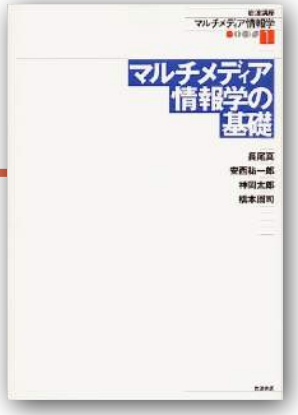
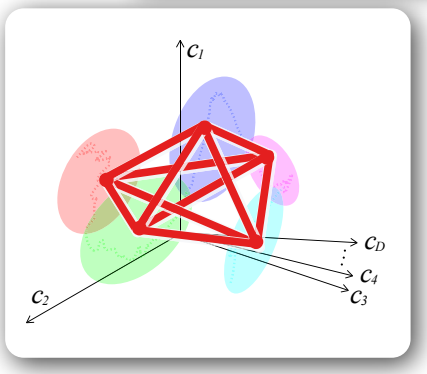
Cognitive Media Processing #9

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



Aim of this class

- Syllabus on the web
  - Cognitive processing of multimedia information by humans and its technical processing by machines are explained and compared. Then, a focus is placed on the fact that a large difference still remains between them. This lecture will enable students to consider deeply what kind of information processing is lacking on machines and has to be implemented on them if students want to create not seemingly but actually “human-like” robots, especially the robots that can understand spoken language.
  - The lectures are divided into three parts. The first part explains the multimedia information processing by human brains. Here, some interesting perceptual characteristics of individuals with autism(自閉症) and synesthesia(共感覚) are shown as examples. The second part describes the conventional technical framework of spoken language processing. The last discusses drawback of the current framework and what kind of new methodology is needed to create really “human-like” robots that can understand spoken language. Then, a possible new framework is introduced and explained.

A new framework for “human-like” speech machines #1

Nobuaki Minematsu

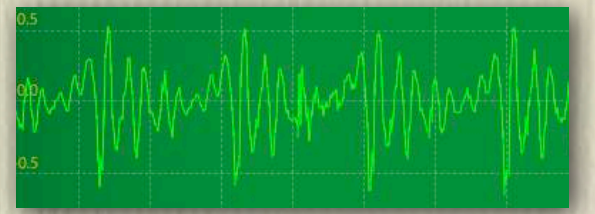
Information that speech can transmit

Three kinds of information

- Linguistic
- Para-linguistic
- Extra-linguistic (non-linguistic)

Speech

- Waveforms, just a sequence of numbers
  - -23, -89, -127, -40, 9, 46, 189, 242, 212, 183, ...



Two major speech applications

- Speech recognition
  - Extraction of linguistic info. from a number sequence
  - Large extra-linguistic variation in speech acoustics is a major problem.
- Speech synthesis
  - Conversion of linguistic info. $+\alpha$ to a number sequence

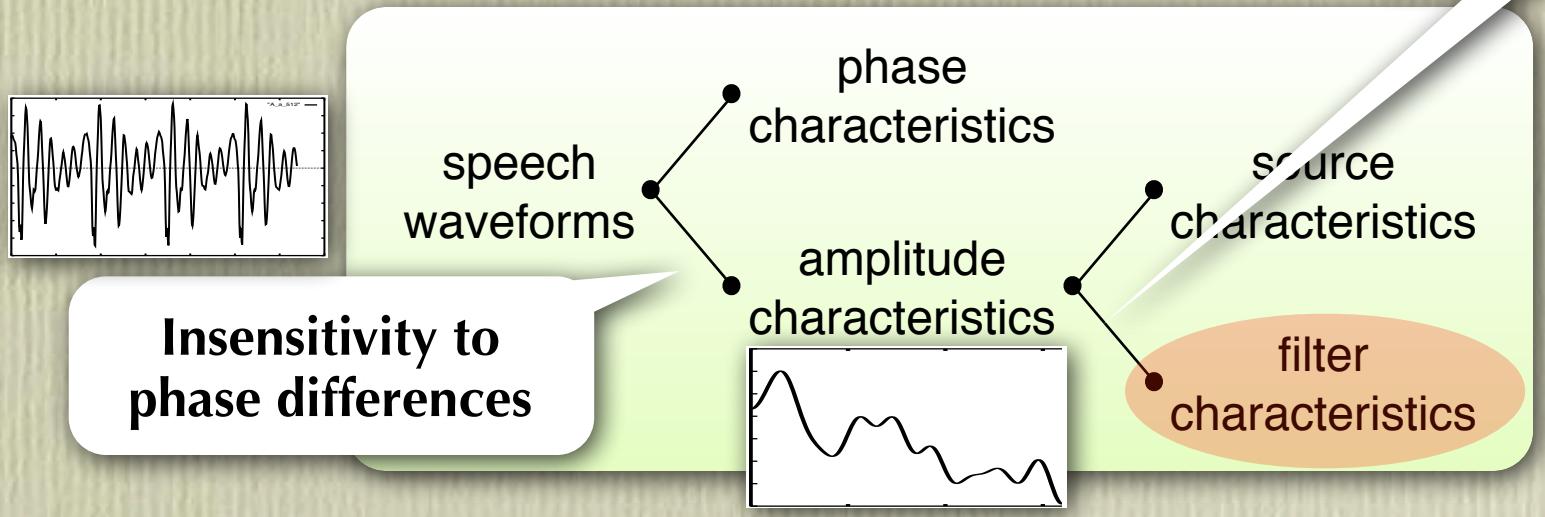
Speech is extremely variable.

Various factors change speech acoustics easily.

The world's tiniest high school girl

Feature separation to find specific info.

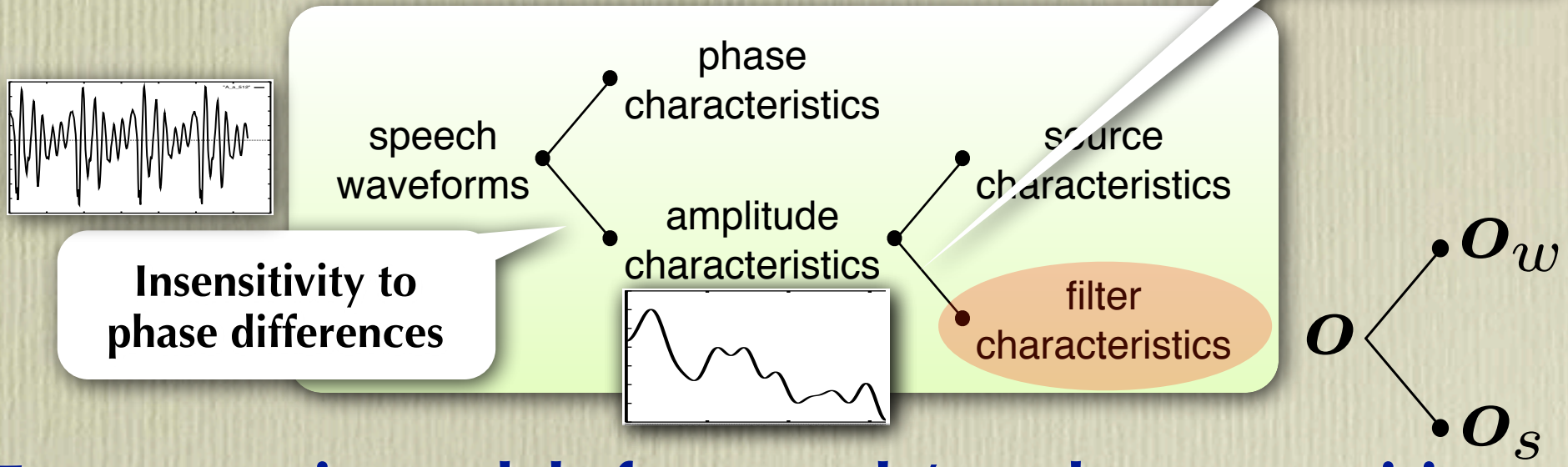
De facto standard acoustic analysis of s



- Spectrum envelope-based feature such as CEP: o
  - But o depends on all the three kinds of info. (ling, para-ling, extra-ling).
- How to suppress extra-linguistic variation in o ?
  - Feature normalization: transforming o to that of the standard speaker
  - Model adaptation: modifying model parameters to fit to the input speaker
  - Statistical independence: hiding these variation through sample collection
  - Physical independence: pursuing features invariant to these variation
  - :

Feature separation to find specific info.

De facto standard acoustic analysis of s

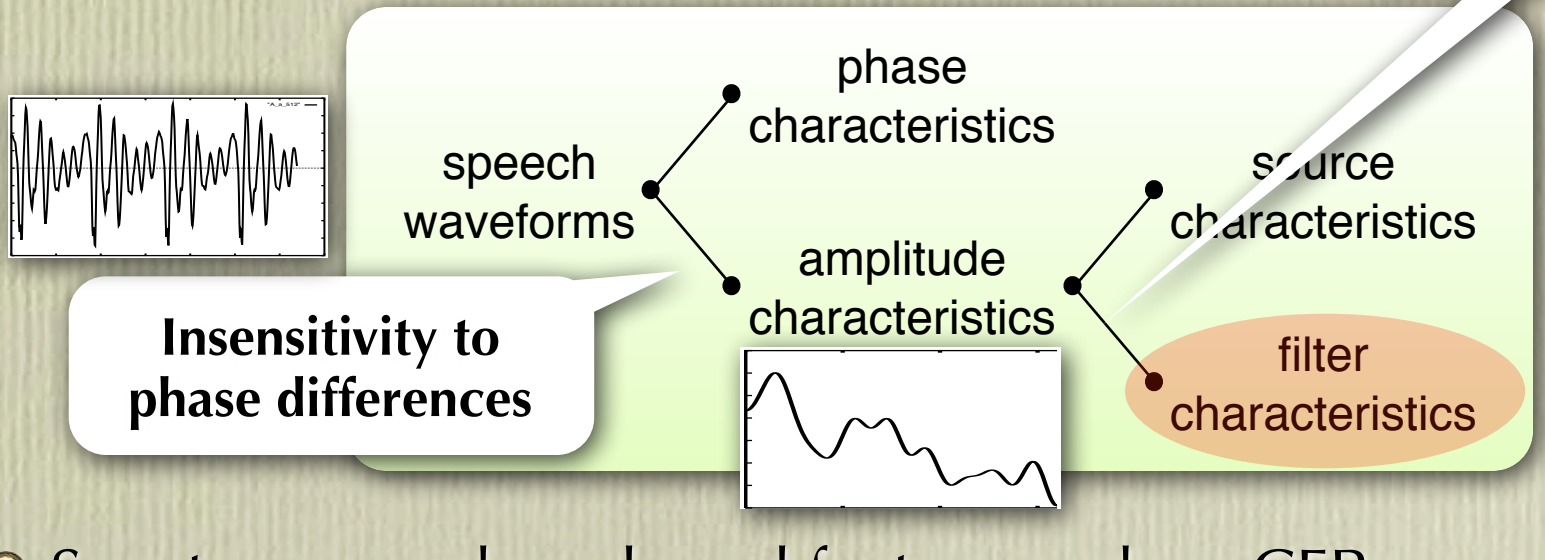


Two acoustic models for speech/speaker recognition

- 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?

Feature separation to find specific info.

De facto standard acoustic analysis of s



- Spectrum envelope-based feature such as CEP: o
  - But o depends on all the three kinds of info. (ling, para-ling, extra-ling).
- How to suppress extra-linguistic variation in o ?
  - Feature normalization: transforming o to that of the standard speaker
  - Model adaptation: modifying model parameters to fit to the input speaker
  - Statistical independence: hiding these variation through sample collection
  - Physical independence: pursuing features invariant to these variation
  - :

A difference bet. machines and humans

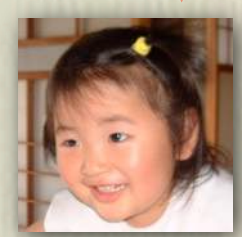
Machine strategy (engineers' strategy): ASR

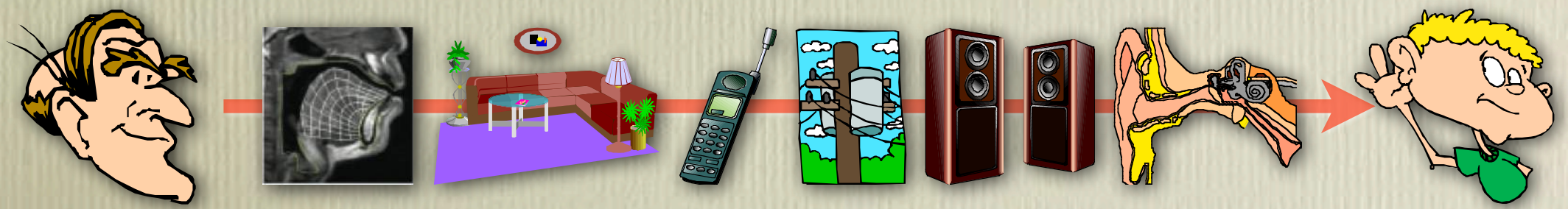
- Collecting a huge amount of speaker-balanced data
  - Statistical training of acoustic models of individual phonemes (allophones)
- Adaptation of the models to new environments and speakers
  - Acoustic mismatch bet. training and testing conditions must be reduced.

?

Human strategy: HSR

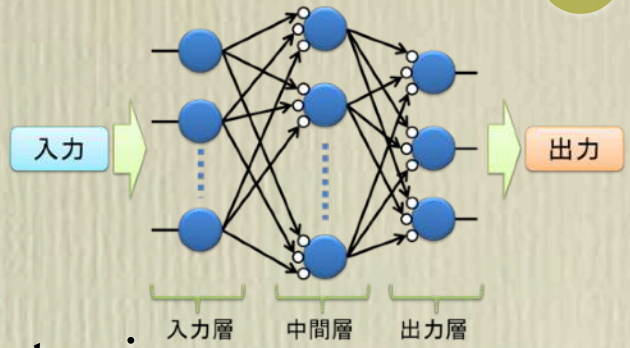
- A major part of the utterances an infant hears are from its parents.
  - The utterances one can hear are extremely speaker-biased.
- Infants don't care about the mismatch in lang. acquisition.
  - Their vocal imitation is not acoustic, it is not impersonation!!





What is the common denominator?

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.
  - More abstract features with extra-linguistic information removed?
- Still difficult to interpret structure and weights of DNN physically.
  - Interpretable DNNs are becoming one of the hot topics [Sim'15].

Simple questions raised by researchers

- "What are ***really*** speaker-independent features?" [Morgan'12, '13]
- "What is ***the common denominator*** bet. speakers?" [Jakobson'79]

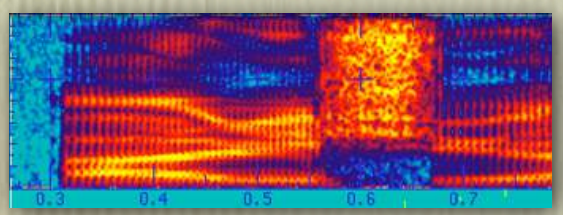
A claim found in classical linguistics

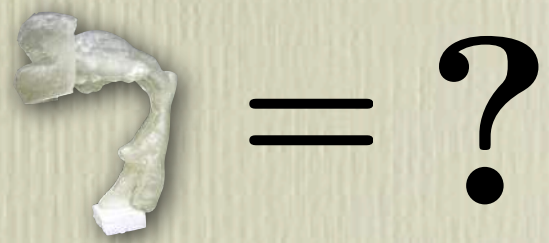
Theory of relational invariance [Jakobson+'79]

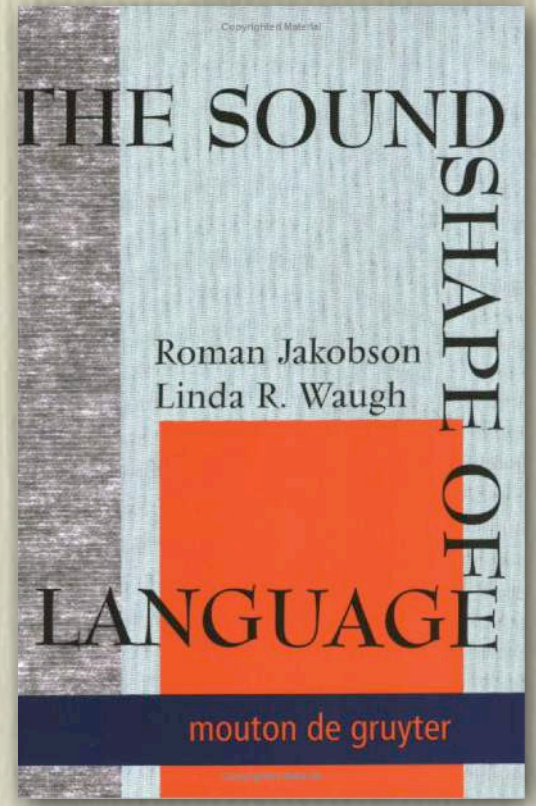
- Also known as theory of distinctive features
- Proposed by R. Jakobson

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.







A difference bet. machines and humans

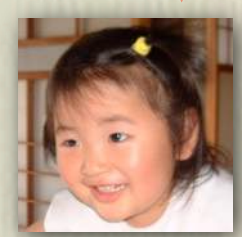
Machine strategy (engineers' strategy): ASR

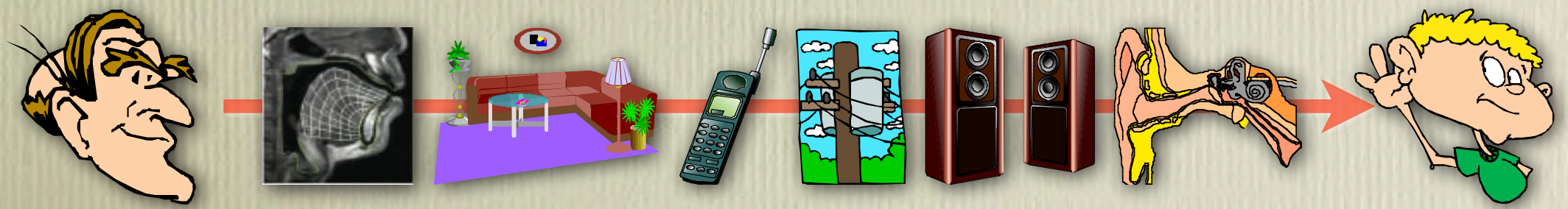
- Collecting a huge amount of speaker-balanced data
  - Statistical training of acoustic models of individual phonemes (allophones)
- Adaptation of the models to new environments and speakers
  - Acoustic mismatch bet. training and testing conditions must be reduced.

?

Human strategy: HSR

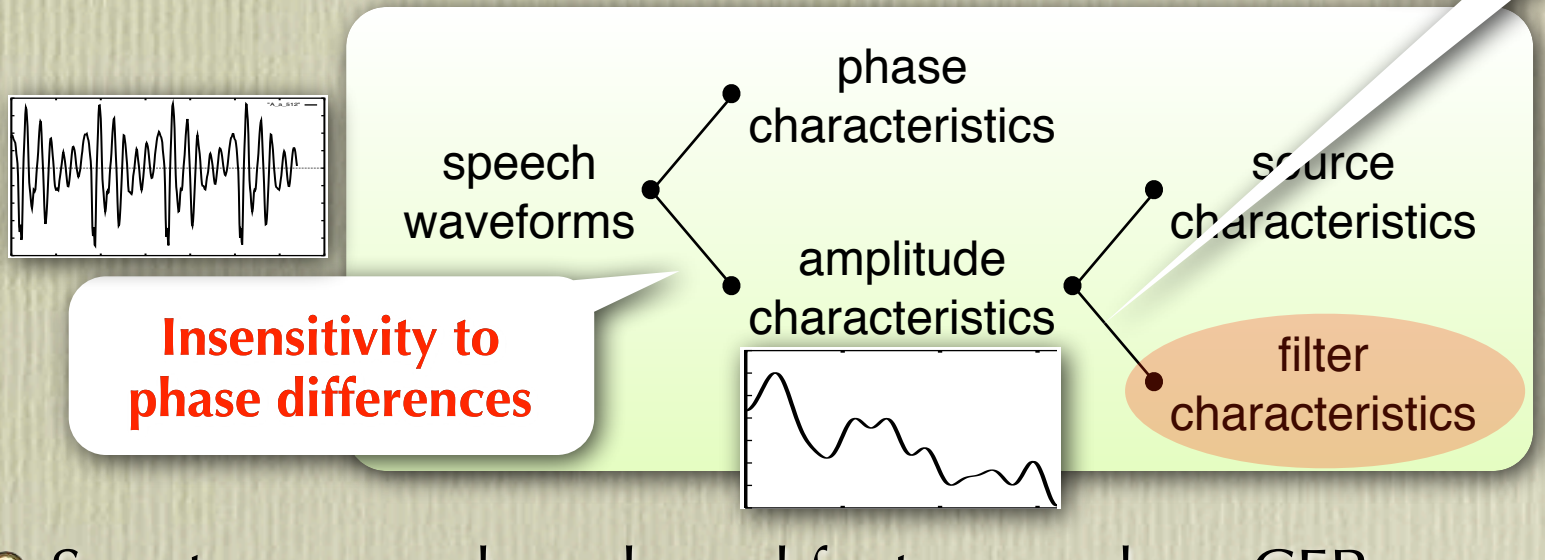
- A major part of the utterances an infant hears are from its parents.
  - The utterances one can hear are extremely speaker-biased.
- Infants don't care about the mismatch in lang. acquisition.
  - Their vocal imitation is not acoustic, it is not impersonation!!





Feature separation to find specific info.

De facto standard acoustic analysis of s



- Spectrum envelope-based feature such as CEP: o
  - But o depends on all the three kinds of info. (ling, para-ling, extra-ling).
- How to suppress extra-linguistic variation in o ?
  - Feature normalization: transforming o to that of the standard speaker
  - Model adaptation: modifying model parameters to fit to the input speaker
  - Statistical independence: hiding these variation through sample collection
  - Physical independence: pursuing features invariant to these variation
  - :

Insensitivity 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 *mis*matched imitation.

- We're very insensitive to speaker identity transmitted via speech.

Acoustically matched imitation is found in

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





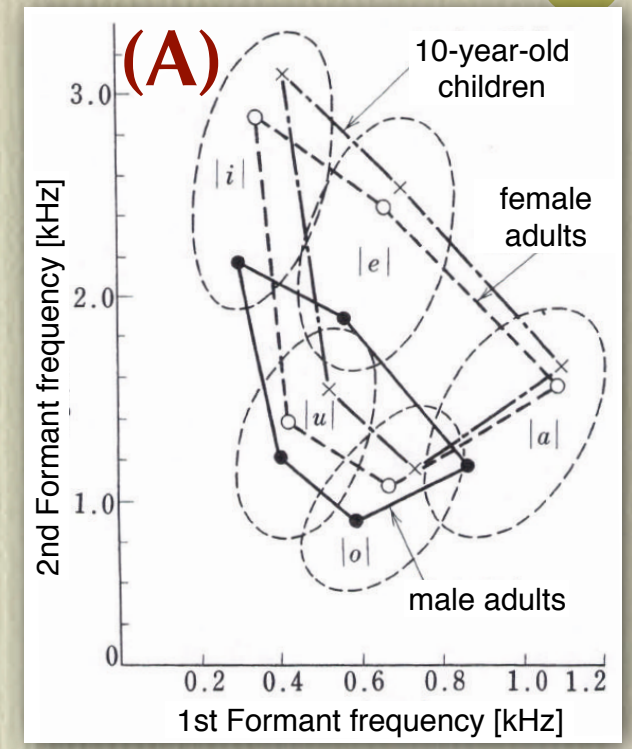

Insensitivity and sensitivity

Infants' vocal learning is

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

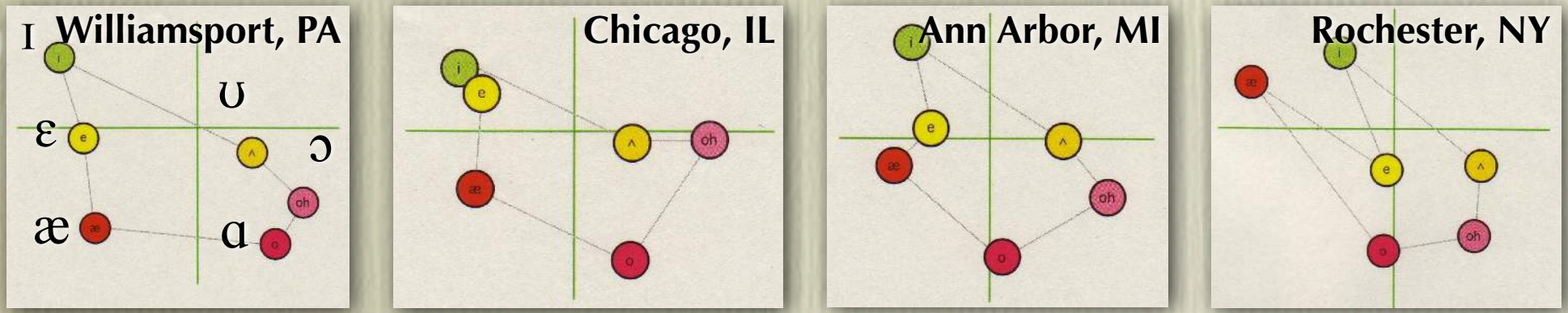
Infants' vocal learning seems to be

- 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

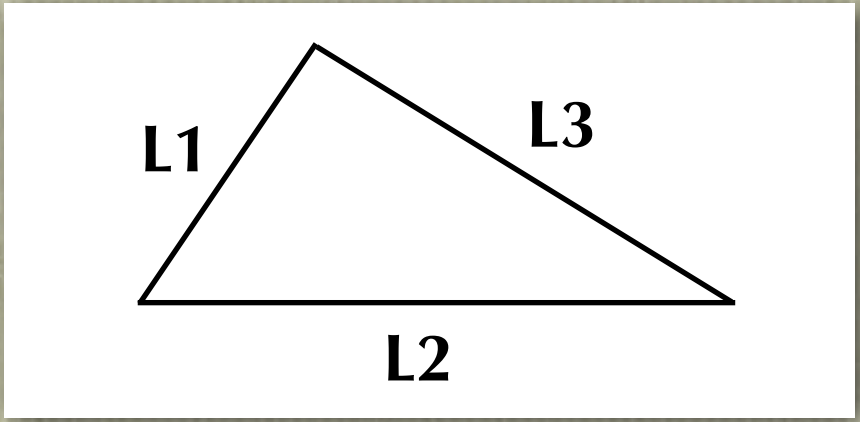
(B)



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

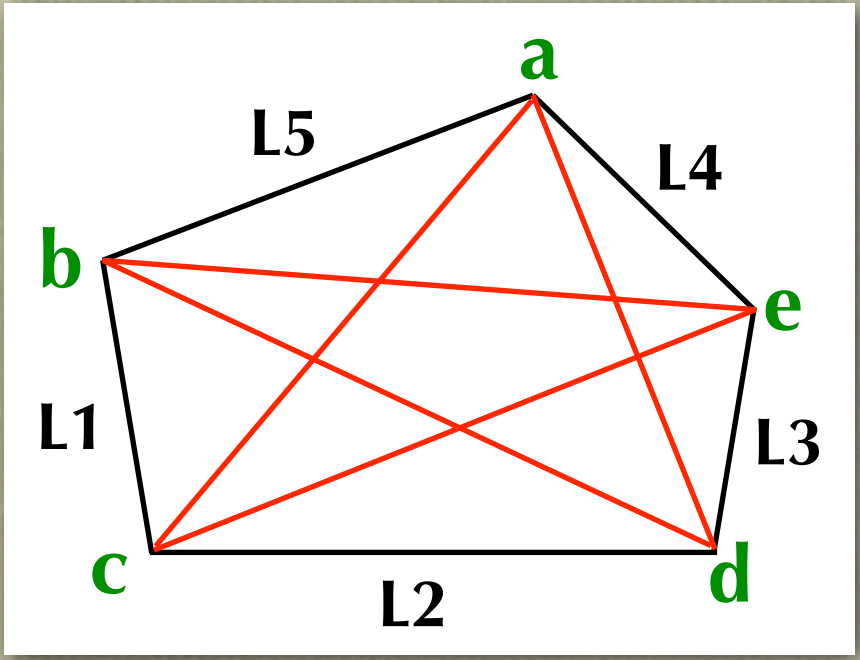
Definition of the shape of a thing

Triangle



(L1, L2, L3)

N-point general geometrical structure



$$
\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array}
\begin{array}{c}
\begin{array}{cccc} a & b & & e \end{array} \\
\left[\begin{array}{cccc}
d_{11} & d_{12} & \cdots & d_{1N} \\
d_{21} & d_{22} & \cdots & d_{2N} \\
d_{31} & & & \\
\vdots & & & \\
d_{N1} & d_{N2} & \cdots & d_{NN}
\end{array}\right]
\end{array}
$$

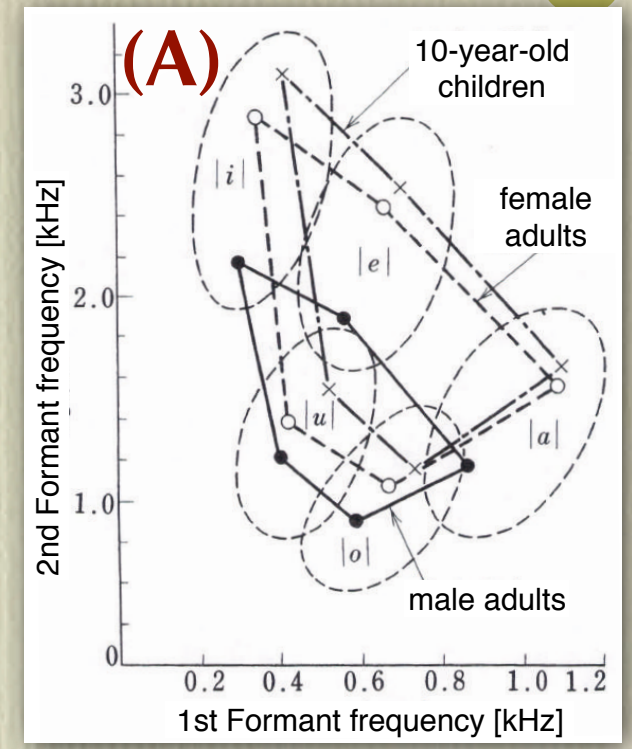
Insensitivity and sensitivity

Infants' vocal learning is

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

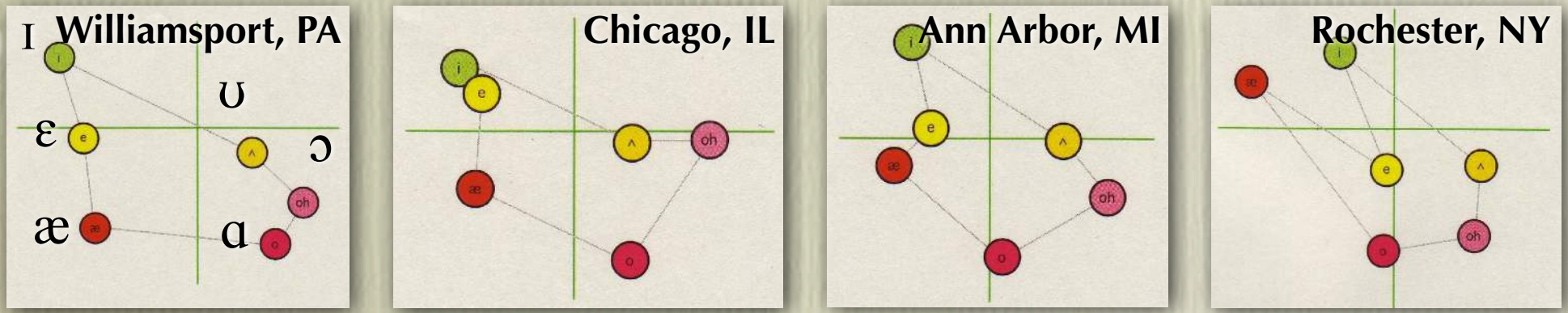
Infants' vocal learning seems to be

- 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]

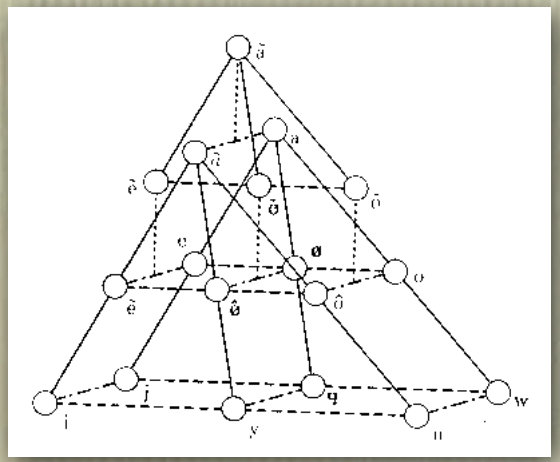
A claim found in classical linguistics

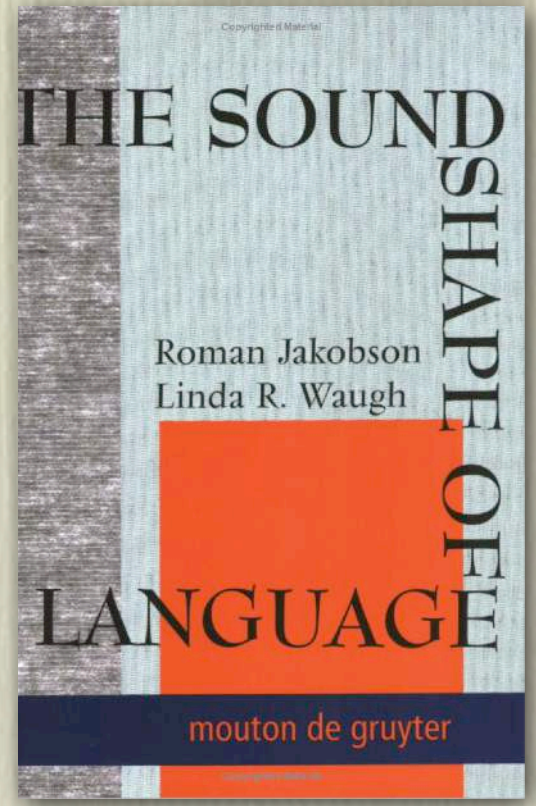
Theory of relational invariance [Jakobson+'79]

- 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.





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

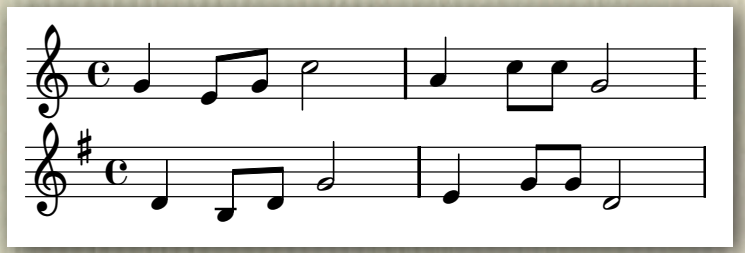
Radical but interesting discussion

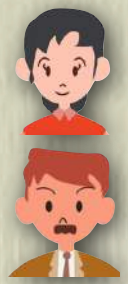
- An interesting link to some behaviors found in language disorder
- An interesting thought experiment

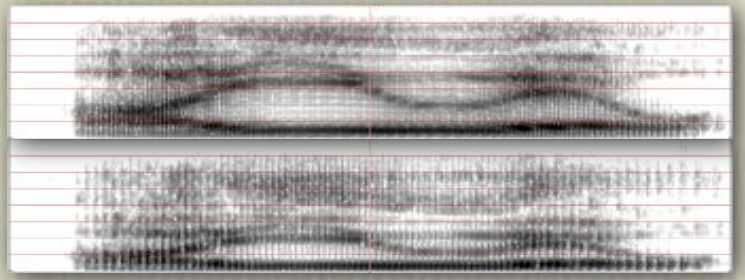
Physical variability and cognitive constancy

Receptors receive very physically-variable stimuli.

- Variability in appearance
  - A dog with different angles
  - A dog with different distances
- Variability in color
  - Flowers at sunrise and those at sunset
  - Flowers seen through colored glasses
- Variability in pitch
  - Humming of a male and that of a female
  - Key change (transposition) of a melody
- Variability in timbre
  - A male's "hello" and a female's
  - An adult's "hello" and a child's







But we can find the equivalence among them easily.

Physical variability and cognitive constancy

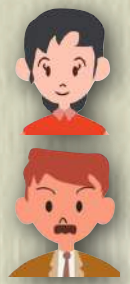
Receptors receive very physically-variable stimuli.

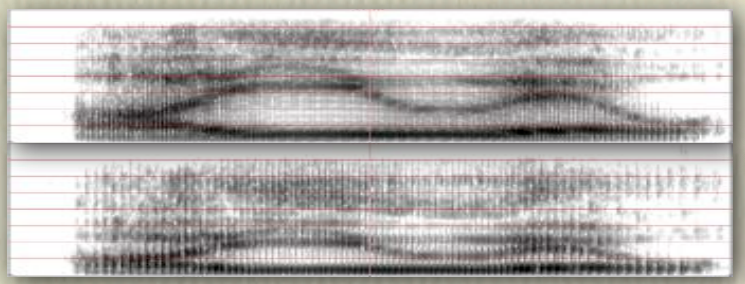
- Variability in appearance
  - A dog with different angles
  - A dog with different distances

- Variability in color

Stimuli deformation caused by static bias and invariant perception of these stimuli

  - Key change (transposition) of a melody



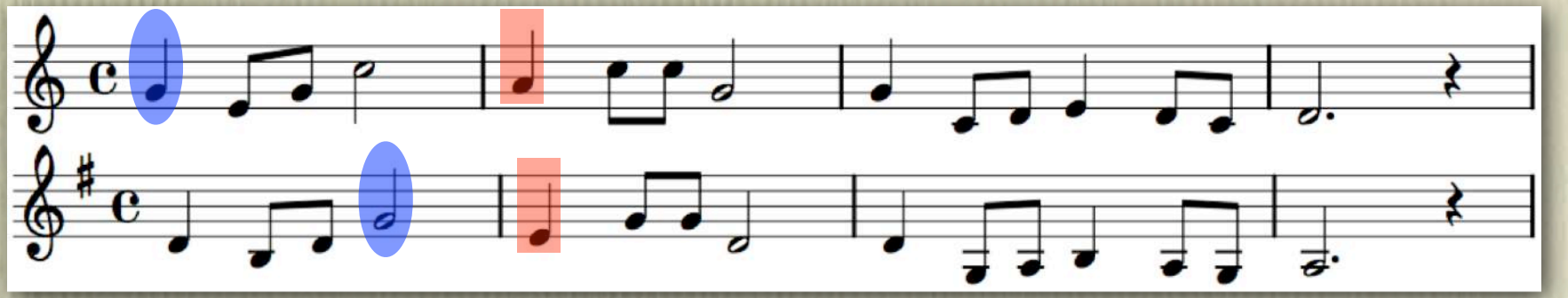


- Variability in timbre
  - A male's "hello" and a female's
  - An adult's "hello" and a child's

But we can find the equivalence among them easily.

Invariant pitch perception against its bias

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



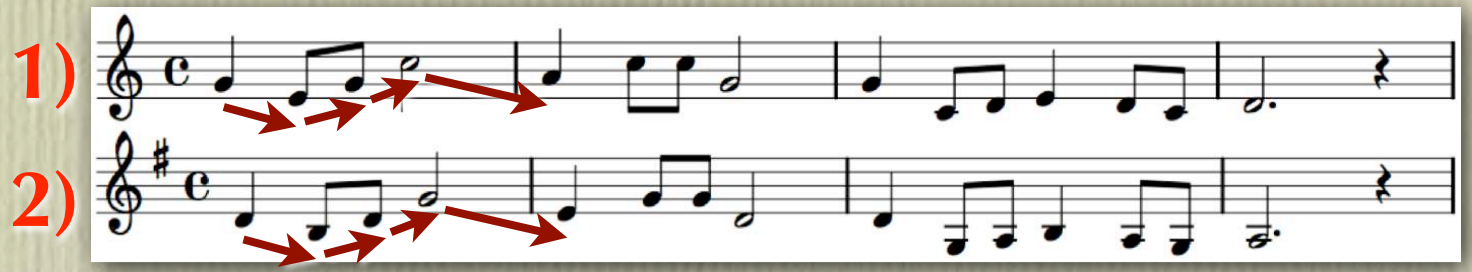
- Absolute (perfect) pitch (Do, Re, Mi... = pitch names) **(音名)**
  - 1 = So, Mi, So, Do, La, Do, Do, So. 2 = Re, Ti, Re, So, Mi, So, So, Re.
- Relative pitch who can transcribe (Do, Re... = syllable names)
  - 1 = So, Mi, So, Do, La, Do, Do, So. 2 = So, Mi, So, Do, La, Do, Do, So.

(階名)

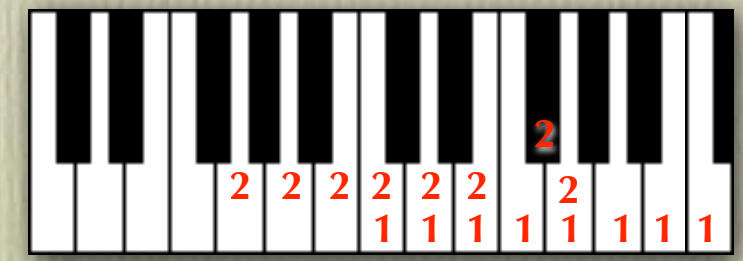
- Relative pitch who cannot transcribe
  - 1 = La, La, La, La, La, La, La, La. 2 = La, La, La, La, La, La, La, La
- Different / identical tones are claimed to be identical / different.
- Not fundamental frequency (absolute property) of each tone, but it only matters **what contrast each tone has to its surrounding tones.**

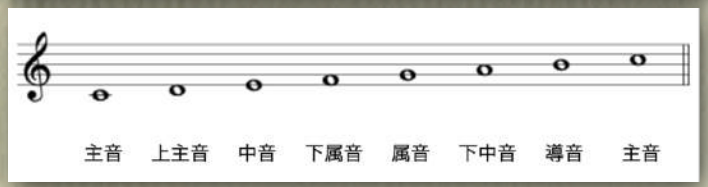
Invariant pitch perception against its bias

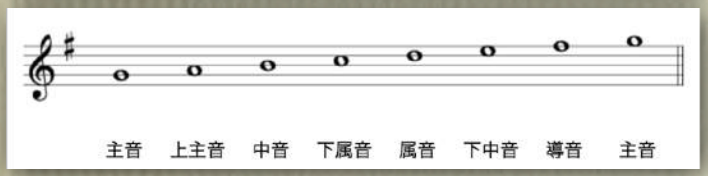
A melody and its transposed version [Higashikawa'05]

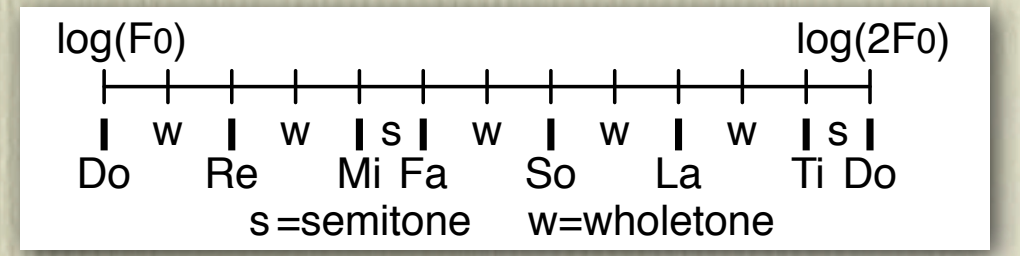


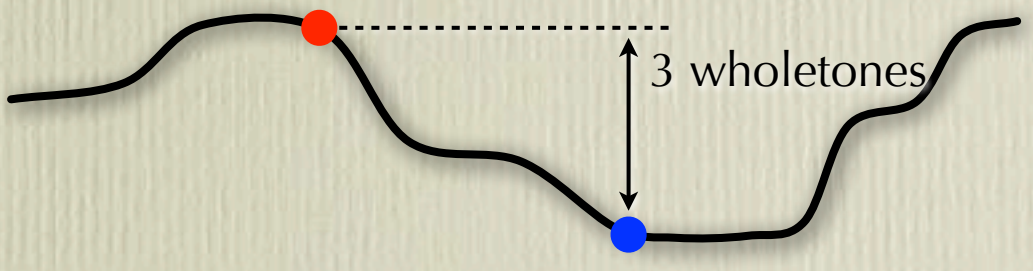
- Listeners with RP can perceive the same sound name sequence.
  - So Mi So Do / Ra Do Do So / So Do Re Mi Re Do / Re
  - The same sound distribution pattern is found in 1) and 2).







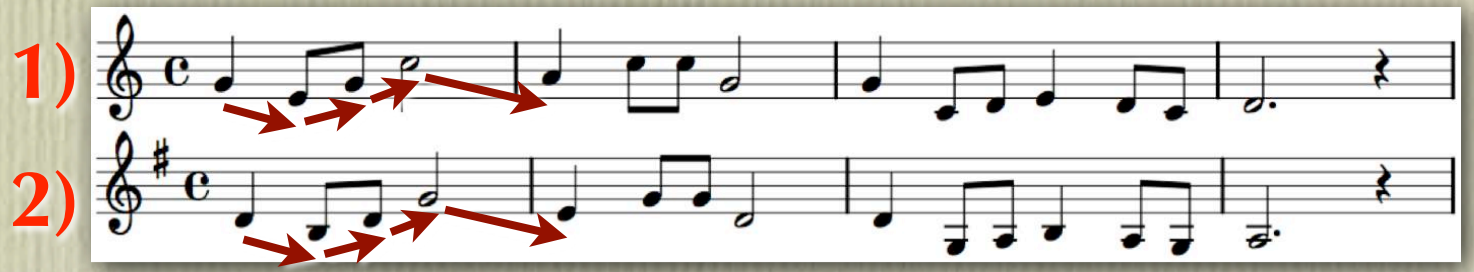




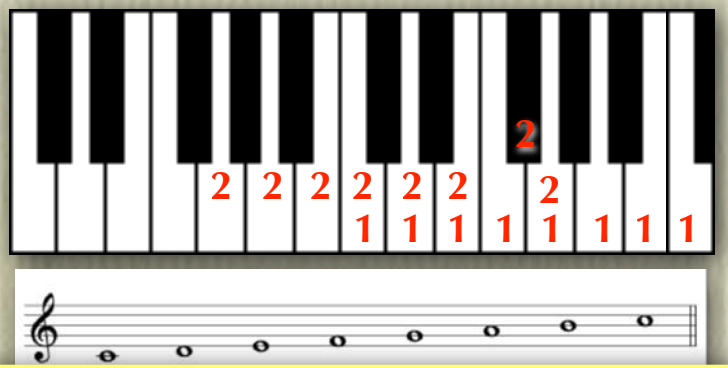
https://ja.wikipedia.org/wiki/音度

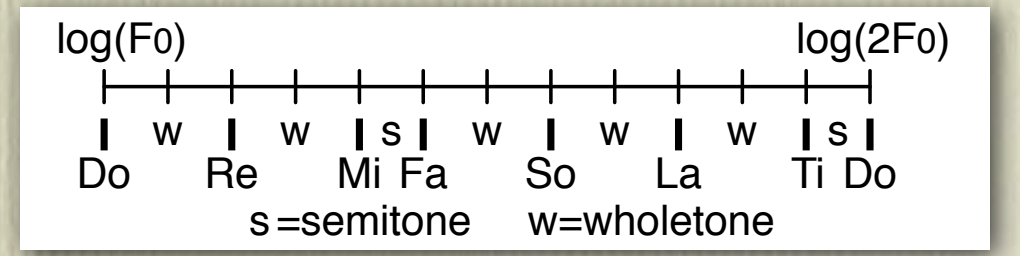
Invariant pitch perception against its bias

A melody and its transposed version [Higashikawa'05]



- 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).





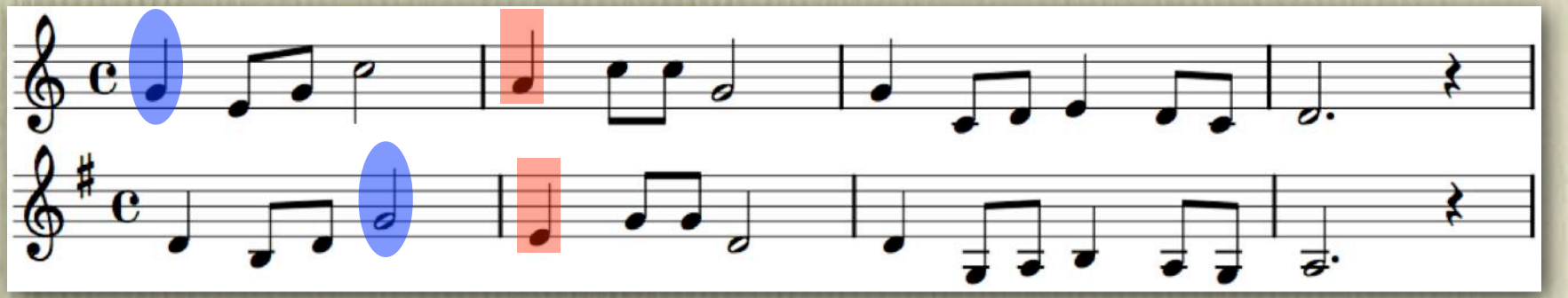
But it is very difficult to label a single tone because there is no contrast at all.

Invariant pitch perception against its bias

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

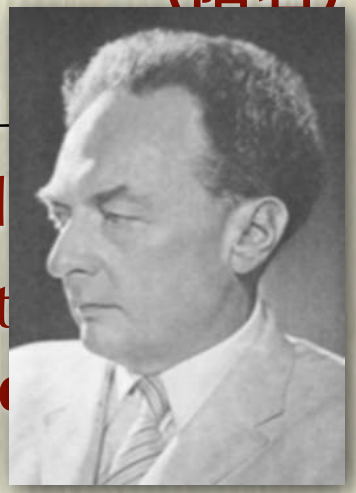
1

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 who can transcribe (Do, Re... = syllable names)
  - 1 = So, Mi, So, Do, La, Do, Do, So. 2 = So, Mi, So, Do, La, Do, Do, So.
- Relative pitch who cannot transcribe
  - 1 = La, La, La, La, La, La, La, La. 2 = La, La, La, La, La, L
- Different / identical tones are claimed to be identical / d
- Not
  only

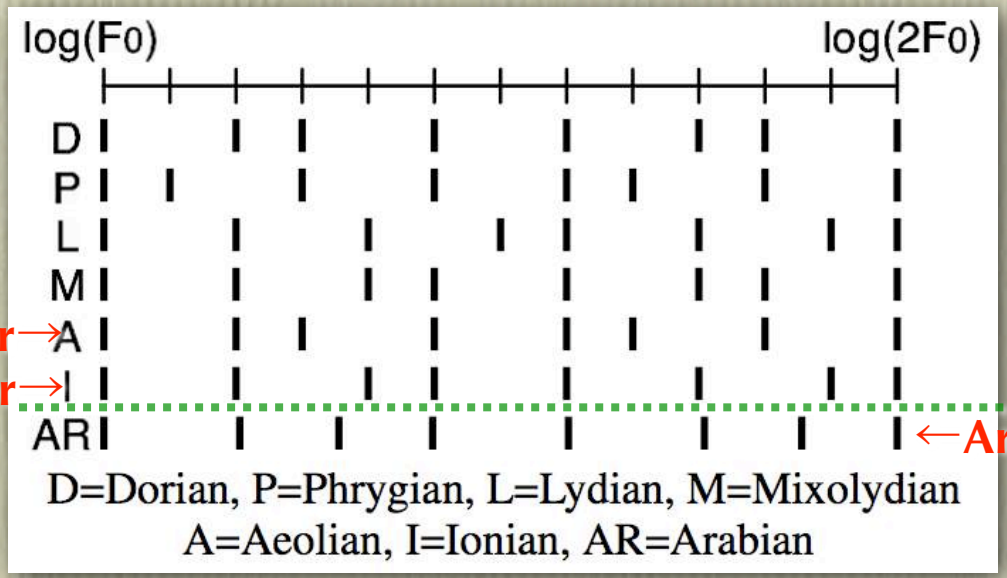
(階名)



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

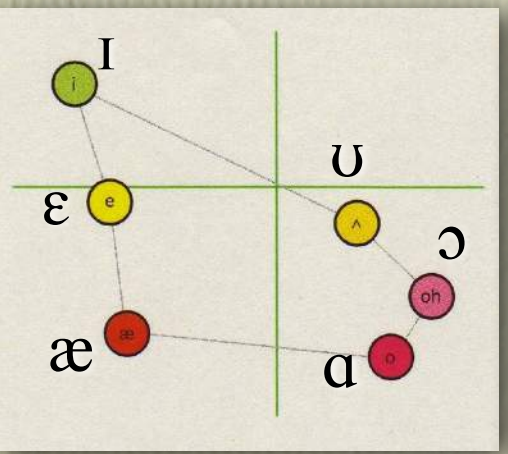
Relative pitch vs. relative timbre

Key-invariant arrangement of tones and its variants

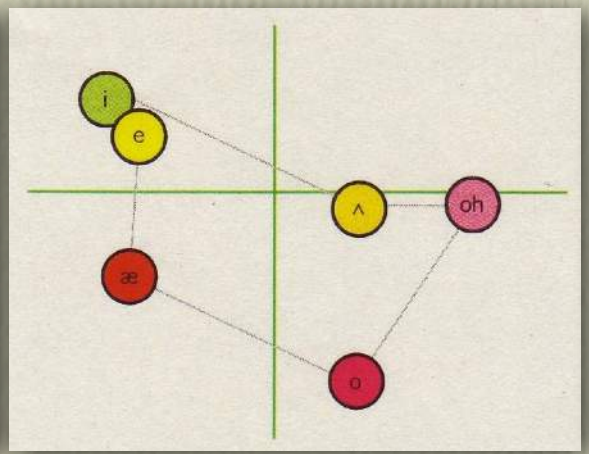


- Western = 5 whole + 2 semi
- D to I = classical church music
- Arabic = with non-semi intervals
  - Western music in Arabic scale

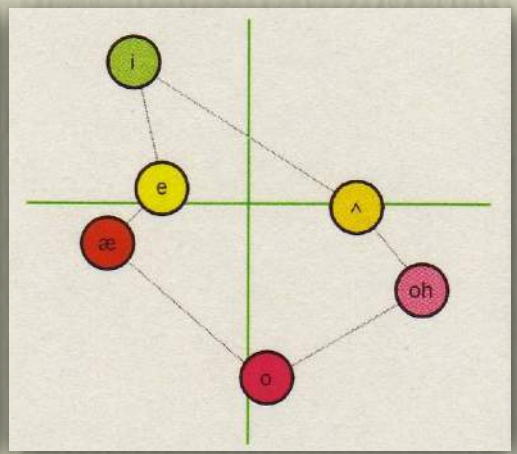
Spk-invariant arrangement of vowels and its variants



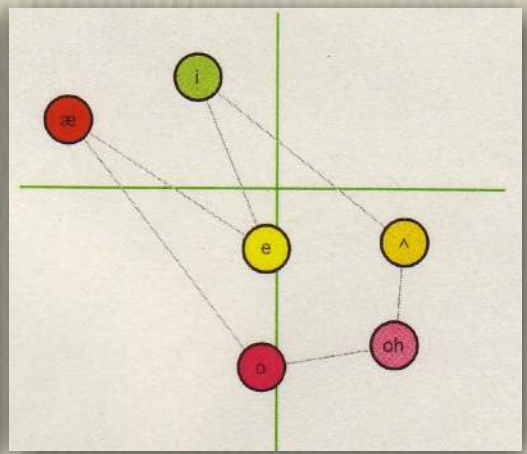
Williamsport, PA



Chicago, IL



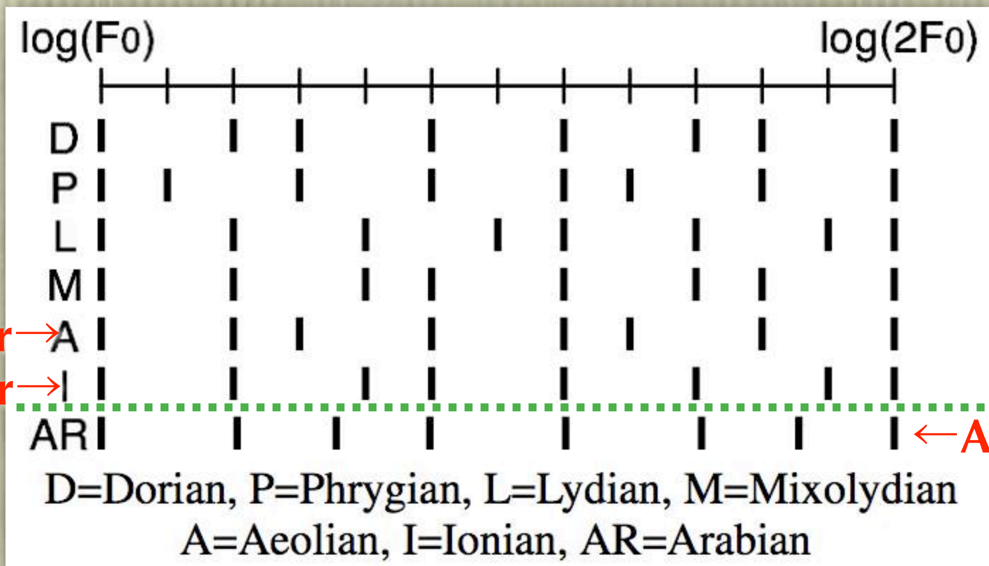
Ann Arbor, MI



Rochester, NY

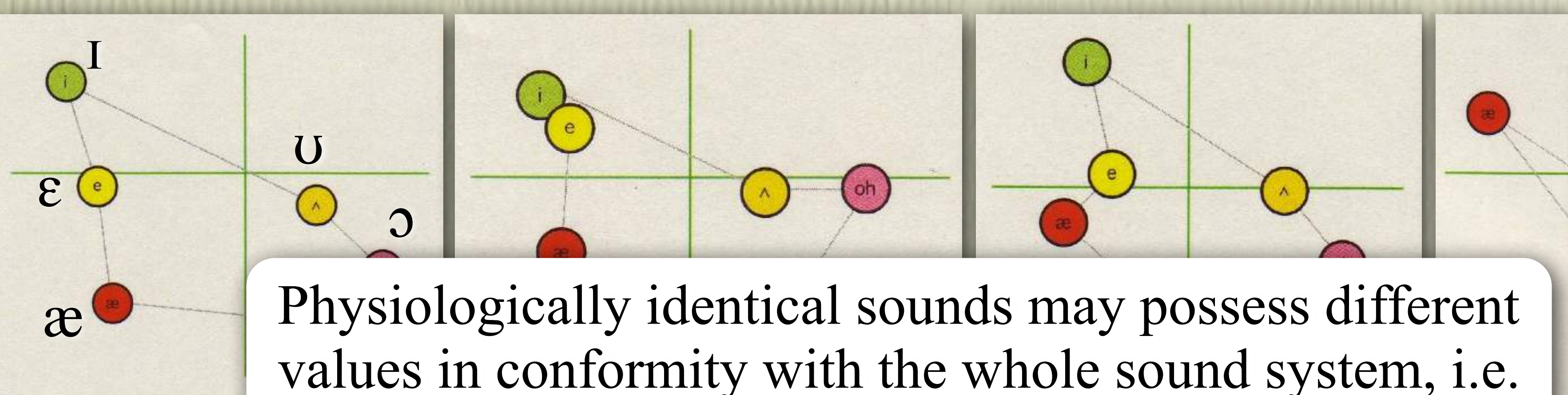
Relative pitch vs. relative timbre

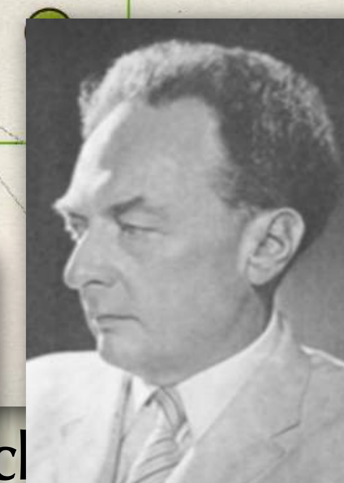
Key-invariant arrangement of tones and its variants



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

Spk-invariant arrangement of vowels and its variants



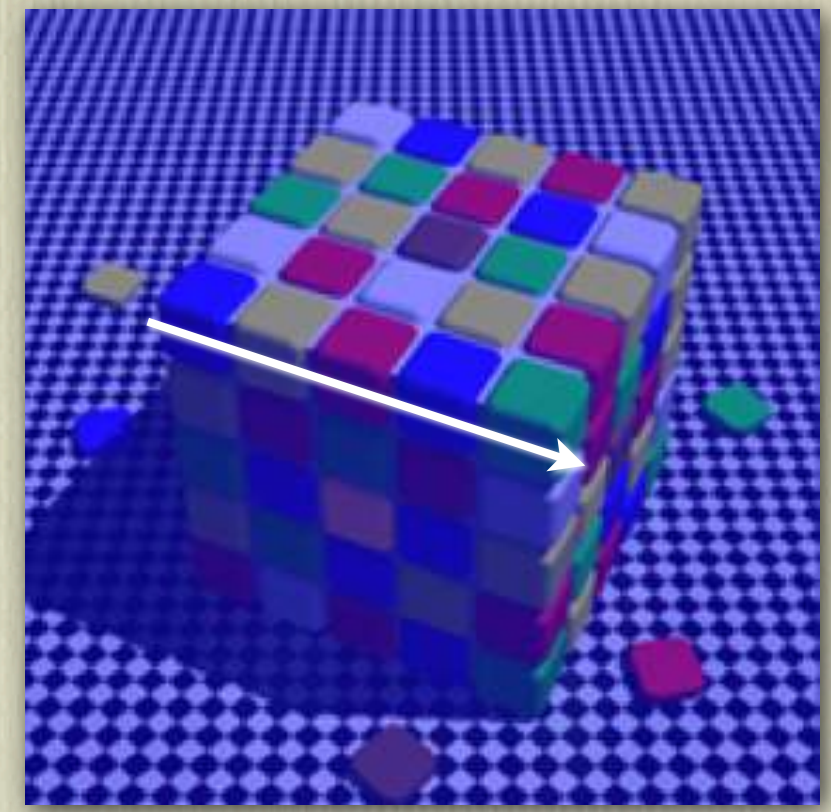


Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. in their relations to the other sounds.

Williamsp

Invariant color perception against its bias

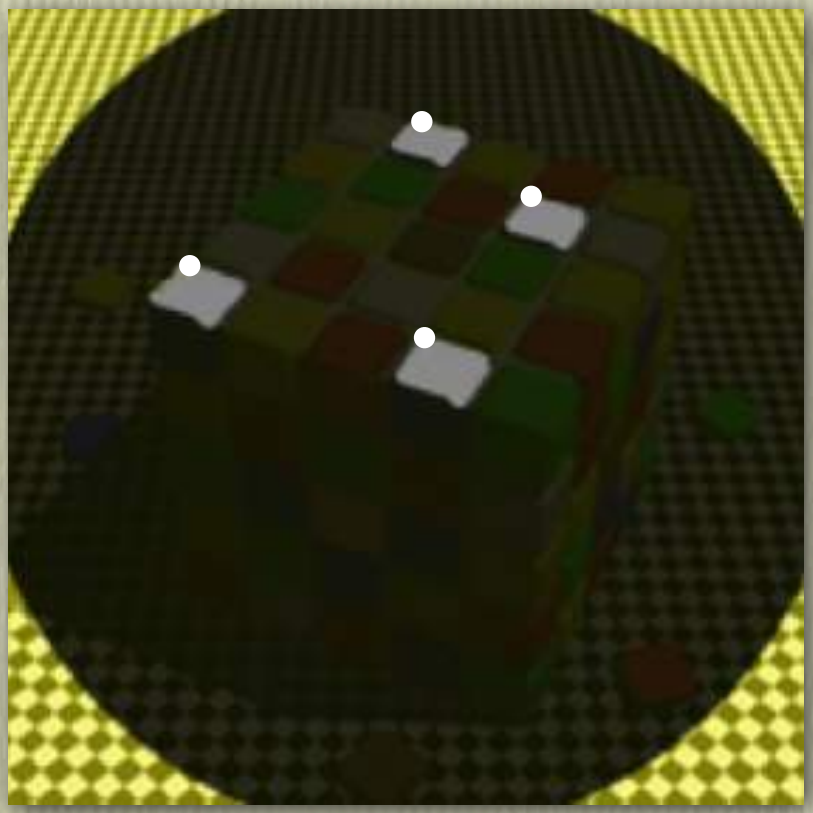
The Rubik's cube seen through colored glasses [Lotto'99]

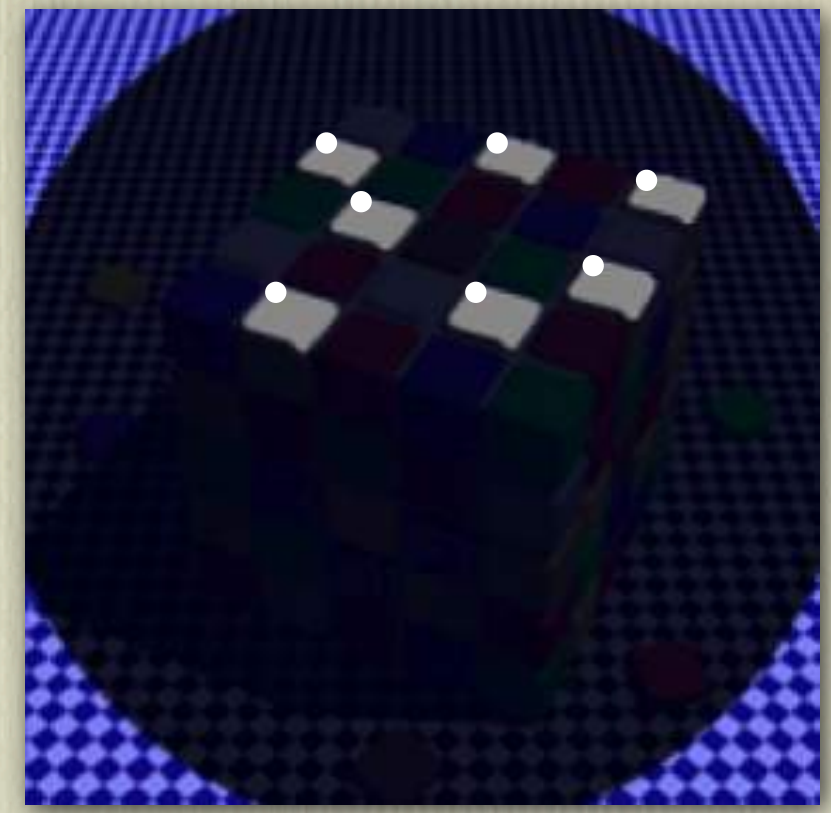


- We perceive that the two cubes are identical.
- Different / identical colors are claimed to be identical / different.
- Not only wavelength (absolute property) of each patch, but also it matters **what contrast each patch has to its surrounding patches.**

Invariant color perception against its bias

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

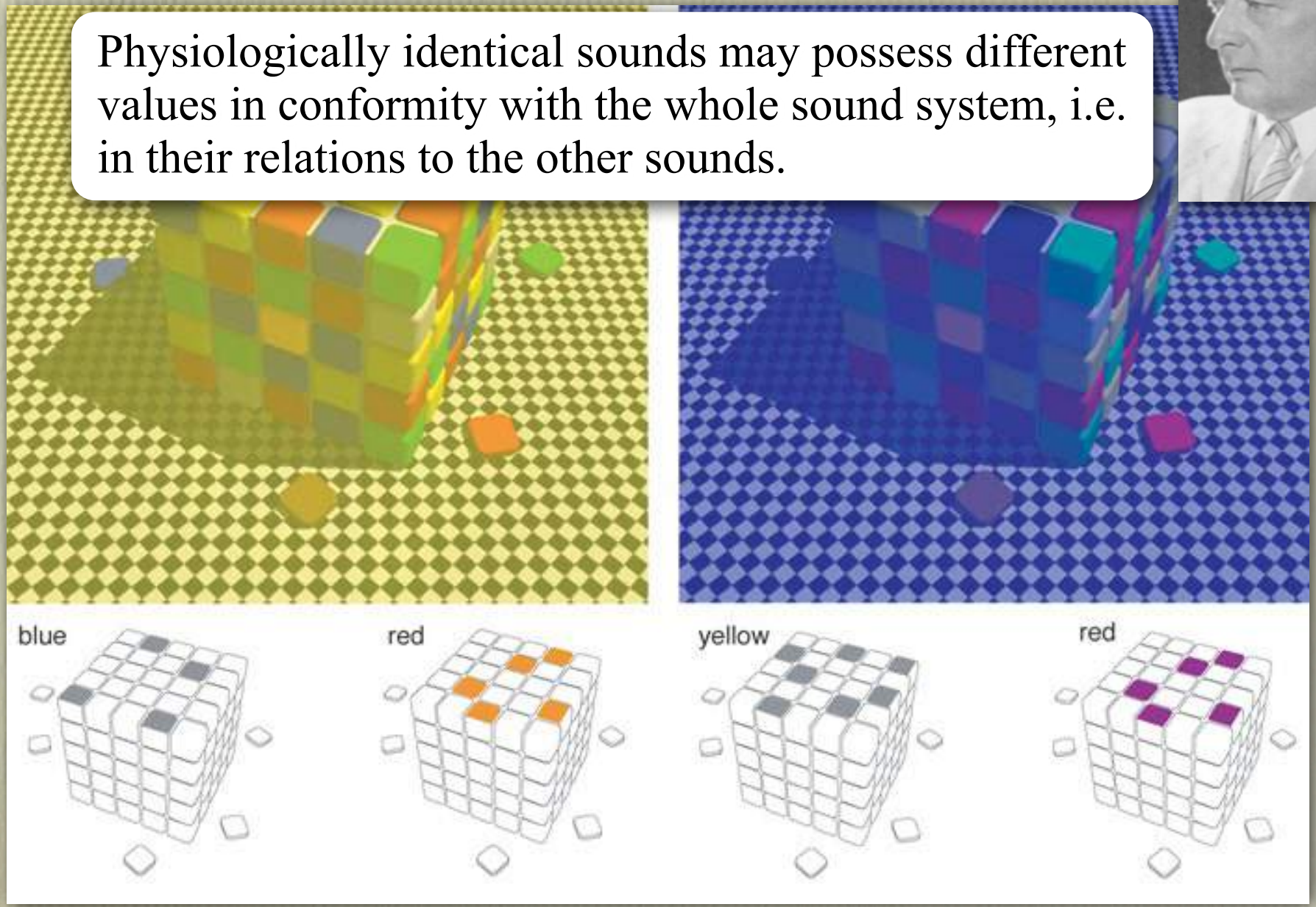




- We perceive that the two cubes are identical.
- Different / identical colors are claimed to be identical / different.
- Not only wavelength (absolute property) of each patch, but also it matters **what contrast each patch has to its surrounding patches.**

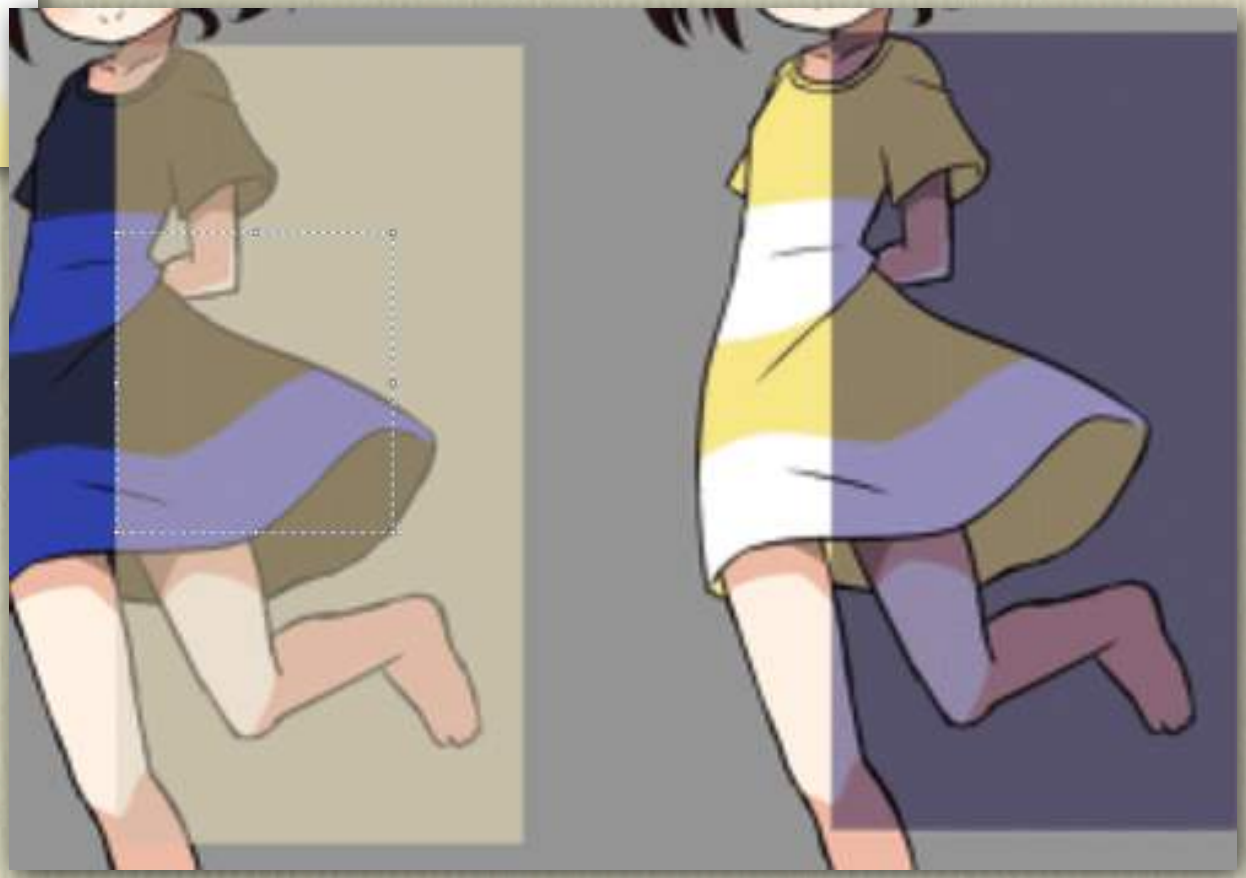
Invariant color perception against it

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



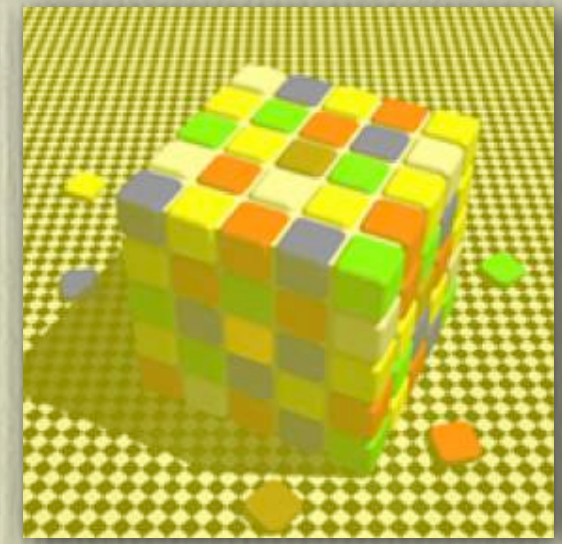
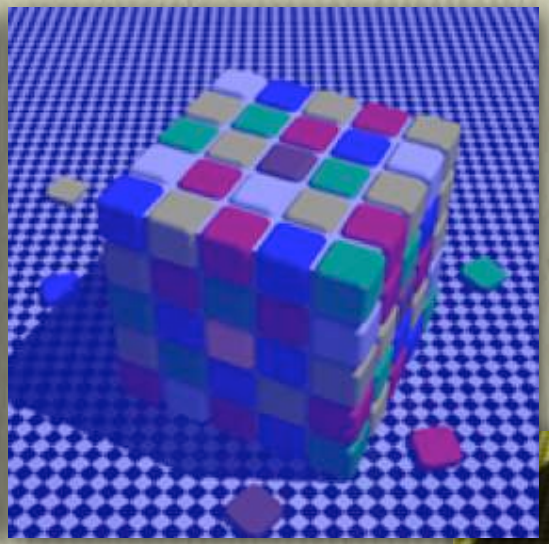
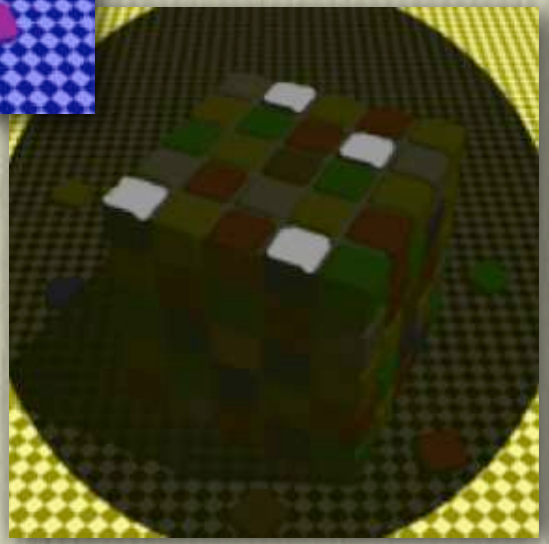
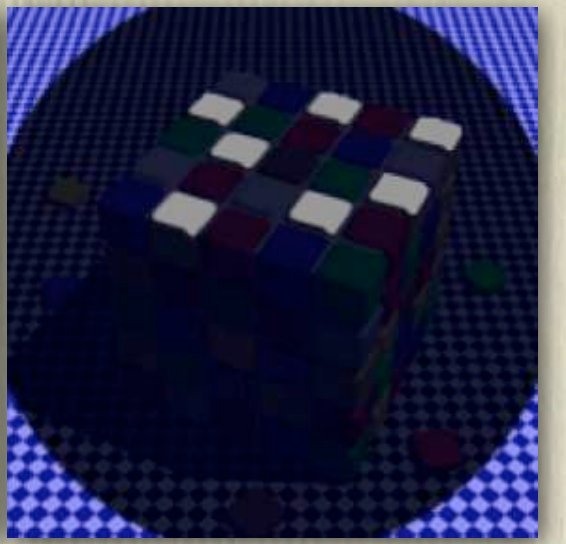
Reprinted from Dale Purves, R. Beau Lotto, Surajit Nundy, "Why We See What We Do,", American Scientist, vol. 90, no. 3, page 236. www.americanscientist.org/template/AssetDetail/assetid/14755.

Do you still remember this?



An evolutional point of view

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





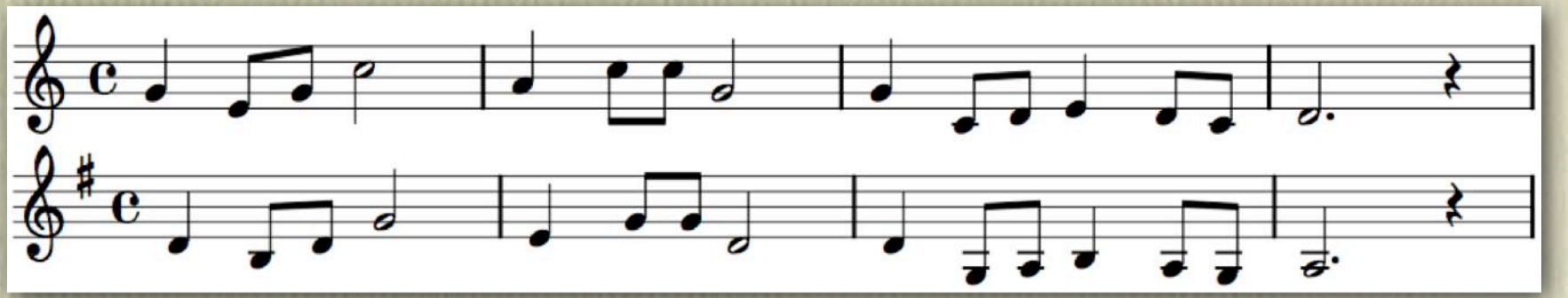



An evolutional point of view

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

1

2



1 = 2

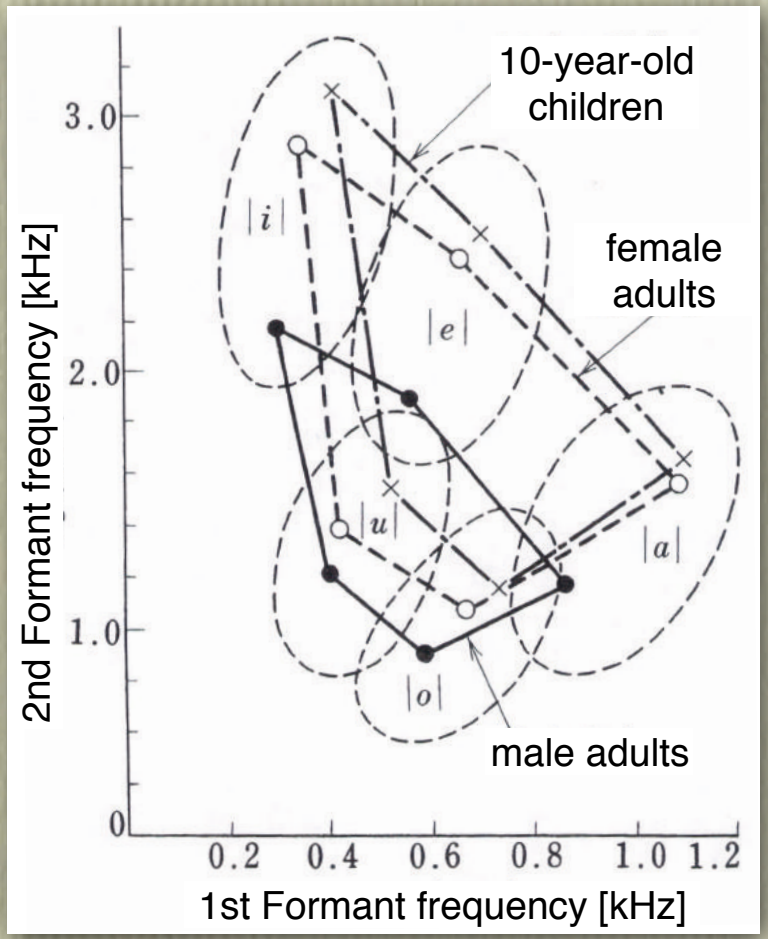







An evolutional point of view

How old is the relative perception in evolution?



1=2=3








?


Insensitivity 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 *mis*matched imitation.

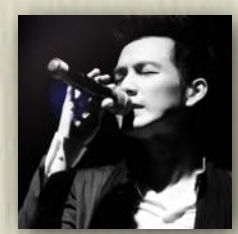
- We're very insensitive to speaker identity transmitted via speech.

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

Invariant timbre perception against its bias

Factors causing static pitch bias in speech

- Length and mass of the vocal chords

Factors causing static timbre bias in speech

- Size and shape of the vocal tract

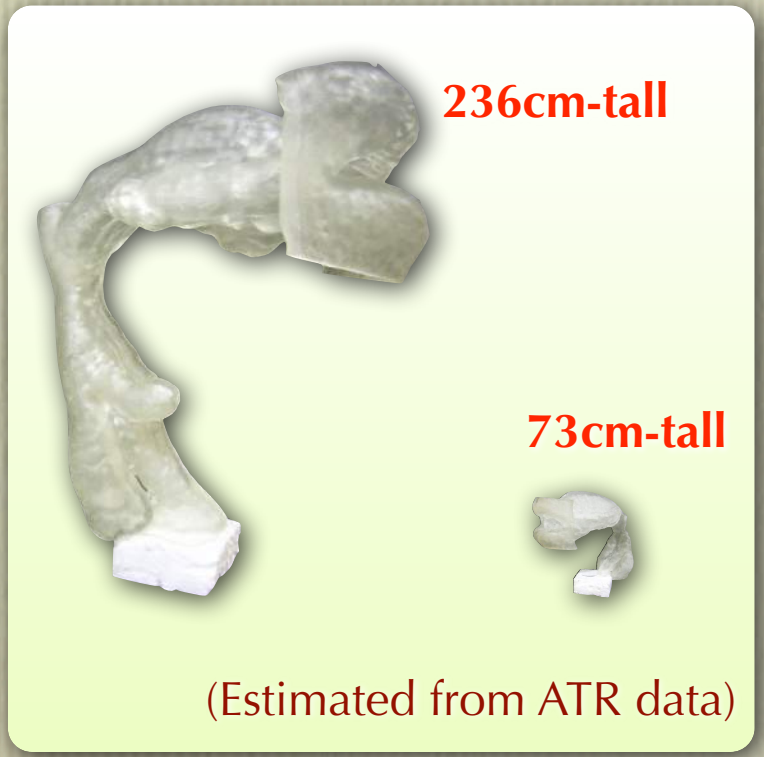
Invariant timbre perception against its bias

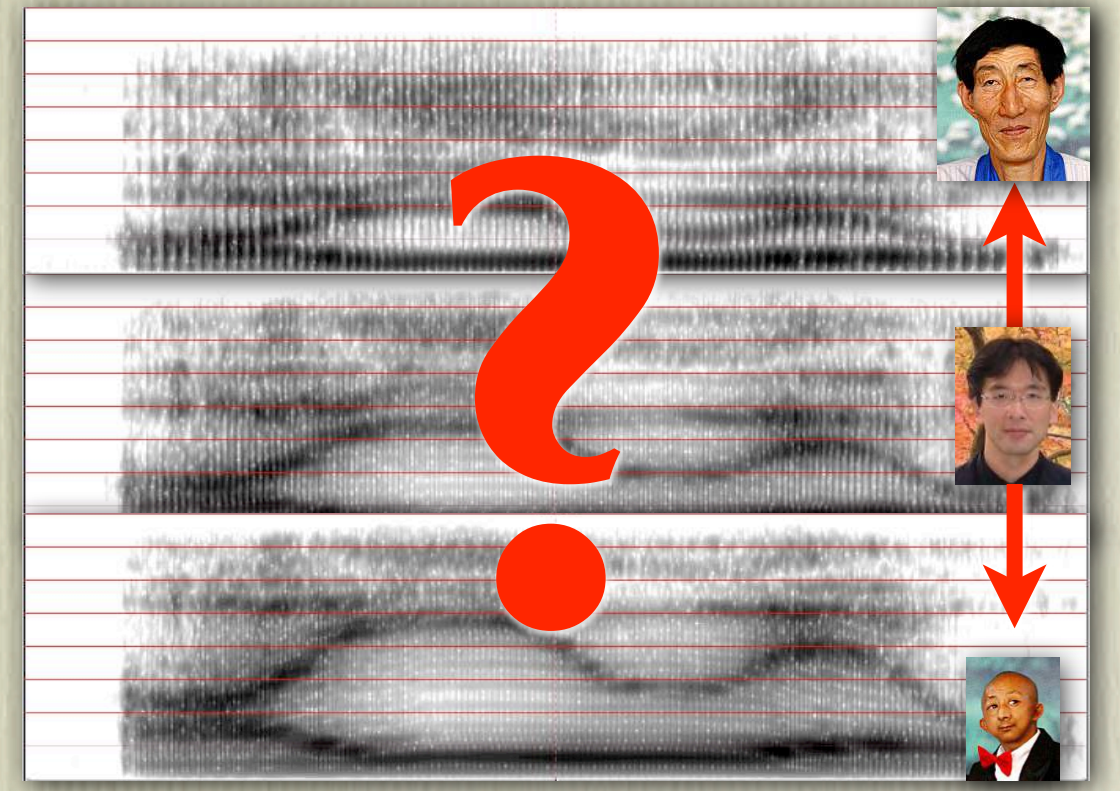
Factors causing static pitch bias in speech

- Length and mass of the vocal chords

Factors causing static timbre bias in speech

- Size and shape of the vocal tract





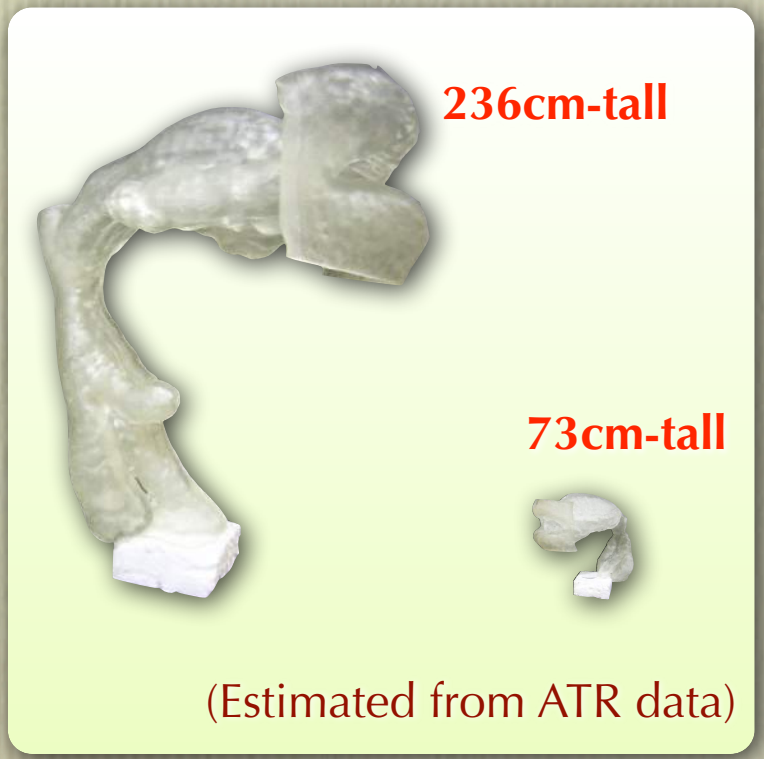
Invariant timbre perception against its bias

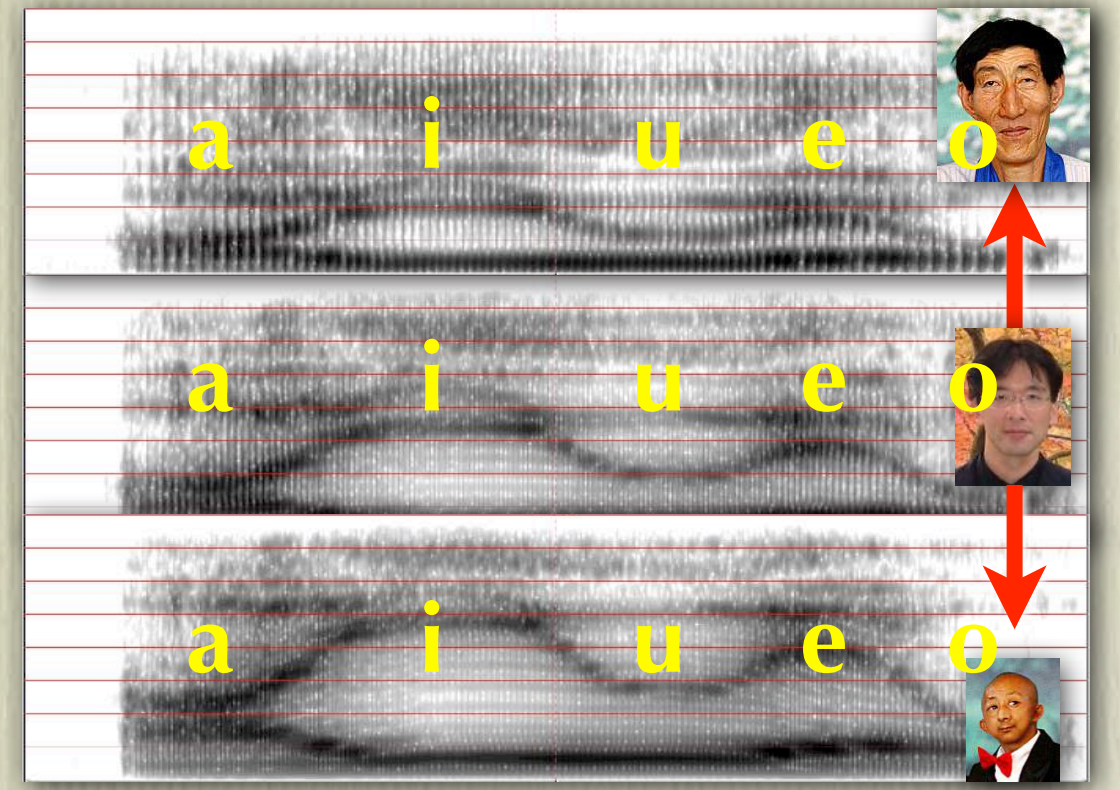
Factors causing static pitch bias in speech

- Length and mass of the vocal chords

Factors causing static timbre bias in speech

- Size and shape of the vocal tract





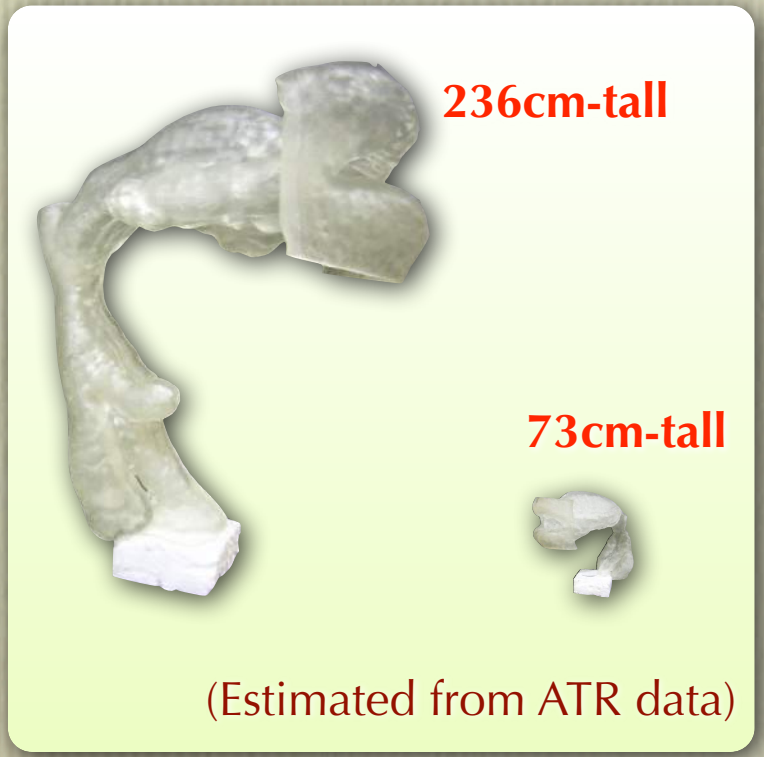
Invariant timbre perception against its bias

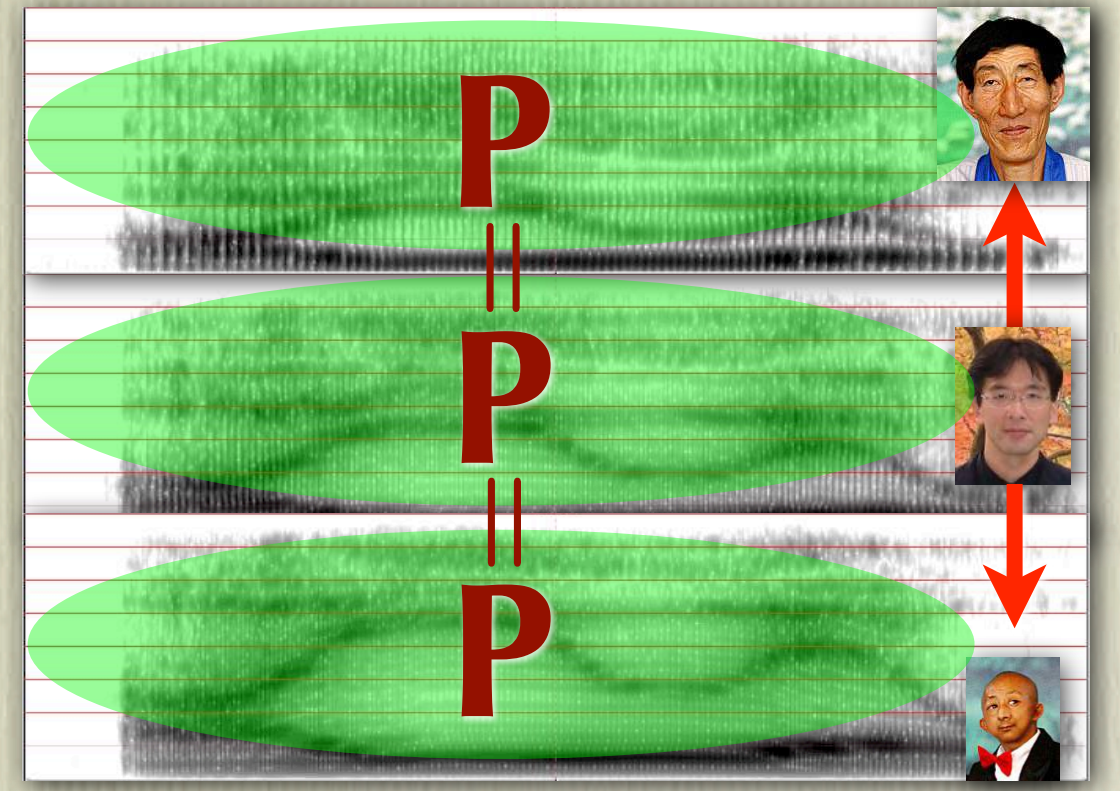
Factors causing static pitch bias in speech

- Length and mass of the vocal chords

Factors causing static timbre bias in speech

- Size and shape of the vocal tract





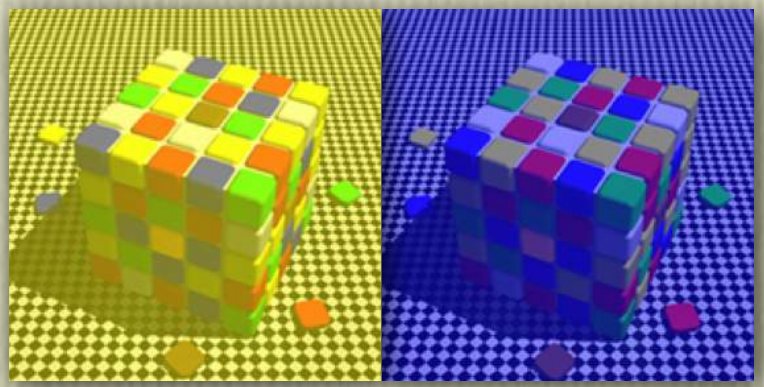
The world-tiniest high school girl!!

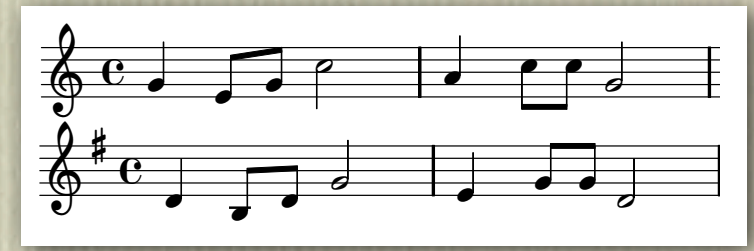
Linearly size-reduced individual!?

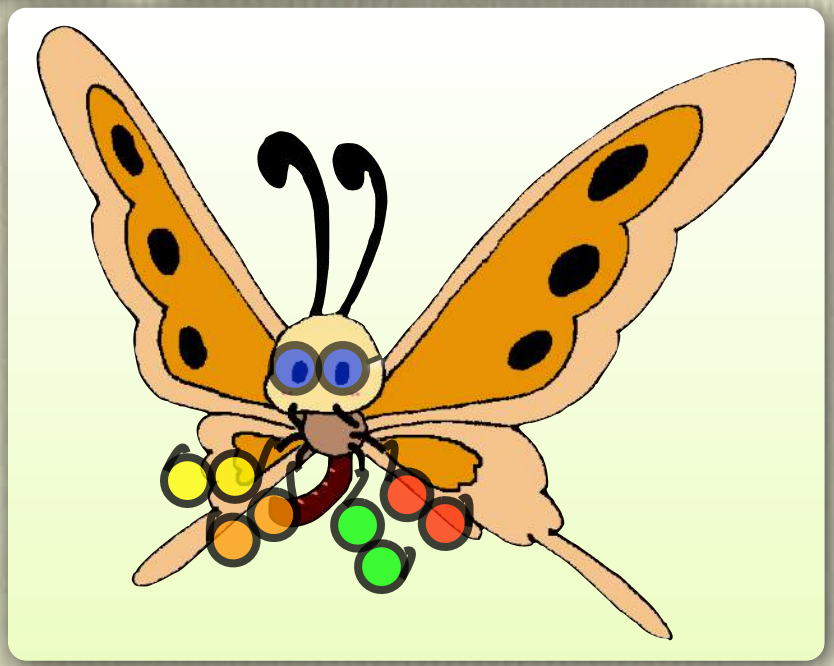
Invariant timbre perception against its bias

Invariant and constant perception wrt. color and pitch

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





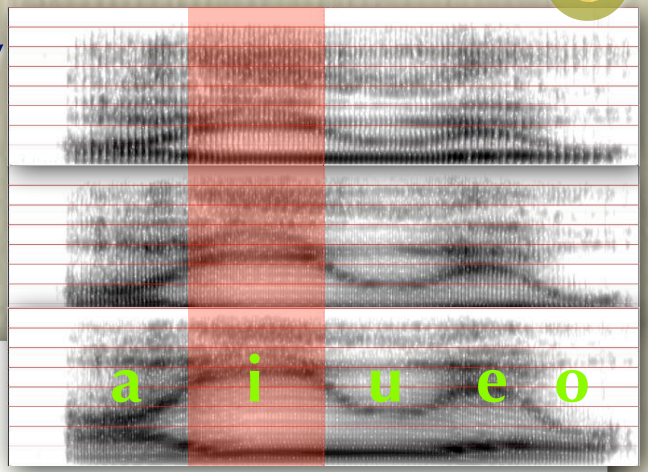


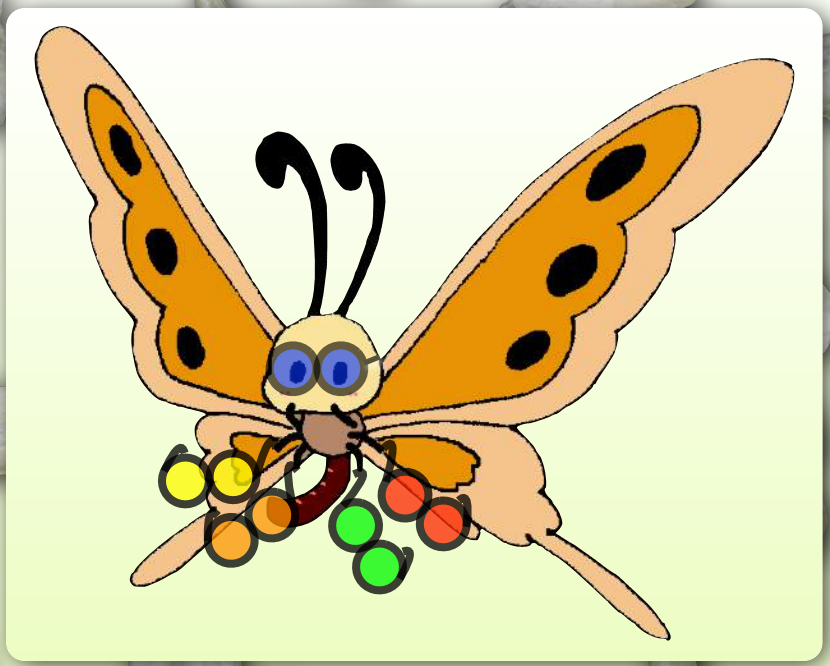
$$P(o|w) \sim \sum_s P(o|w,s)P(s)$$

Invariant timbre perception against its bias

De facto standard for timbre variability

- Segmentation of speech into elements
- Statistical models for individual elements



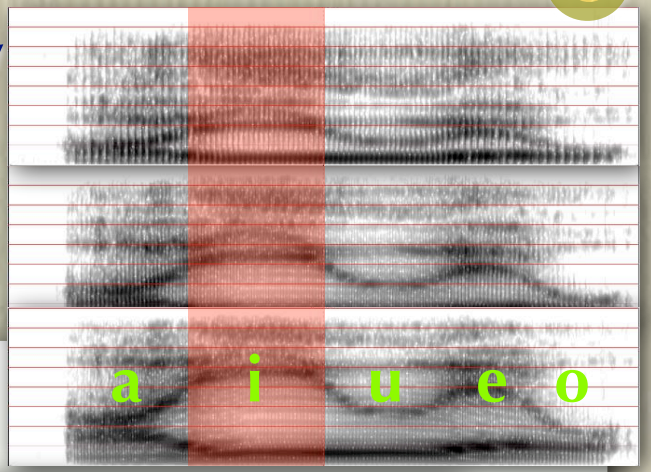


hundreds to thousands

Invariant timbre perception against its bias

De facto standard for timbre variability

- Segmentation of speech into elements
- Statistical models for individual elements



to thousands

A difference bet. machines and humans

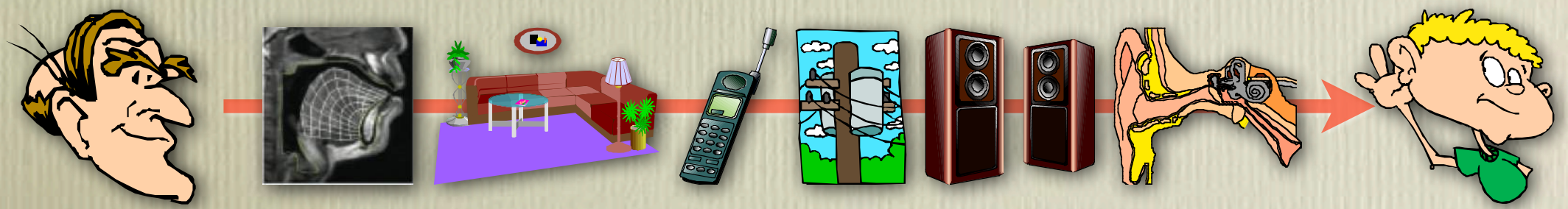
Machine strategy (engineers' strategy): ASR

- 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.

?

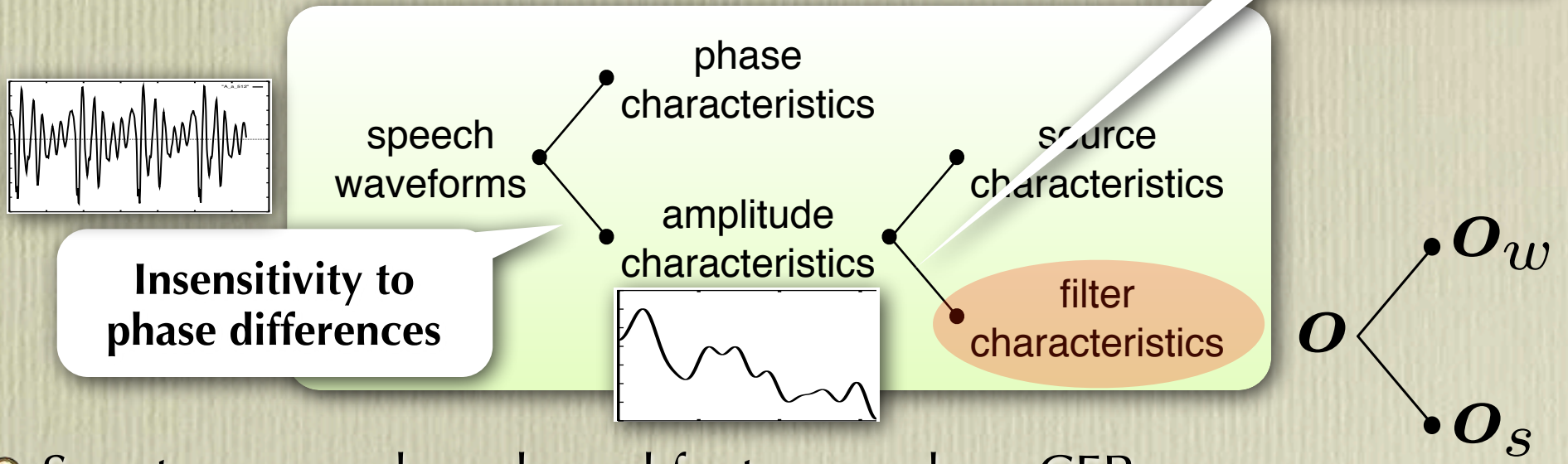
Human strategy: HSR

- 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



- Spectrum envelope-based feature such as CEP: o
  - But o depends on all the three kinds of info. (ling, para-ling, extra-ling).
- How to suppress extra-linguistic variation in o ?
  - Feature normalization: transforming o to that of the standard speaker
  - Model adaptation: modifying model parameters to fit to the input speaker
  - Statistical independence: hiding these variation through sample collection
  - Physical independence: pursuing features invariant to these variation
  - :

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 ≠ acoustic = ??

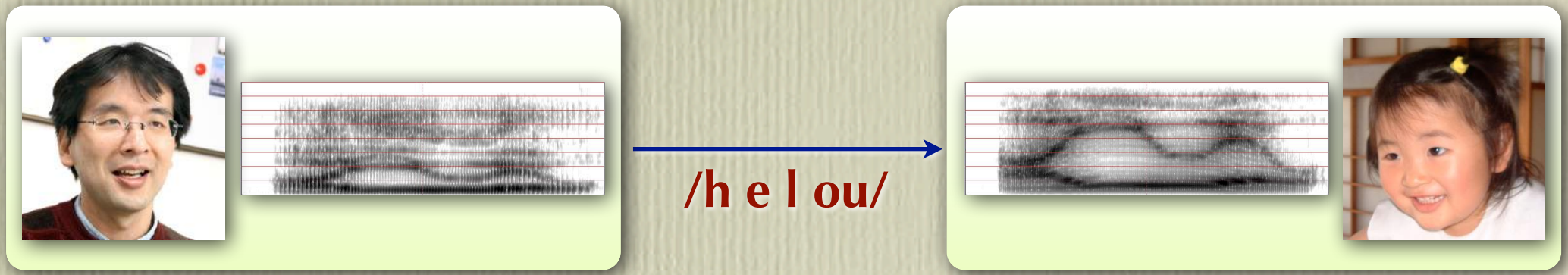
- Acoustic imitation performed by myna birds [Miyamoto'95]
  - They imitate the sounds of cars, doors, dogs, cats as well as human voices.
  - Hearing a very good myna bird say something, one can guess its owner.
- Beyond-scale imitation of utterances performed by children
  - No one can guess a parent 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.



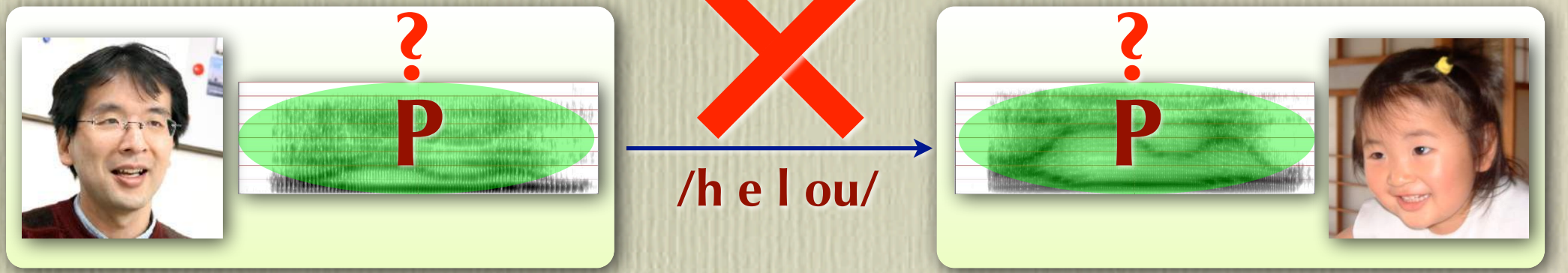
- Phonemic awareness is too poor to decompose an utterance.

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?

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?

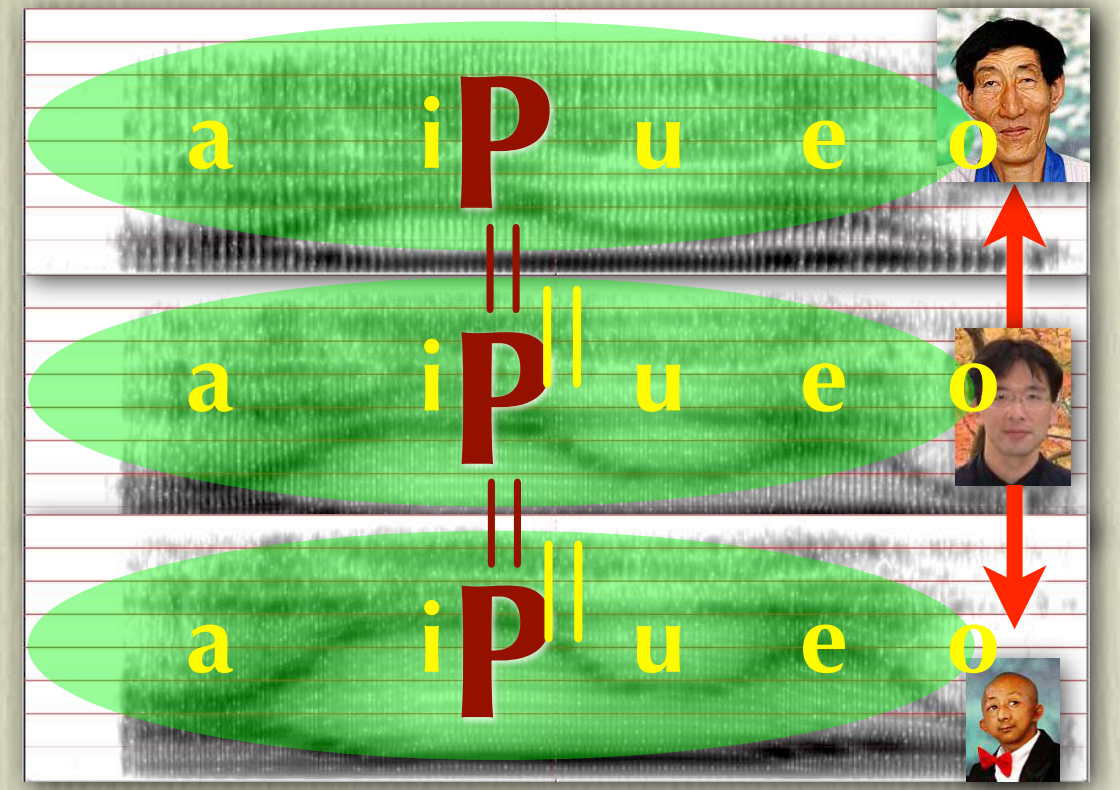
Invariant timbre perception against its bias

Factors causing static pitch bias in speech

- Length and mass of the vocal chords

Factors causing static timbre bias in speech

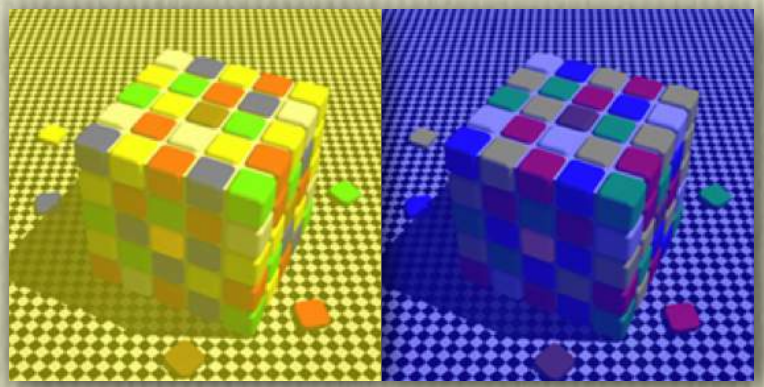
- Size and shape of the vocal tract

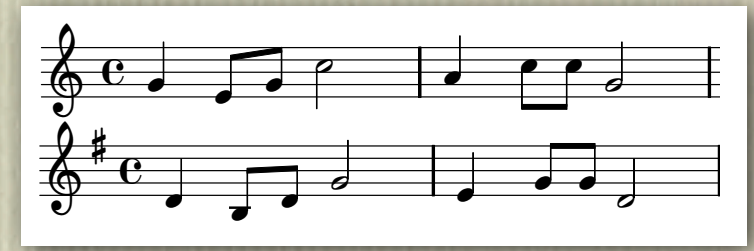


Invariant timbre perception against its bias

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