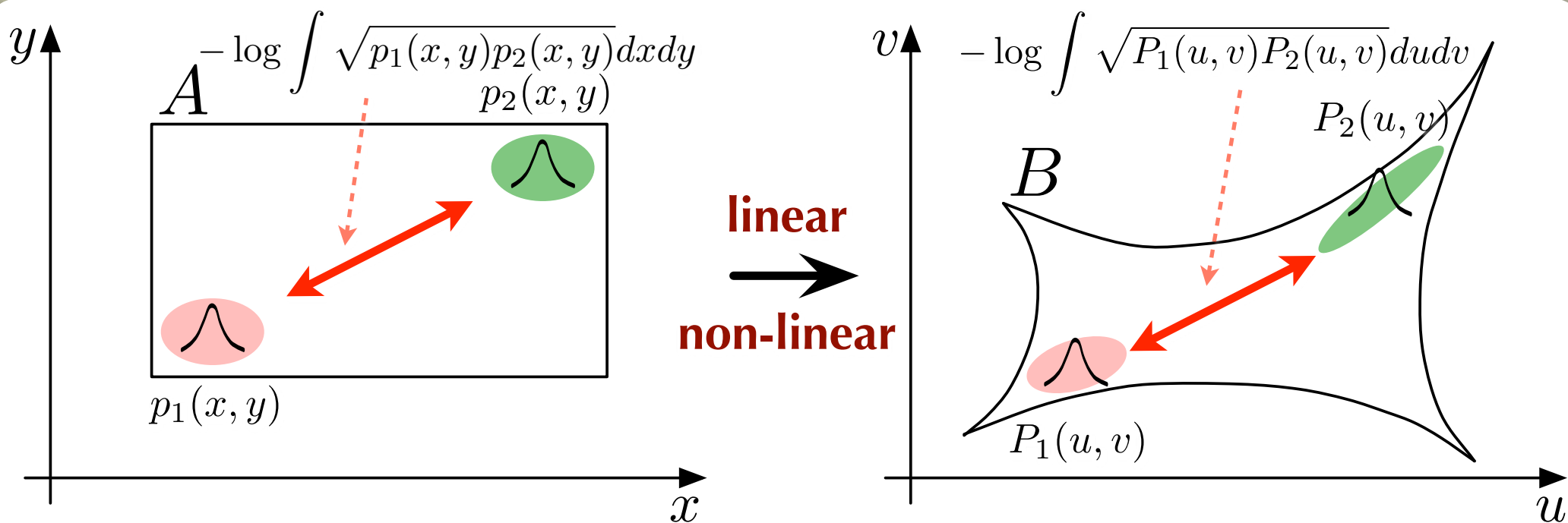
$$q_1(u, v) = p_1(x(u, v), y(u, v)), \quad J = \text{Jacobian}$$



# Complete transform-invariance

## Complete invariance between two spaces

- An assumption
  - The transform is convertible and differentiable anywhere.
- An event in a space should be represented as distribution.
  - Event  $p$  in space  $A$  is transformed into event  $P$  in space  $B$
  - $p$  and  $P$  are physically different (/a/ of speaker  $A$  and /a/ of speaker  $B$ )





# 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.

- $f_{div}(p_1, p_2) = \int p_2(\mathbf{x}) g\left(\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}\right) d\mathbf{x}$

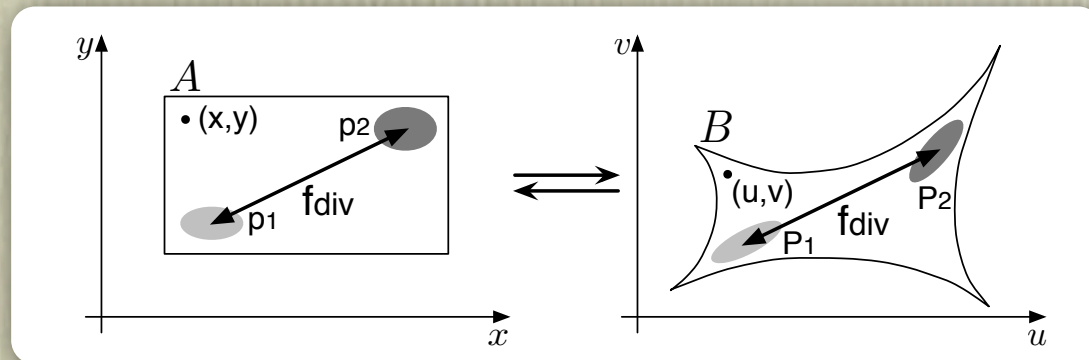
- $g(t) = t \log(t) \rightarrow f_{div} = \text{KL} - \text{div.}$        $g(t) = \sqrt{t} \rightarrow -\log(f_{div}) = \text{BD}$

- $f_{div}(p_1, p_2) = f_{div}(P_1, P_2)$

- Invariant features have to be f-divergence.

- If  $\int M(p_1(\mathbf{x}), p_2(\mathbf{x})) d\mathbf{x}$  is invariant with any transformation,

- The following condition has to be satisfied.  $M = p_2(\mathbf{x}) g\left(\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}\right)$

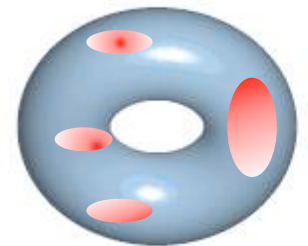
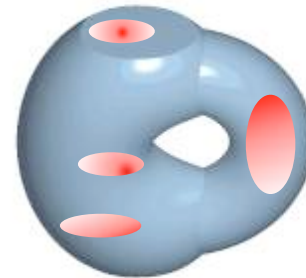
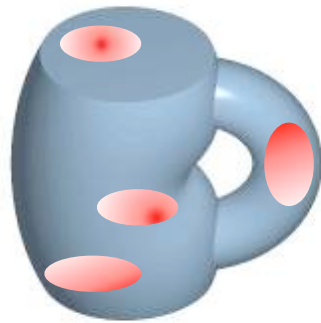
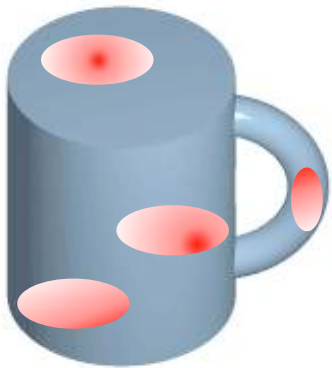




# Invariance in variability

## Topological invariance [Minematsu'09]

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

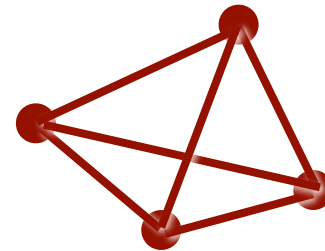
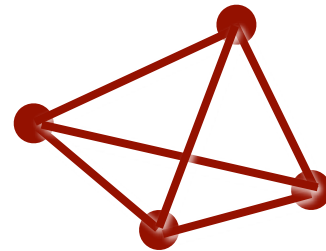
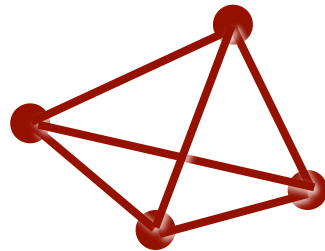
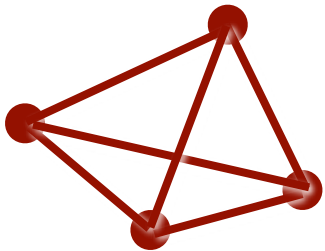
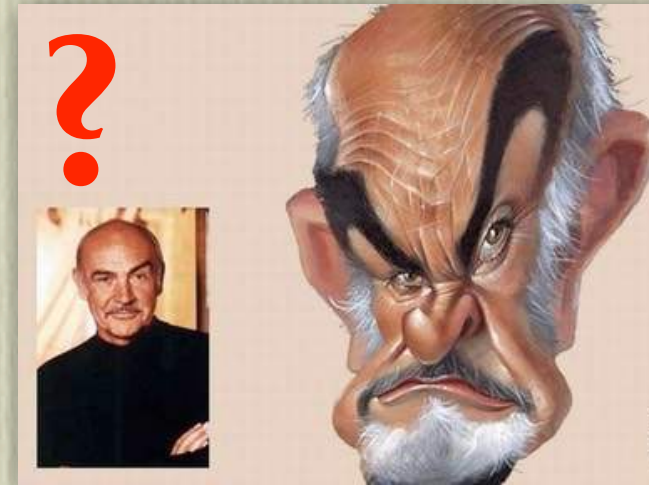




# Invariance in variability

## Topological invariance [Minematsu'09]

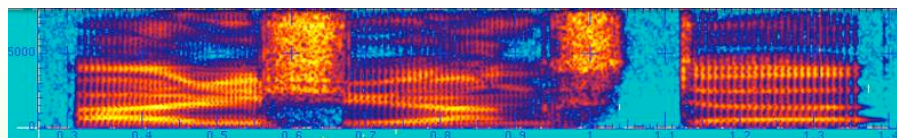
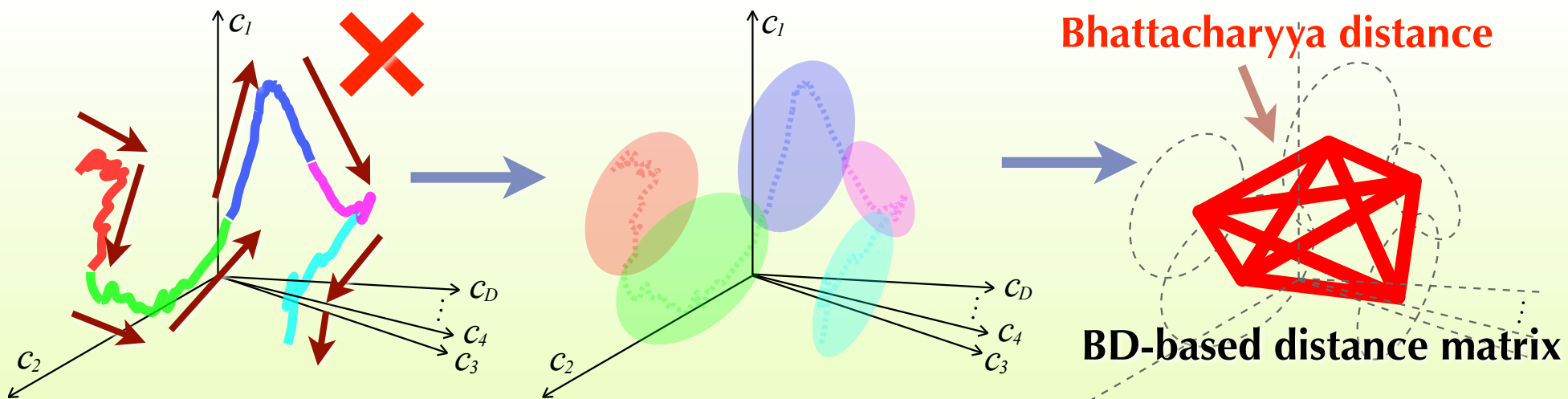
- Topology focuses on invariant features wrt. any kind of deformation.



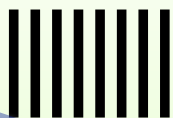


# Invariant speech structure

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



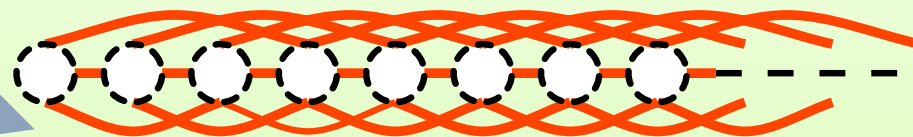
spectrogram (spectrum slice sequence)



cepstrum vector sequence



distribution sequence

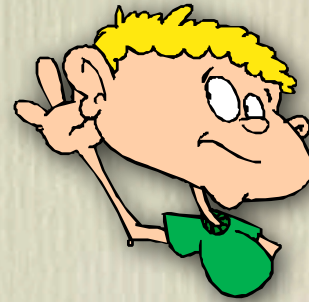
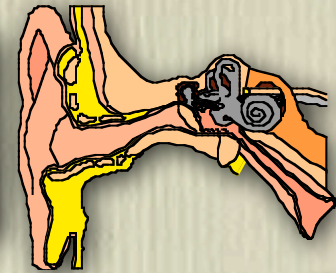


An event (distribution) may be smaller than a phoneme.



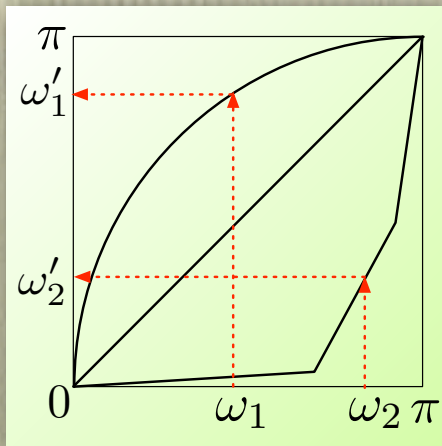
# Speech modification by VTLD

## Speech modification by non-linguistic factors

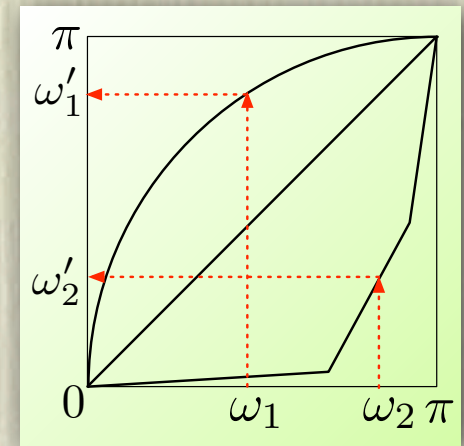
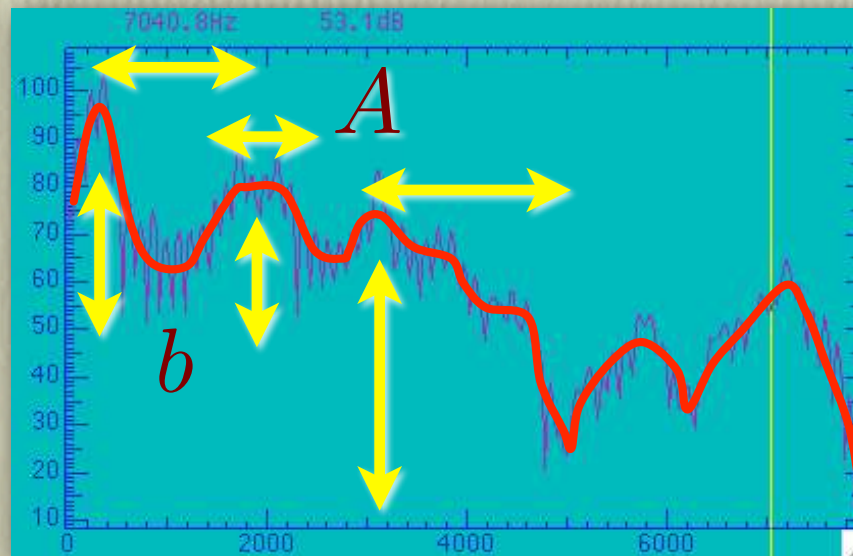


$$\times H(s)$$

$$c' = c + b$$



$$c' = Ac$$



$$c' = Ac$$

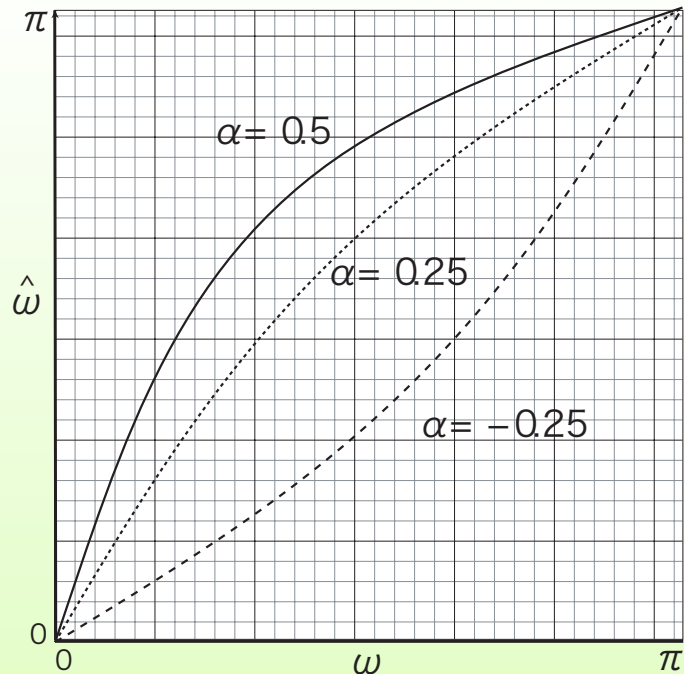


# VTL-based variation = $\times$ matrix A

## Vocal tract length variation

Can be approximated as multiplication of matrix A in cep. domain.

**A is represented as warping parameter  $\alpha$ .**



$$\hat{\mathbf{c}} = (\hat{c}_1 \hat{c}_2 \hat{c}_3 \hat{c}_4 \dots)^t$$

$$\mathbf{A} = \begin{pmatrix} 1 - \alpha^2 & 2\alpha - 2\alpha^3 & \dots & \dots \\ -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$$\mathbf{c} = (c_1 c_2 c_3 c_4 \dots)^t$$

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=\max(0, j-i)}^j \binom{j}{m} \times \frac{(m+i-1)!}{(m+i-j)!} (-1)^m \alpha^{(2m+i-j)}$$

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad z = e^{j\omega}, \quad \hat{z} = e^{j\hat{\omega}}$$

$$\mathbf{c}' = \mathbf{A}\mathbf{c}$$



# Geometrical characteristics of A

$$\begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

$$T = R + O$$

$$R = \begin{pmatrix} 1-2\alpha^2 & 2\alpha(1-\frac{1}{2}\alpha^2) \\ -2\alpha(1-\frac{1}{2}\alpha^2) & 1-2\alpha^2 \end{pmatrix}$$

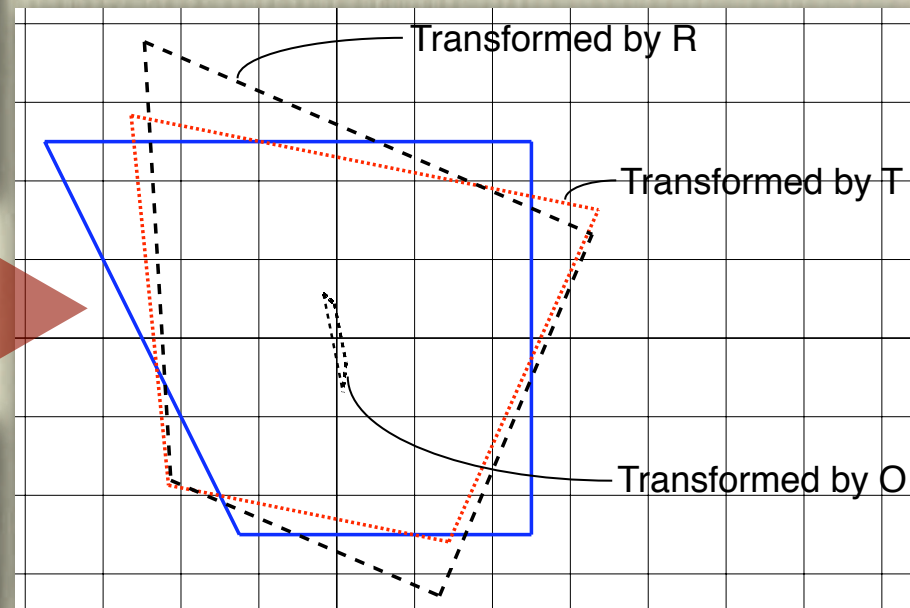
$$O = \begin{pmatrix} \alpha^2 & -\alpha^3 \\ -\alpha & -2\alpha^2+3\alpha^4 \end{pmatrix}.$$

$$A = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \dots & \dots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$



$$R \simeq \begin{pmatrix} 1-2\alpha^2 & 2\alpha\sqrt{1-\alpha^2} \\ -2\alpha\sqrt{1-\alpha^2} & 1-2\alpha^2 \end{pmatrix}$$

$$= \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{pmatrix} \quad (\alpha = \sin \theta)$$



Is it the case in N dimensions?



# Geometrical characteristics of A

What is the rotation matrix in an N dimensional space?

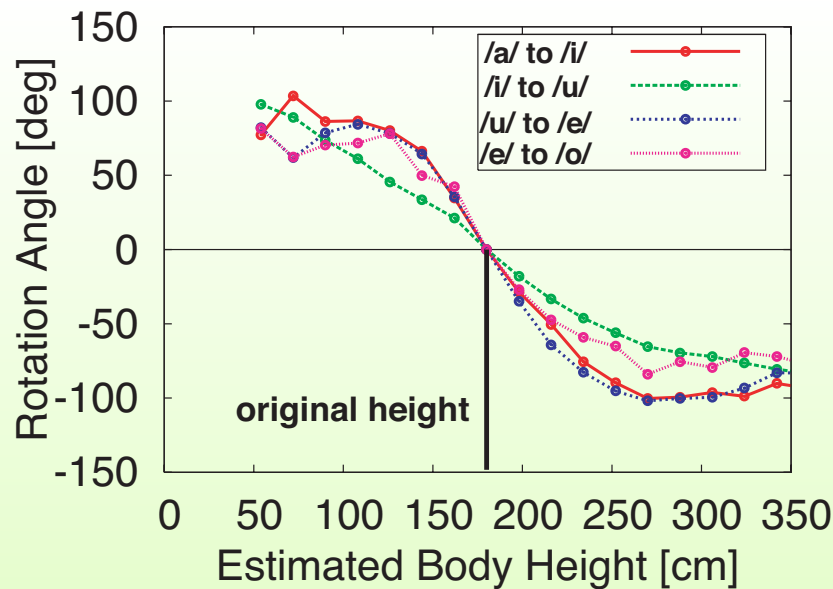
$$R^t R = R R^t = I$$

$$\det R = +1.$$

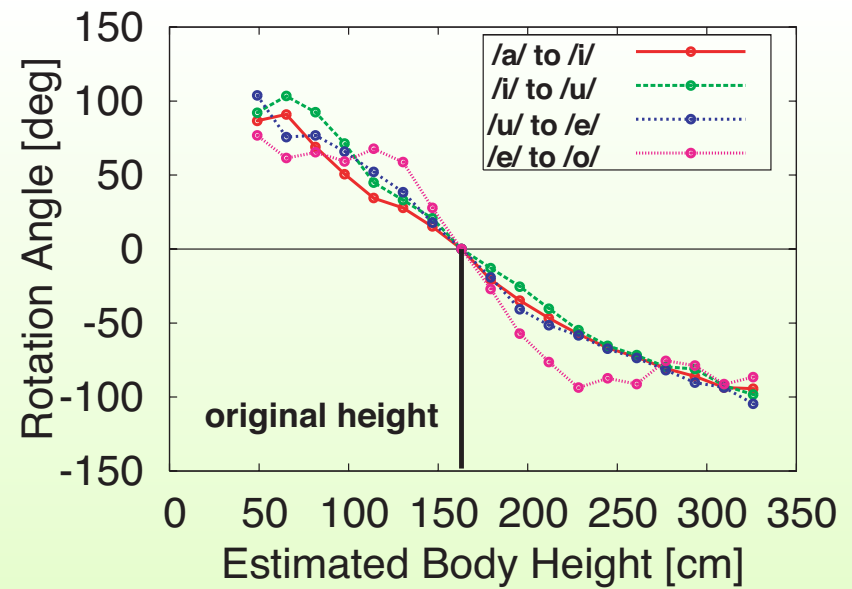
$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=\max(0, j-i)}^j \binom{j}{m} \times \frac{(m+i-1)!}{(m+i-j)!} (-1)^m \alpha^{(2m+i-j)}$$

satisfied approximately.

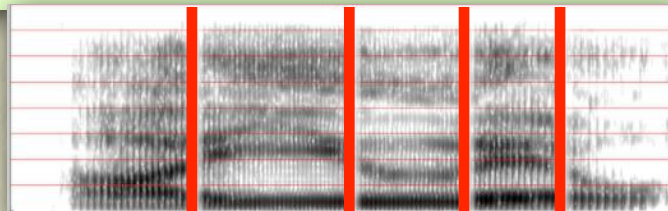
Frequency warping can rotate any cepstrum trajectory.



(a):MFCC (male)



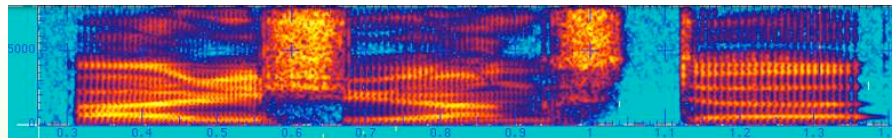
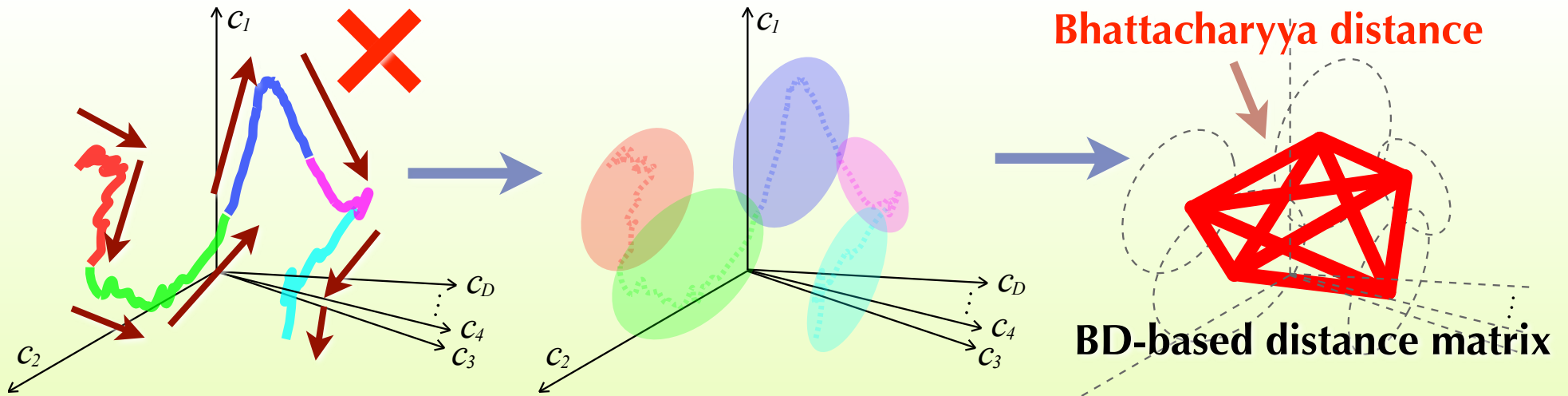
(d):MFCC (female)





# Invariant speech structure

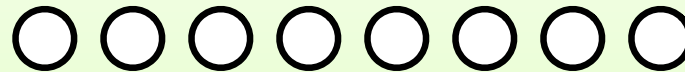
Utterance to structure conversion using  $f$ -div. [Minematsu'06]



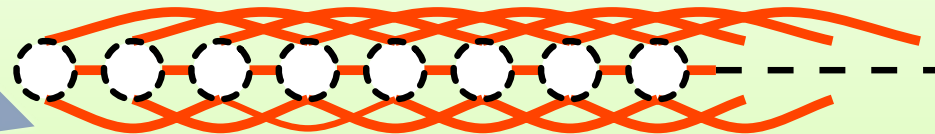
spectrogram (spectrum slice sequence)



cepstrum vector sequence



distribution sequence



An event (distribution) may be smaller than a phoneme.

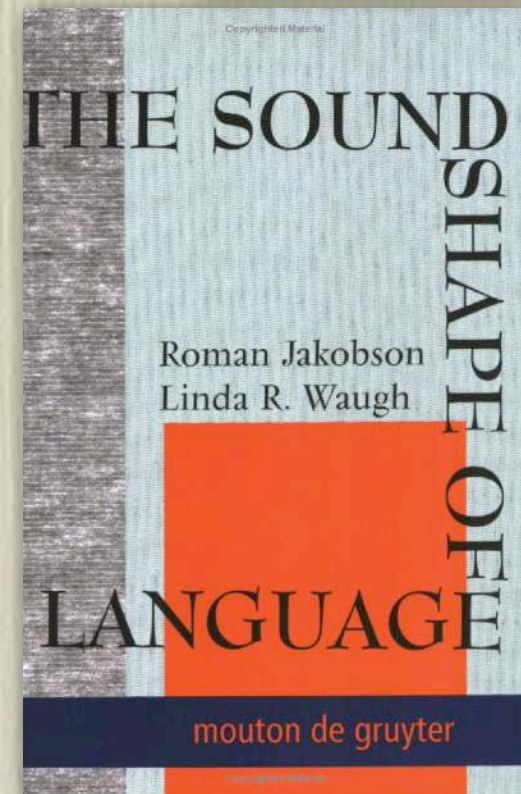
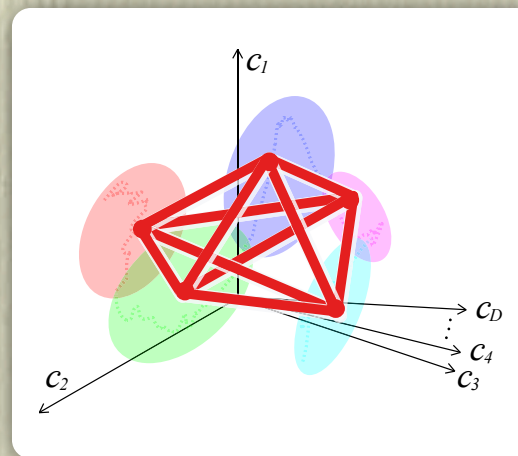
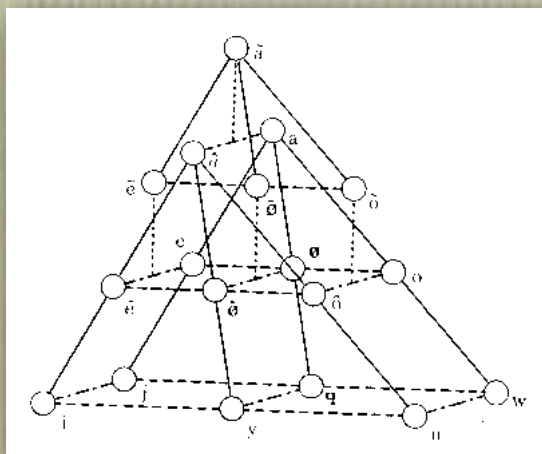
# A claim found in classical linguistics

## 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.

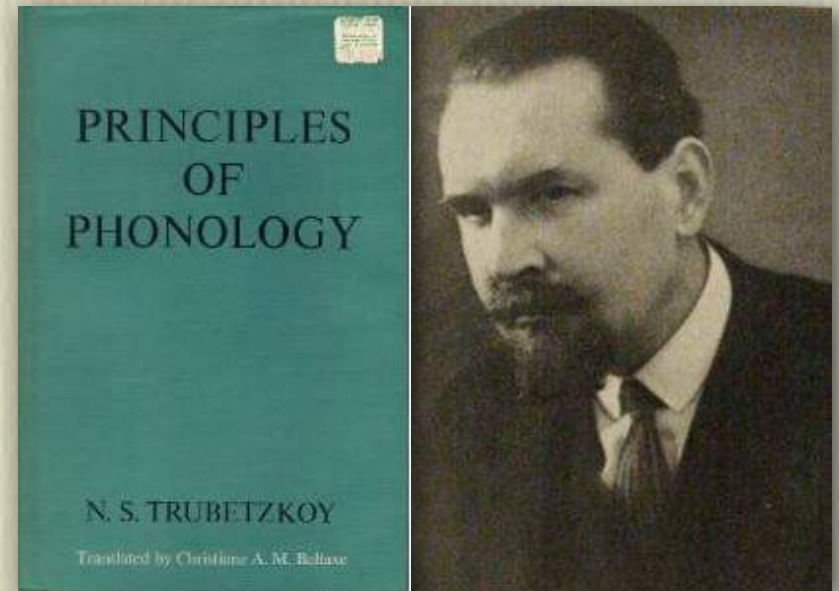
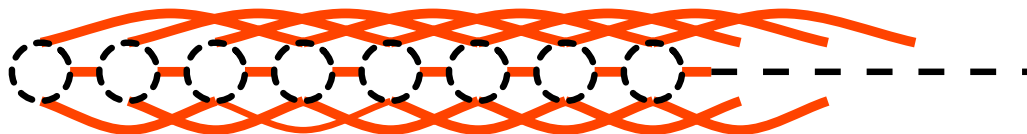
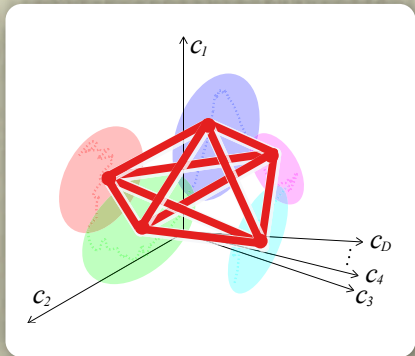




# More classical claims in linguistics

## Nikolay Sergeevich Trubetskoj (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.

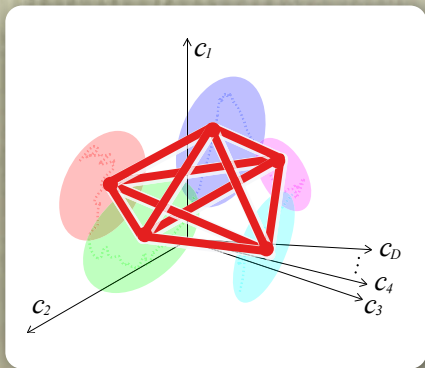




# More classical claims in linguistics

## 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.



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

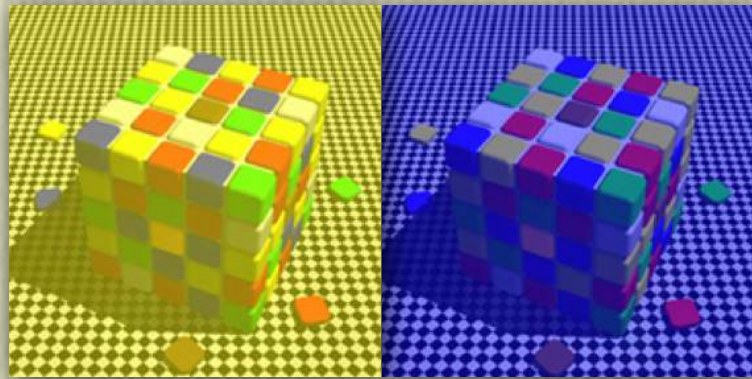




# Invariant **timbre** perception against its bias

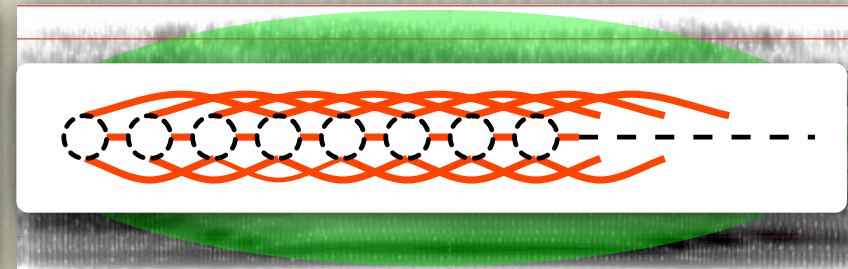
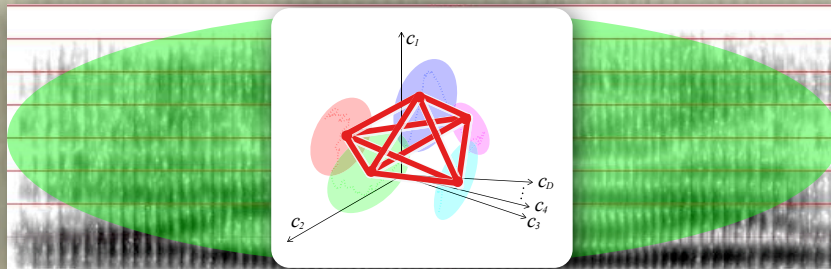
## Invariant and constant perception wrt. **color and pitch**

- **Contrast-based** information processing is important.
- **Holistic & relational** processing enables **element** identification.



## Invariant and constant perception wrt. **timbre**

- **Contrast-based** information processing is important.
- **Holistic & relational** processing enables **element** identification.





# 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

- An interesting link to some behaviors found in language disorder
- An interesting thought experiment



# !! Note !!

**Next Tuesday (Dec 26) is a day for Friday classes.**

**The next lecture will be held on Dec 25 (Monday).**

- Dec 25 is a day for supplementary lectures.

- From 13:00 to 14:45 in room 246.

# Dec 25 (Mon)

# 13:00-14:45