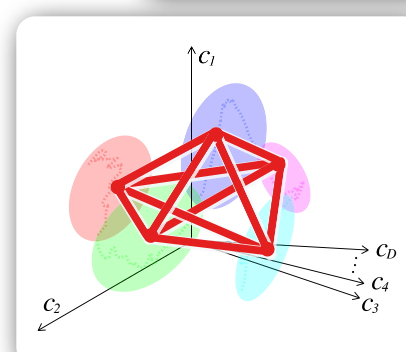
Cognitive Media Processing #9

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 new framework is introduced and explained.

A new framework for "human-like" speech machines #1

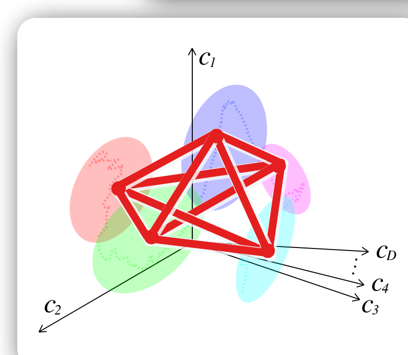
Nobuaki Minematsu

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 -

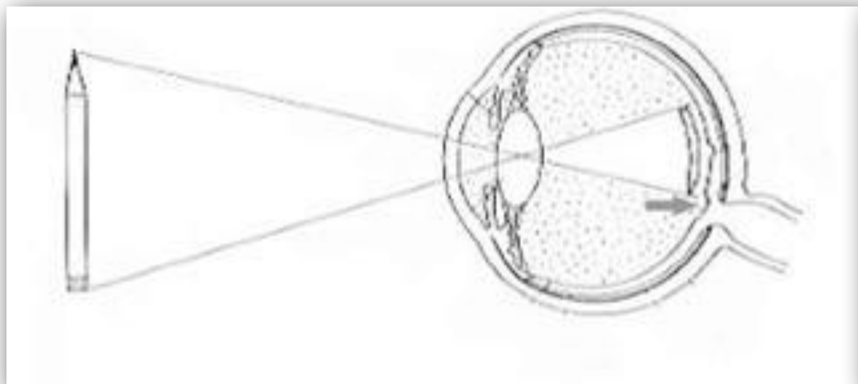
- Theme-2
  - Speech communication technology - articulatory & acoustic phonetics -
  - Speech communication technology - speech analysis -
  - Speech communication technology - speech recognition -
  - Speech communication technology - speech synthesis -

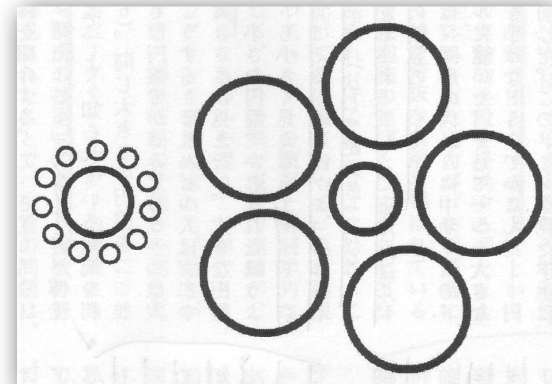
- Theme-3
  - A new framework for “human-like” speech machines #1
  - A new framework for “human-like” speech machines #2
  - A new framework for “human-like” speech machines #3
  - A new framework for “human-like” speech machines #4

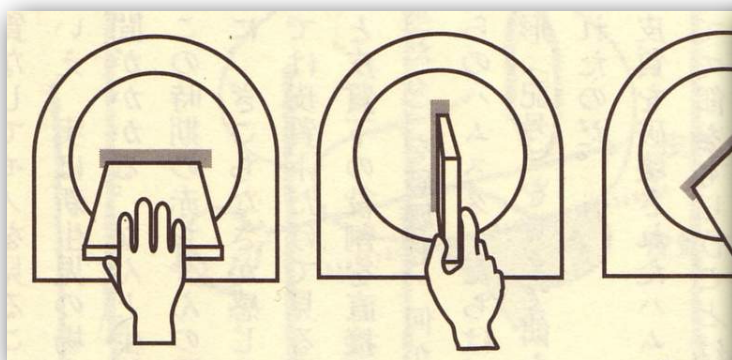


Human media information processing

- Unconscious processing
  - Blind spot, blind sight, color illusion, size illusion, etc

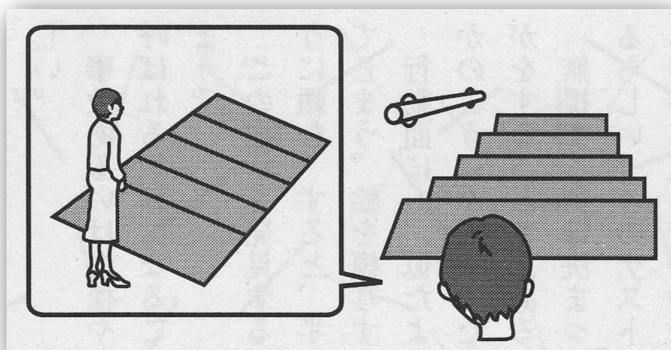


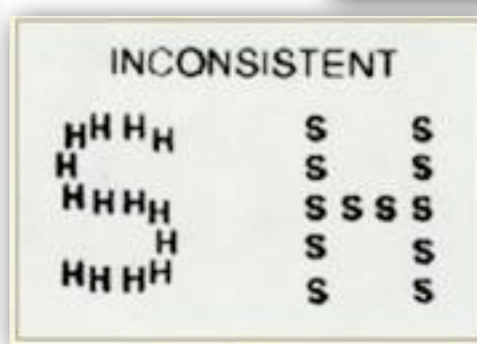




Human media information processing

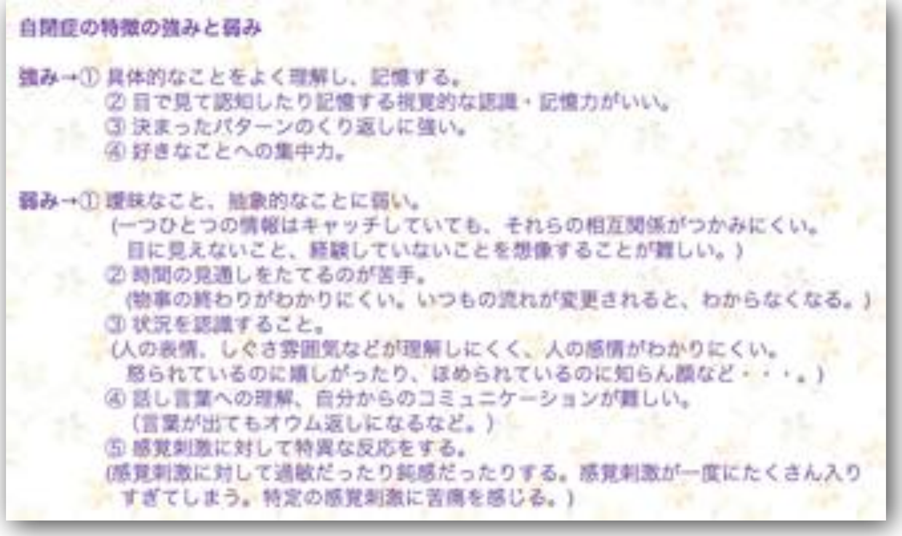
- Unconscious processing
  - Visual sensation described by a medical doctor with brain damage
  - Paying attention only to some specific objects
  - Some interesting behaviors of autistics (detailed memorization and rote learning?)





Sensation by autistics

- What are autistics good at and poor at?
  - Good at
    - remembering very detailed aspects of stimuli.
      - Especially their visual memory is often extraordinary.
    - processing constantly repeated patterns.
    - concentrating a (given) specific task.
  - Poor at
    - dealing with something abstract or invisible.
    - capturing the relations of things although good at capturing a specific one thing.
      - Good at capturing an element but poor at capturing them as a whole.
    - dealing with temporal development including future planning
    - understanding the environments properly.
      - Hidden messages are difficult to detect, ex. facial expressions, metaphors, etc.
    - understanding spoken language.
      - In cases of severely damaged autistics, their first language is written language.
    - smooth communication with others.
    - dealing properly with sensory stimuli.
      - Their sensitivity of sensory stimuli is too good. Can hear the sounds that non-autistics cannot hear.
      - Difficult to select important stimuli / difficult to ignore irrelevant stimuli.

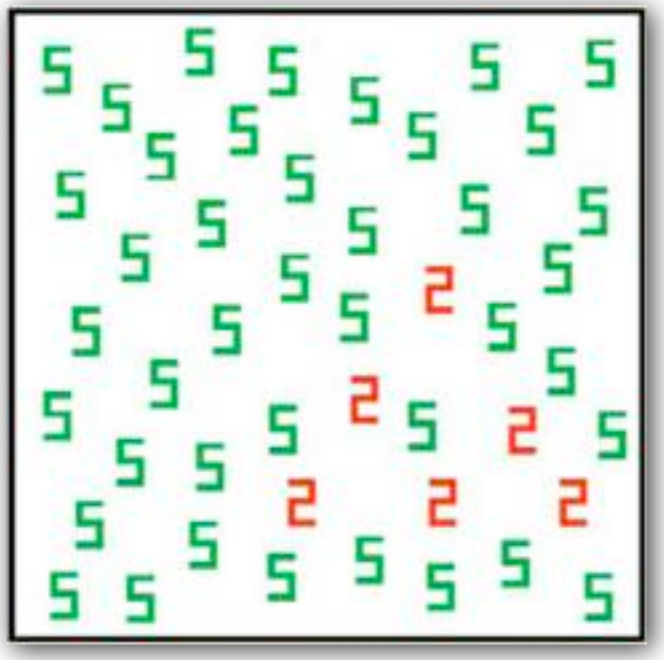

自閉症の特徴の強みと弱み

強み→① 具体的なことをよく理解し、記憶する。
② 目で見て認知したり記憶する視覚的な認識・記憶力がいい。
③ 決まったパターンのくり返しに強い。
④ 好きなことへの集中力。

弱み→① 曖昧なこと、抽象的なことに弱い。
(一つひとつの情報はキャッチしていても、それらの相互関係がつかみにくい。
目に見えないこと、経験していないことを想像することが難しい。)
② 時間の見通しをたてるのが苦手。
(物事の終わりがわかりにくい。いつもの流れが変更されると、わからなくなる。)
③ 状況を認識すること。
(人の表情、しぐさ雰囲気などが理解しにくく、人の感情がわかりにくい。
怒られているのに嬉しがったり、ほめられているのに知らん顔など・・・。)
④ 話し言葉への理解、自分からのコミュニケーションが難しい。
(言葉が出てもオウム返しになるなど。)
⑤ 感覚刺激に対して特異な反応をする。
(感覚刺激に対して過敏だったり鈍感だったりする。感覚刺激が一度にたくさん入り
すぎてしまう。特定の感覚刺激に苦痛を感じる。)

Human media information processing

- Unconscious processing
  - Mixed media processing
    - "I can see through my tongue."
    - Mixed sensation of synesthesia
  - Organizing principle for cerebral function (V. Mountcastle, 1978)
    - The unit of the cerebral cortex, called "column", has a very similar anatomical structure.
    - It implies that a universal information algorithm (common framework) exists in the cortex.

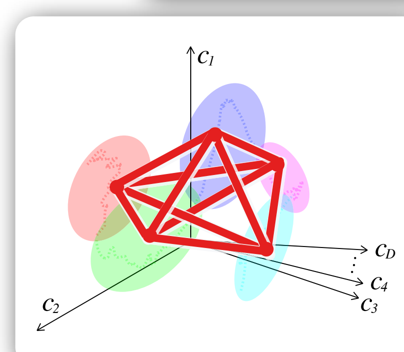


Title of each lecture

- 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 -

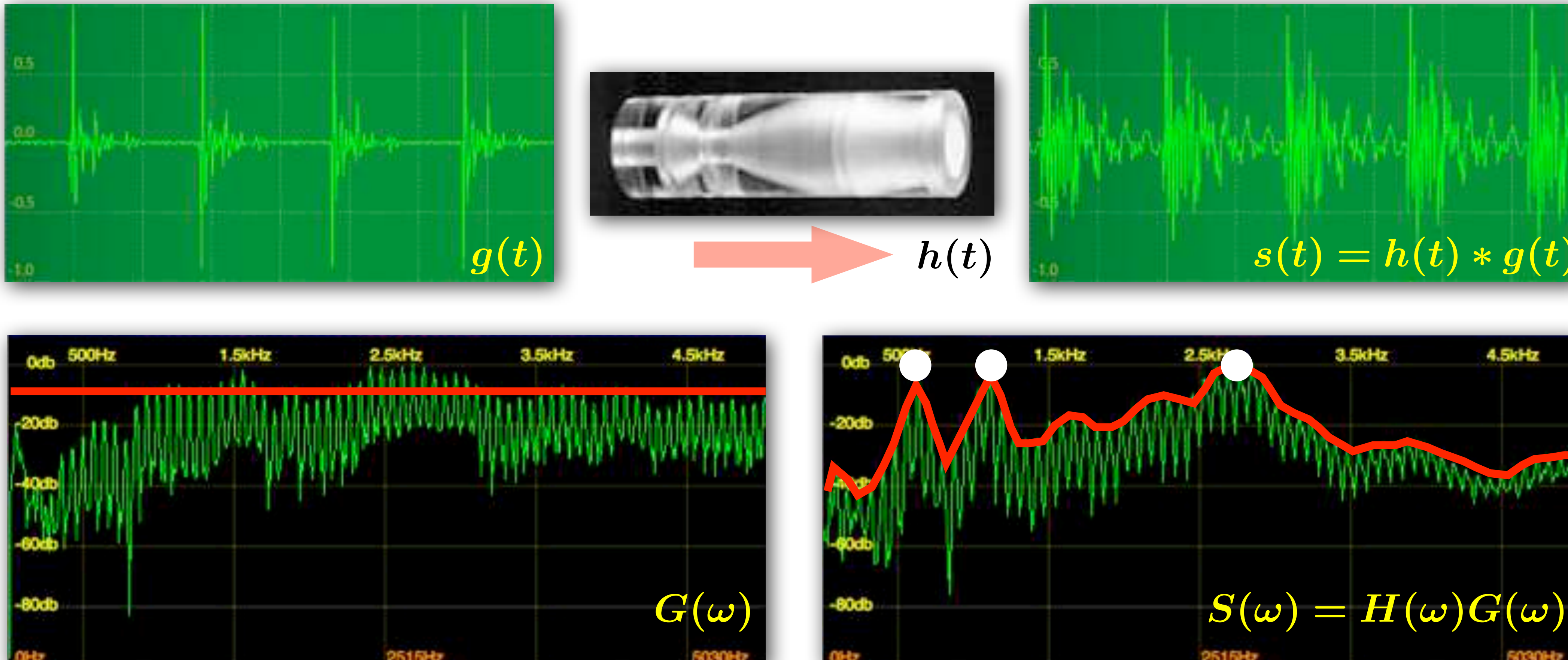
- Theme-2
  - Speech communication technology - articulatory & acoustic phonetics -
  - Speech communication technology - speech analysis -
  - Speech communication technology - speech recognition -
  - Speech communication technology - speech synthesis -

- 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



Acoustic phonetics

- Spectrum of a vowel sound



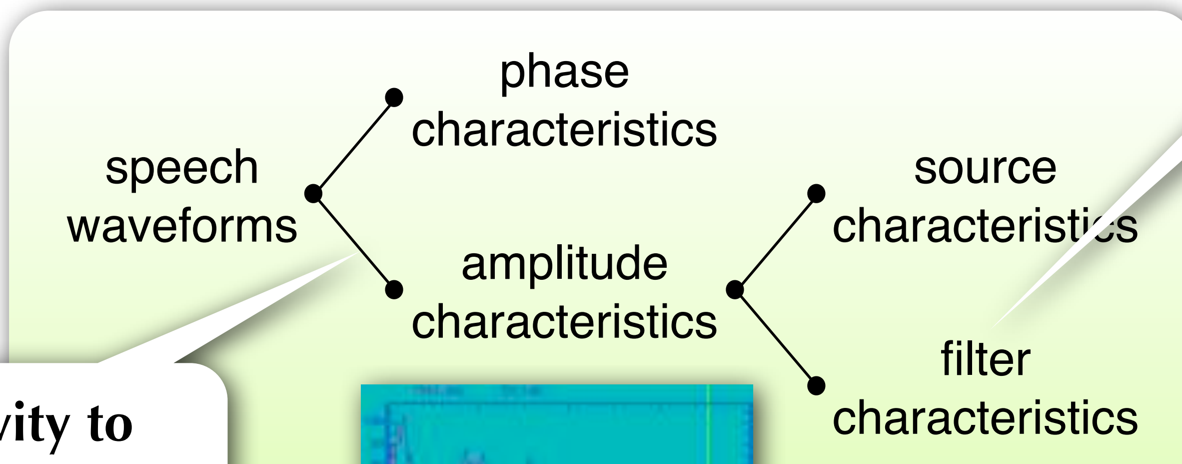
Resonance = concentration of the energy on specific bands that are determined only by the shape of a tube used for sound generation.

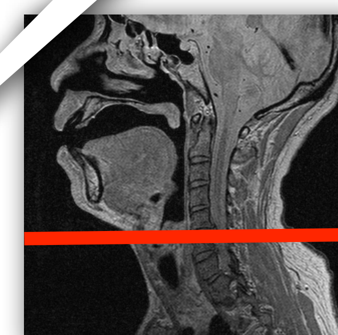
Timbre = energy distribution pattern over the frequency axis

Waveform to spectrum

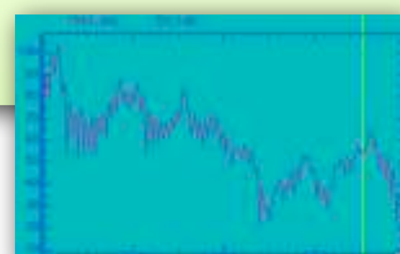
- From waveforms to spectrums
  - Windowing + FFT + log-amplitude
- Insensitivity of human ears on phase characteristics of speech
  - Human ears are basically “deaf” to phase differences in speech.
  - It is not impossible for us to discriminate acoustically two sounds with different phase characteristics but we don’t discriminate them linguistically.
    - No languages have those two sounds as two different *phonemes*.

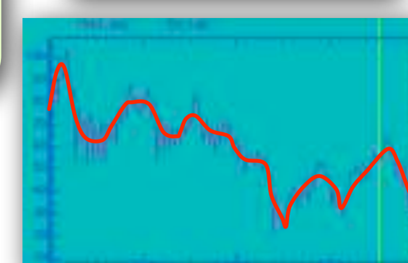
Insensitivity to pitch differences





Insensitivity to phase differences





Acoustic phonetics

- Other vowels = standing waves generated through a complicated tube

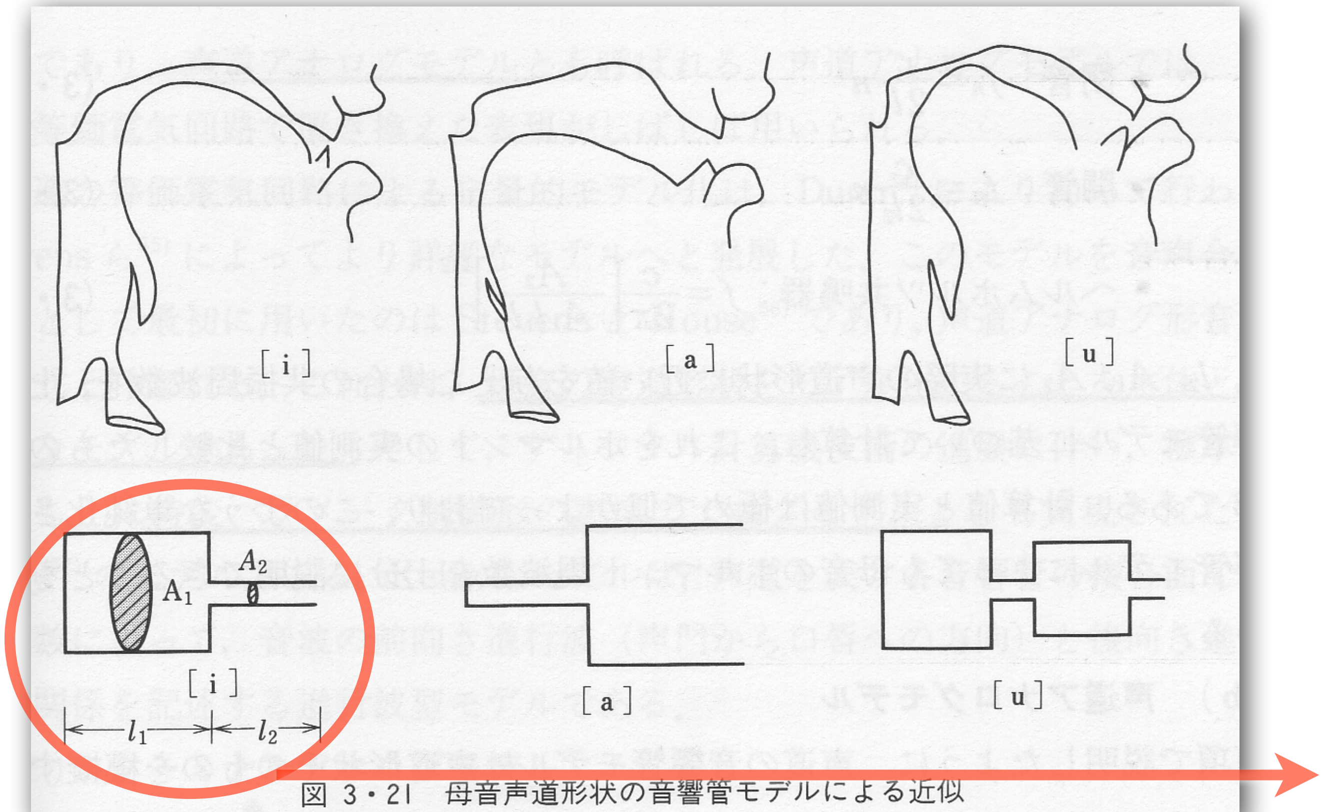


図 3・21 母音声道形状の音響管モデルによる近似

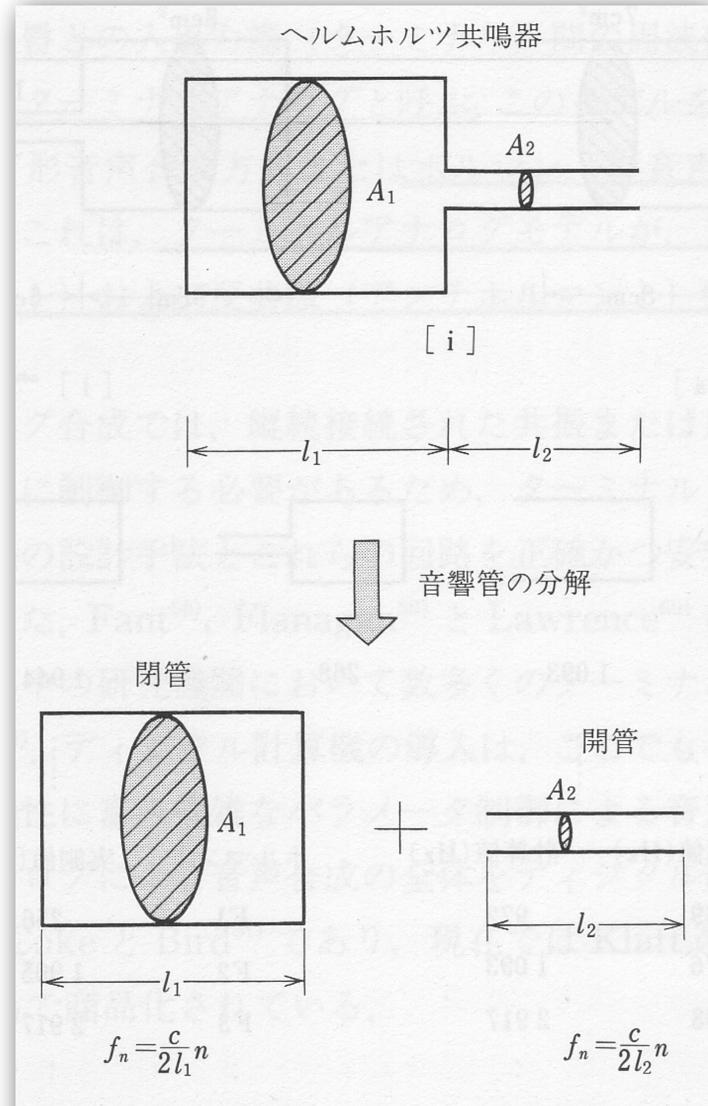
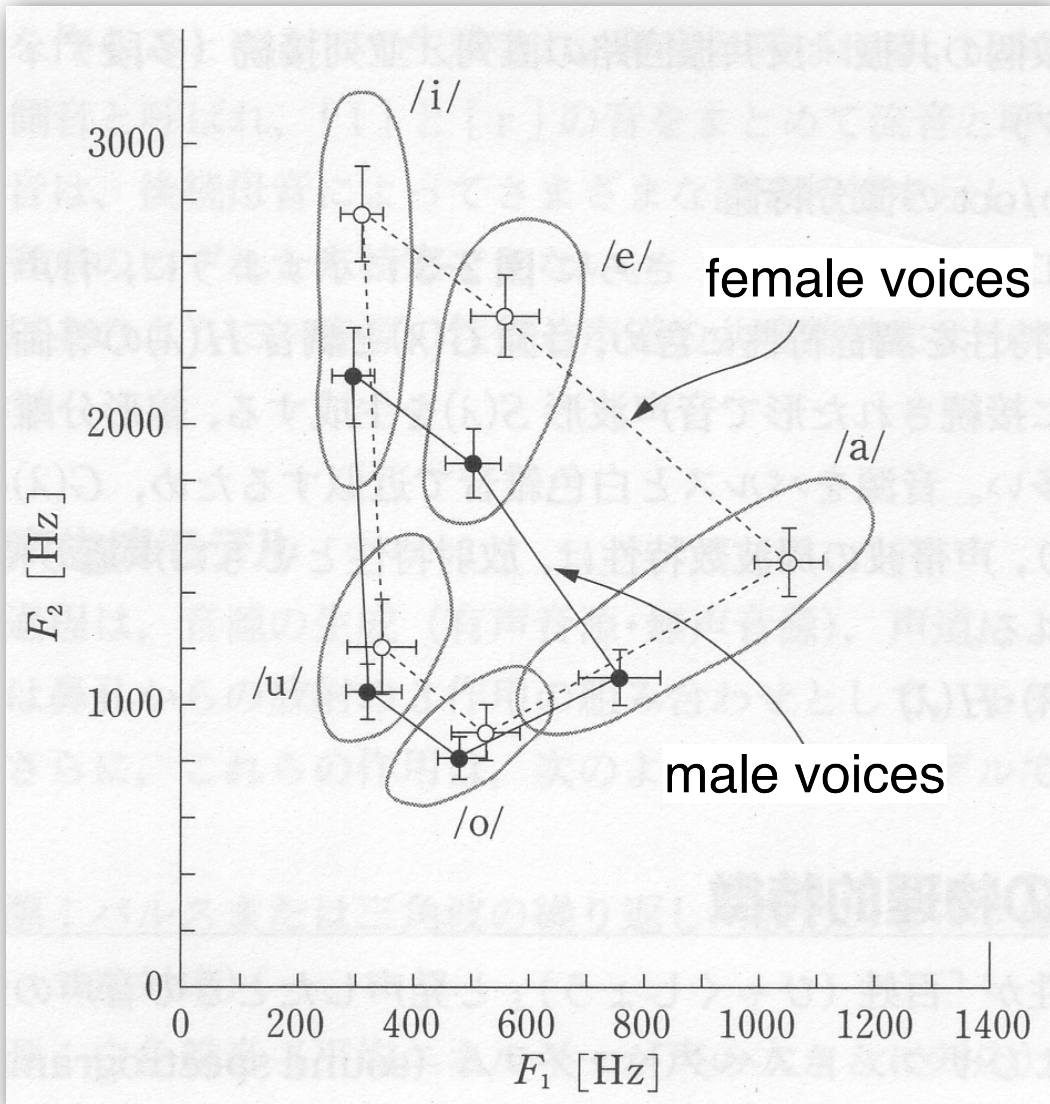


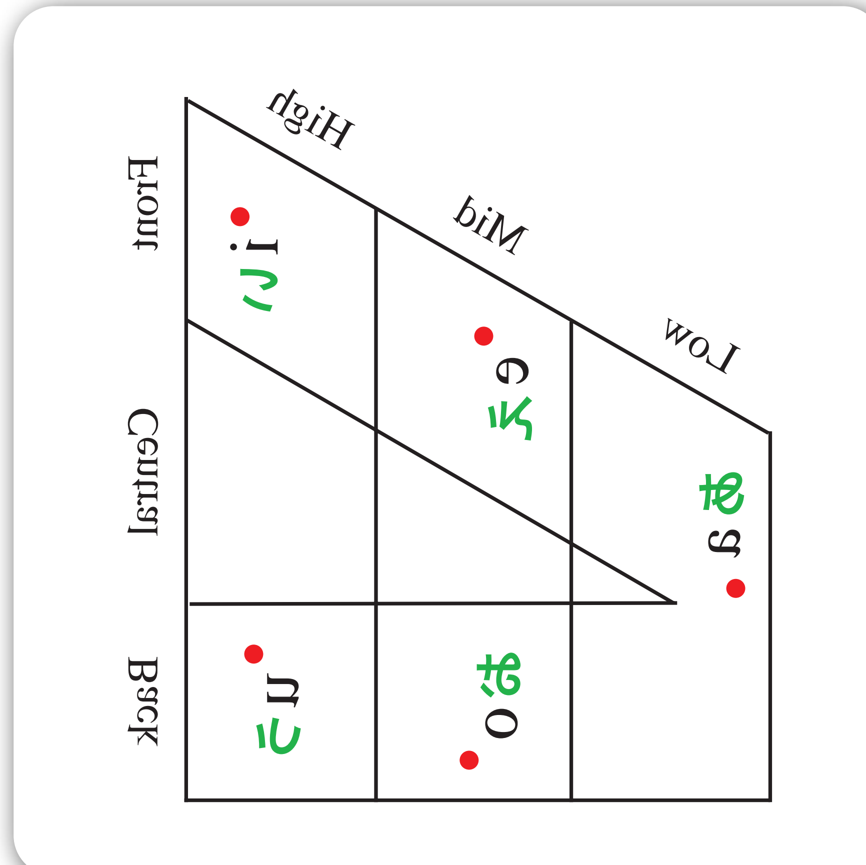
図 3・22 音響管モデルにおける音響管の分解

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

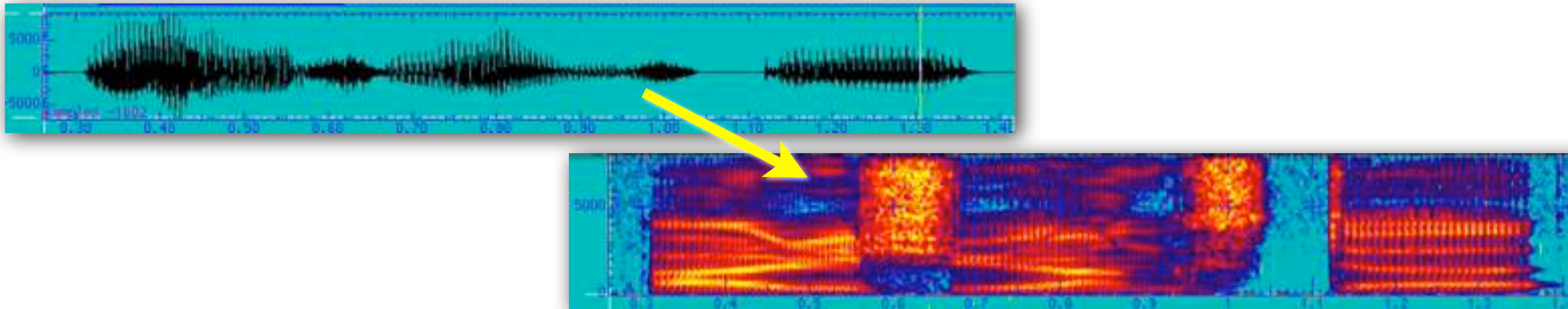
Acoustic and articulatory phonetics

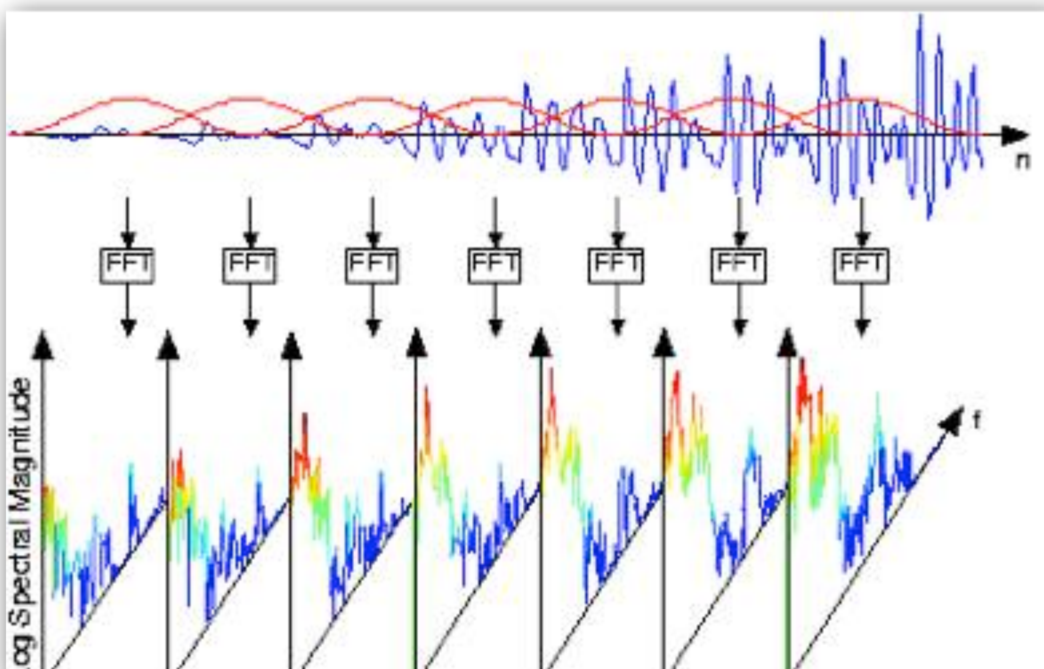
- Shape difference = resonance frequency difference
  - /a/ and /i/ /a/ and /a/

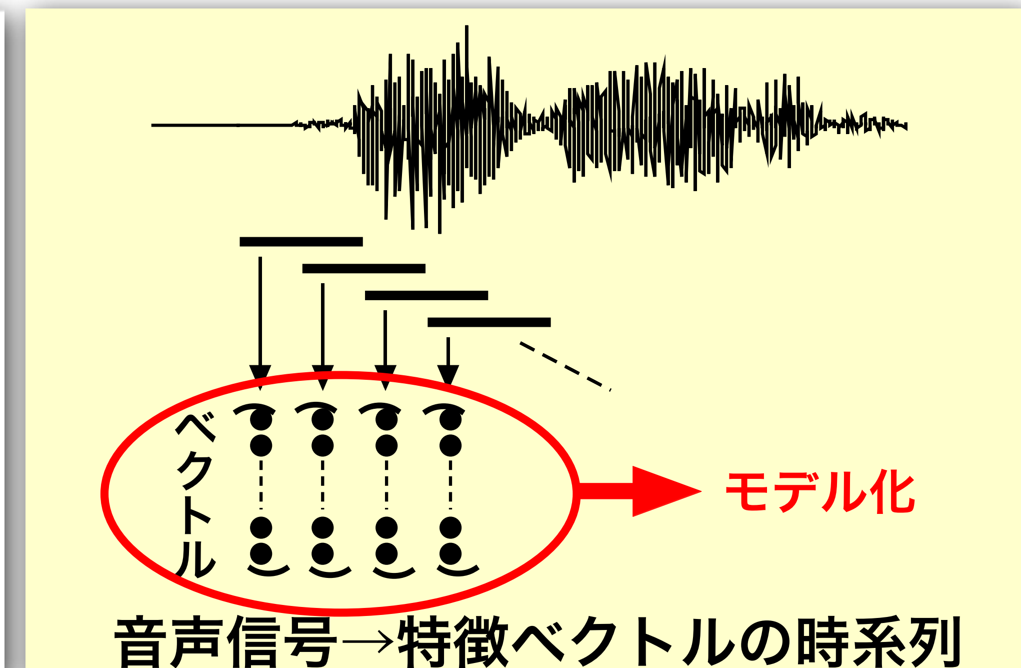




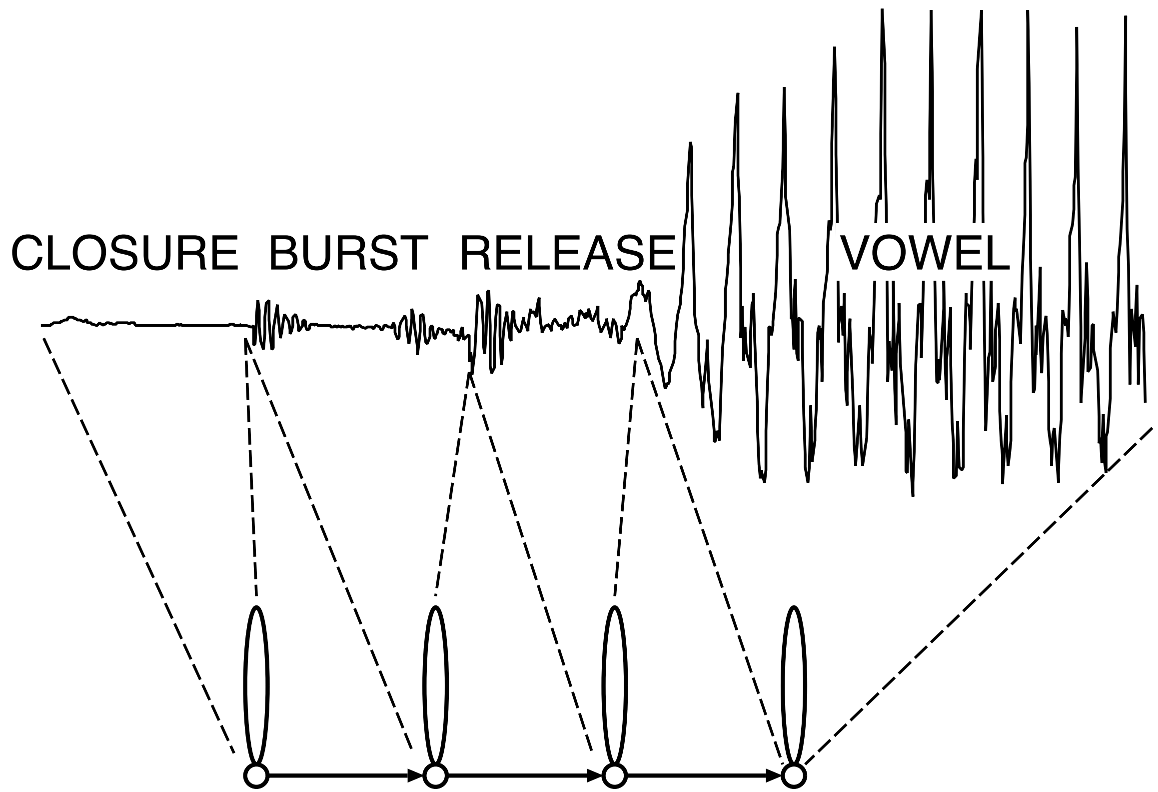
Waveforms --> spectrums --> sequence of feature vectors







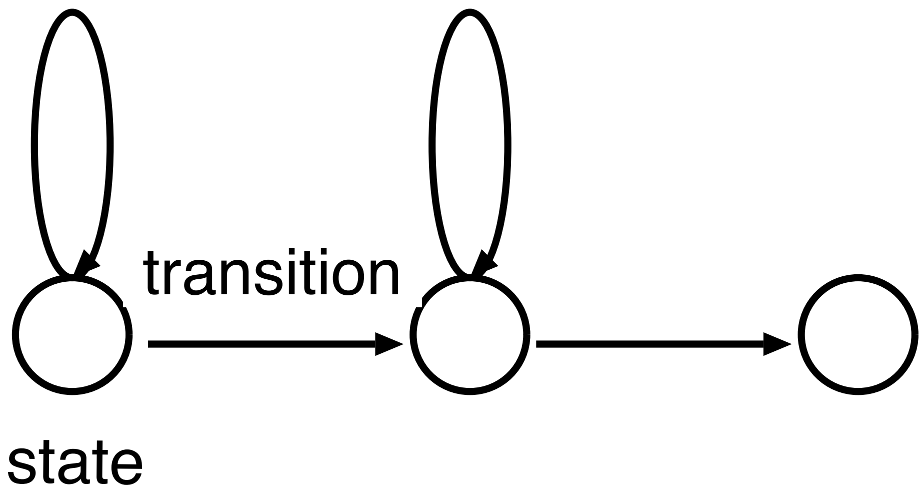
HMM as generative model



Probabilistic generative model

State transition is modeled as transition probability.
Output features are modeled as output probability.

Parameters of HMM



- Transition prob. : $P(s_{t+1}|s_t = i) = \{a_{1i}, a_{2i}, ..., a_{ji}, ..., a_{Si}\}$
- Output prob. : $P(o|s_t = i) = b_i(o)$

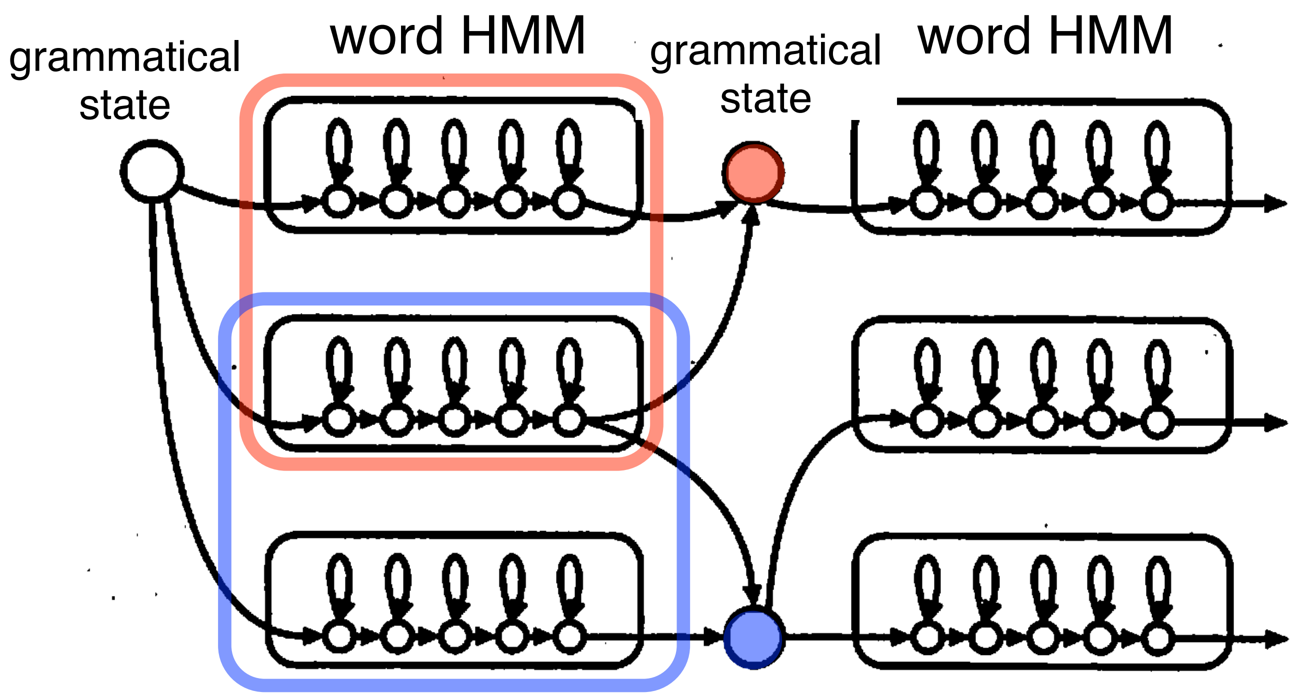
Forward prob.

$$\alpha_j(t) = P(o_1, \cdots, o_t, s(t) = j|M) \quad = \sum_i \alpha_i(t-1) a_{ij} b_j(o_t)$$

Backward prob.

$$\beta_j(t) = P(o_{t+1}, \cdots, o_T|s(t) = j, M) \quad = \sum_i a_{ji} b_i(o_{t+1}) \beta_i(t+1)$$

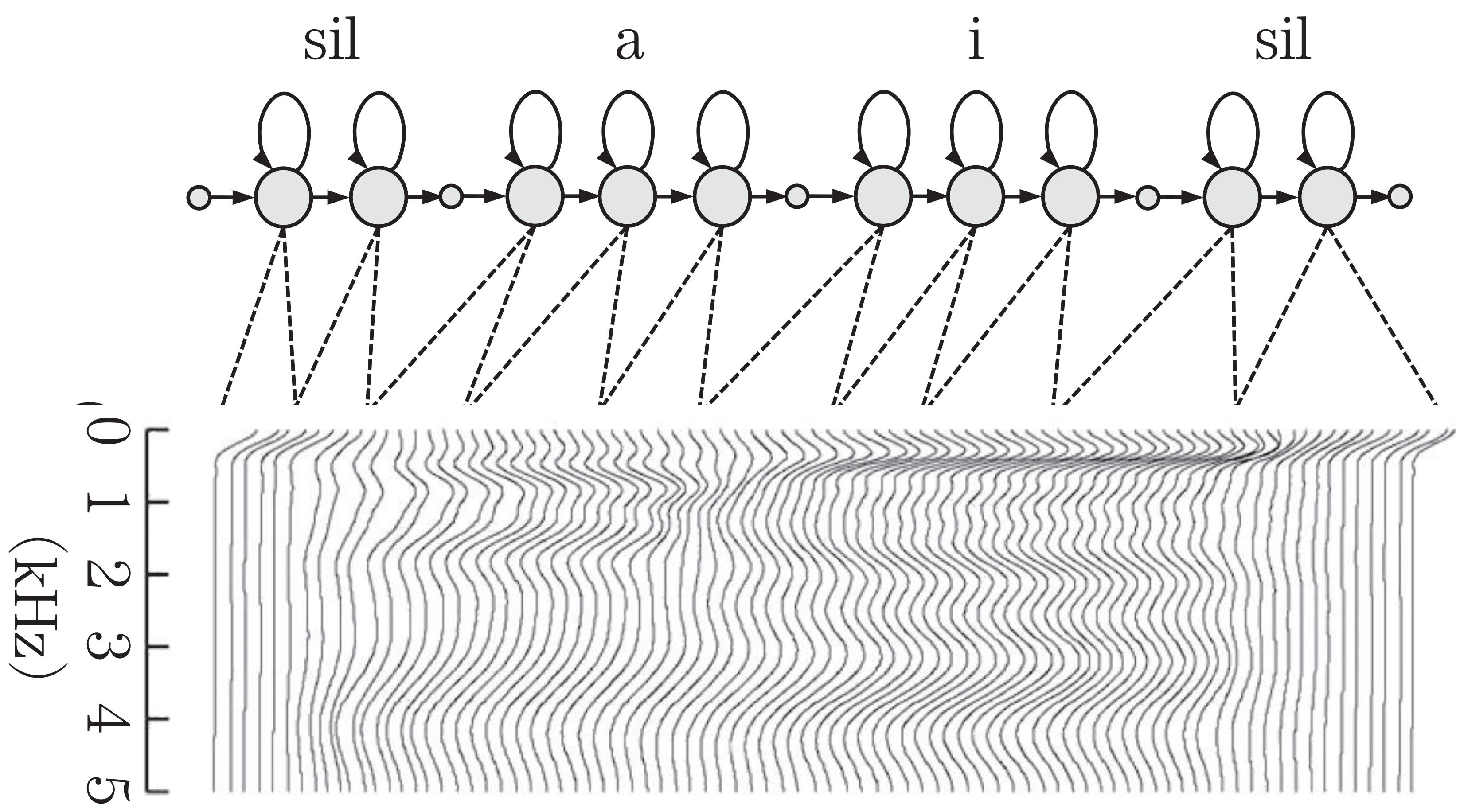
Speech recognition using a network grammar



When a grammatical state has more than one preceding words, the word of the maximum probability (or words with higher probabilities) is adopted and it will be connected to the following candidate words.

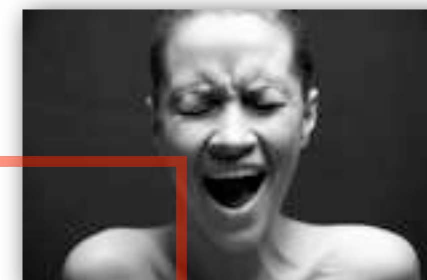
Spectrum generated from HMMs

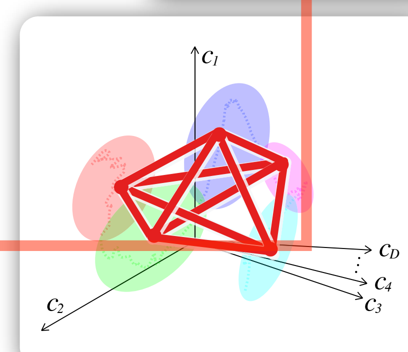
- Text -> HMM seq. -> most likely state seq. -> most likely spectrum seq.
  - The most likely spectrum from a state = mean vector (spectrum) of the state
    --> the spectrum sequence has to have stepwise abrupt changes.



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





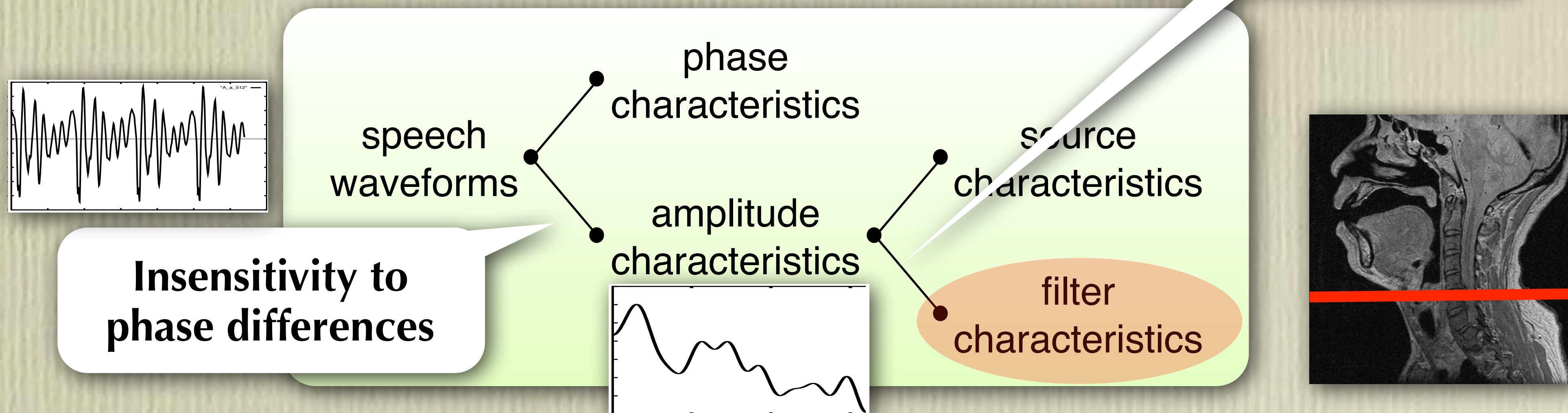
Speech is extremely variable.

- Various factors change speech acoustics easily.

- The world's tiniest high school girl

Feature separation to find specific info.

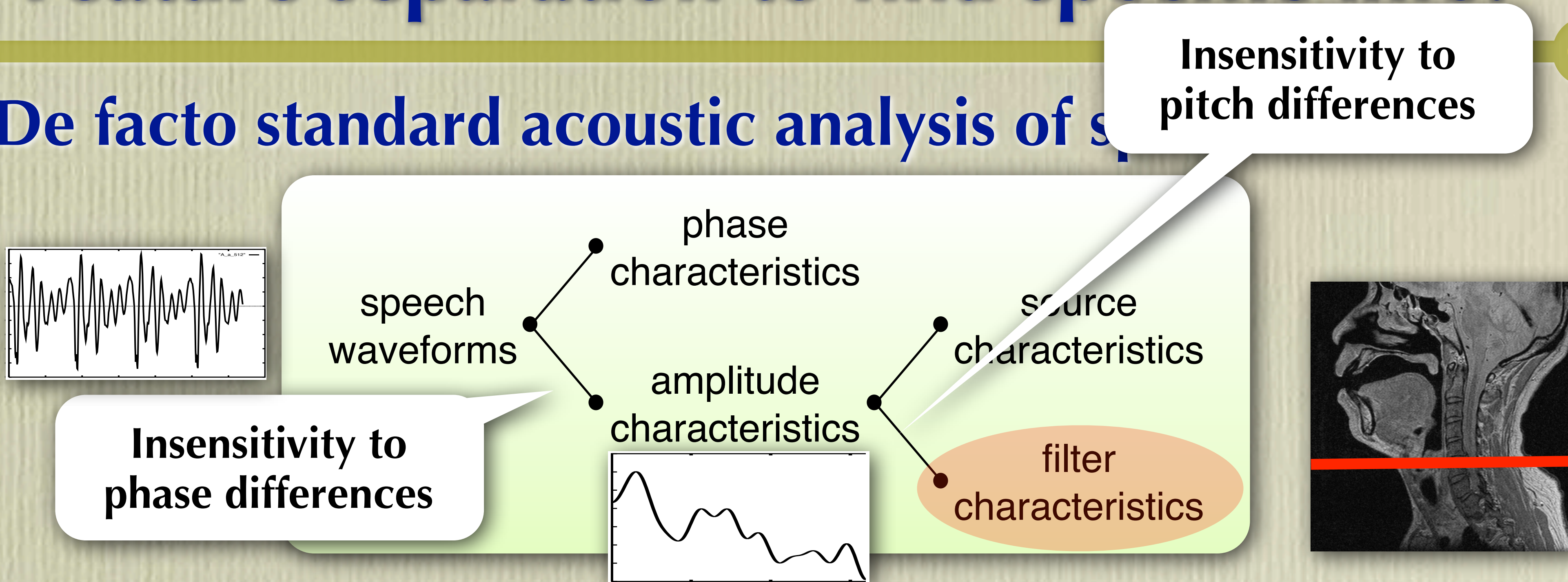
De facto standard acoustic analysis of s



- 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 those variation through sample collection
  - Physical independence: pursuing features invariant to those variation
  - :

Feature separation to find specific info.

De facto standard acoustic analysis of s

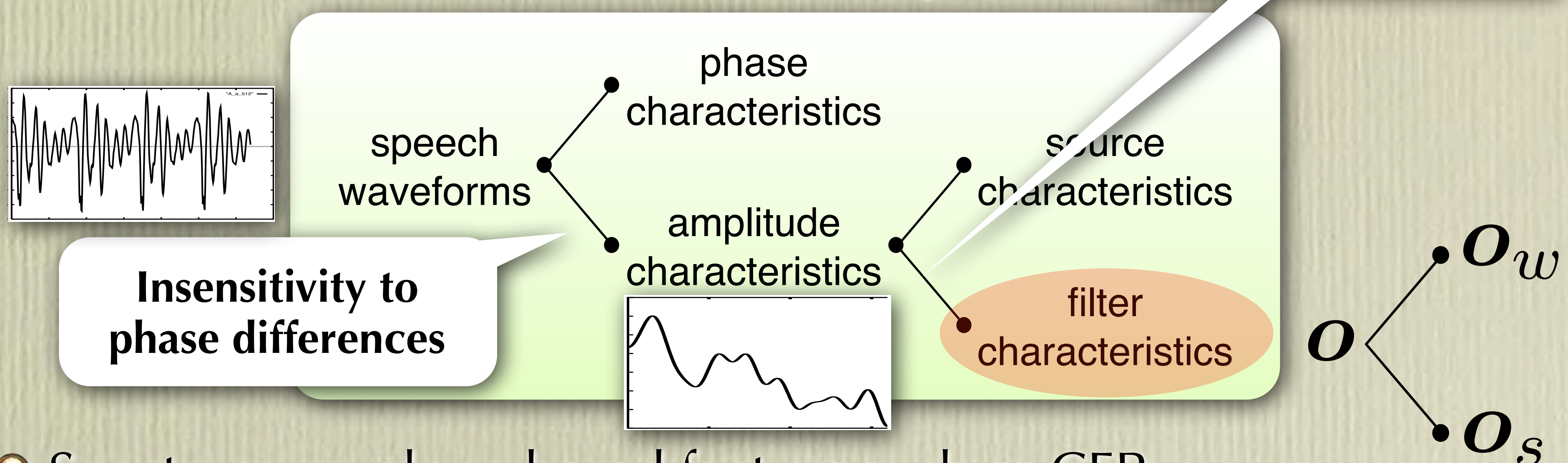


Two acoustic models for speech/speaker recognition

- 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?

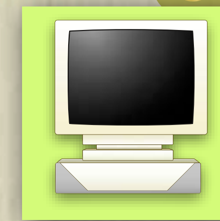
Feature separation to find specific info.

De facto standard acoustic analysis of s



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

A difference bet. machines and humans



Machine strategy (engineers' strategy): ASR

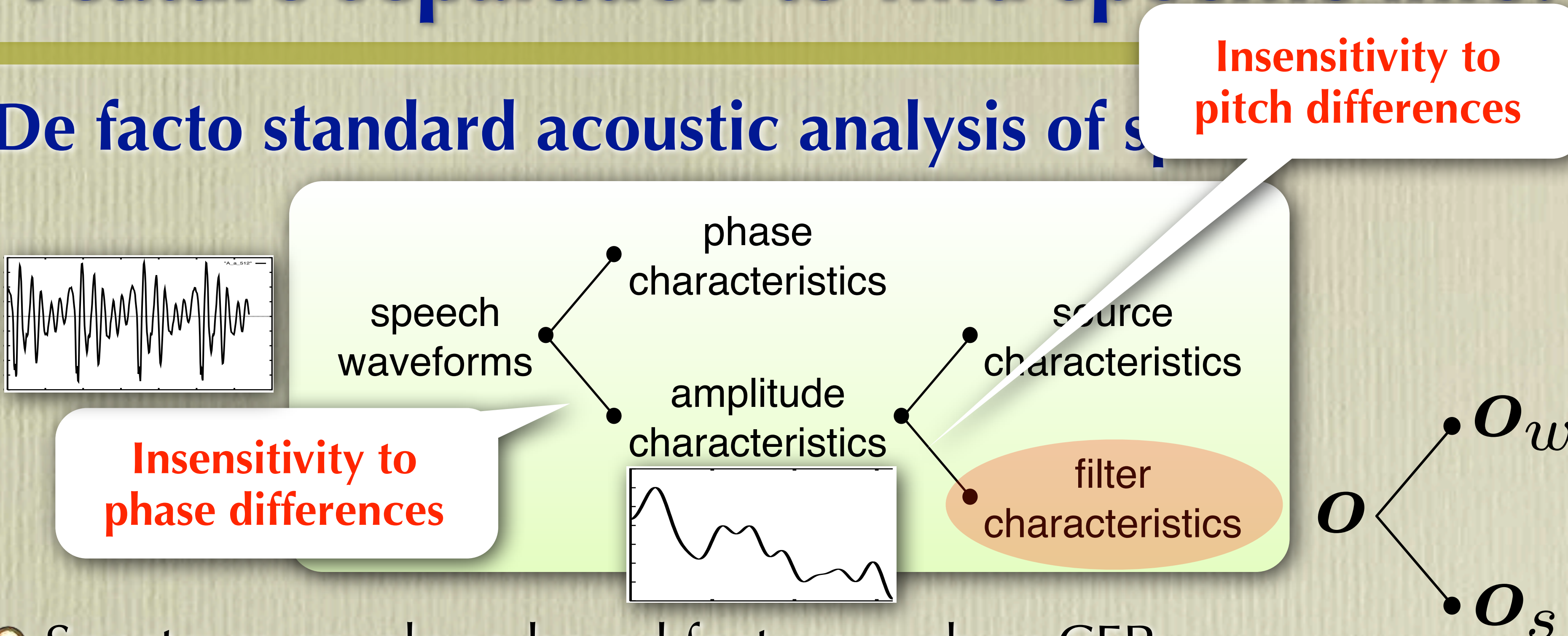
- Collecting a huge amount of speaker-balanced data
  - Statistical training of acoustic models of individual phonemes (allophones)
- Adaptation of the models to new environments and speakers
  - Acoustic mismatch bet. training and testing conditions must be reduced.

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!!

Feature separation to find specific info.

De facto standard acoustic analysis of s



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

Insensitivity in our language learning

Vocal learning (including vocal imitation)

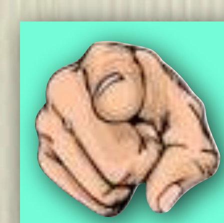
- A imitate(s) B vocally.
  - A: students and B: teachers
  - A: infants and B: parents (caretakers)
  - A: you and B: professional singer (Karaoke)
  - But A do not impersonate B.
    - Acoustically *mis*matched imitation.
- We're very insensitive to speaker identity transmitted via speech.

Acoustically matched imitation is often found in

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

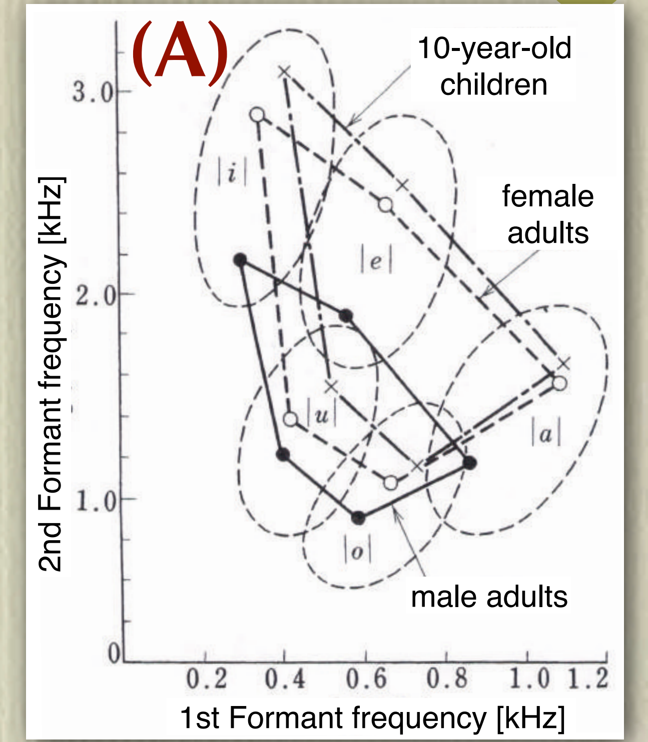
 → 

 → 

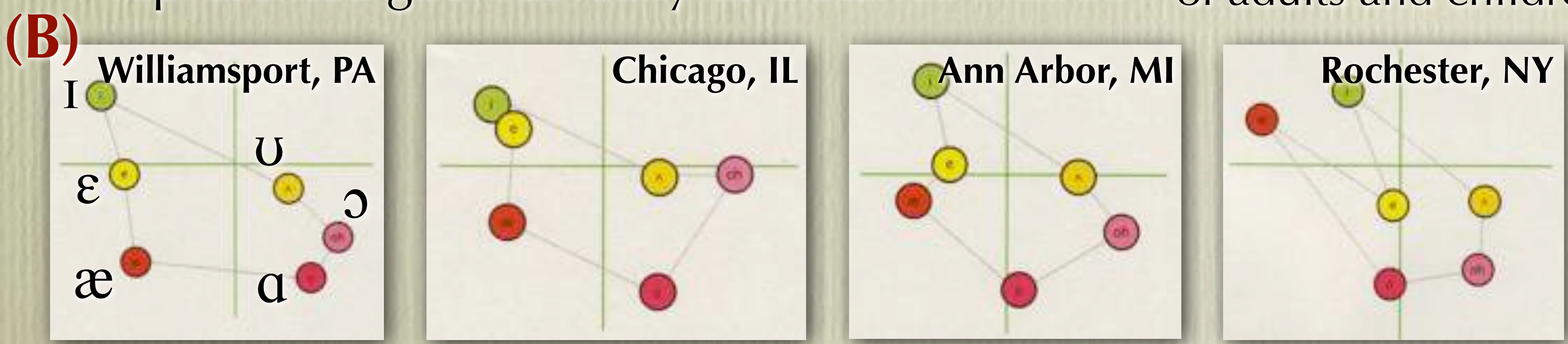
 → 

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



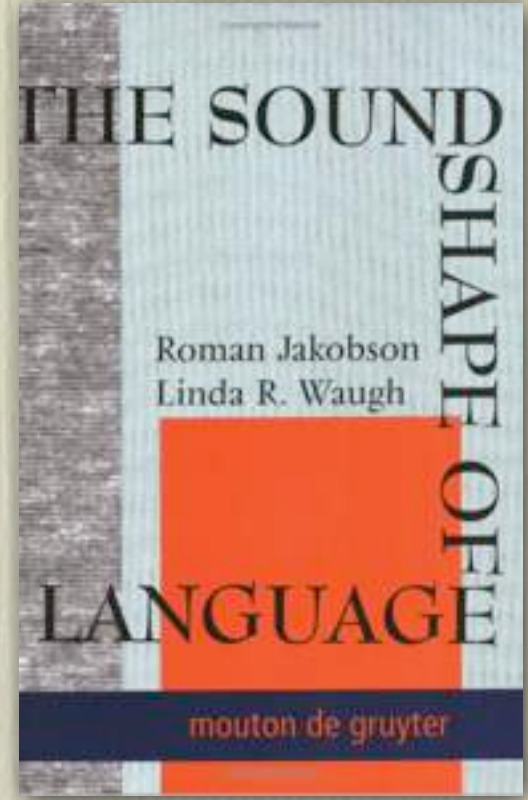
Distribution of normalized formants among AE dialects [Labov et al.'05]

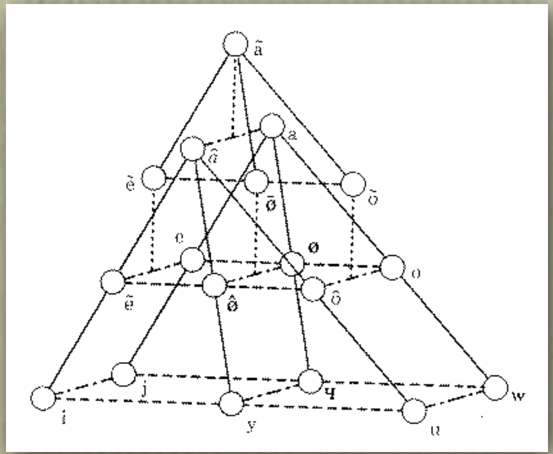
A claim found in classical linguistics

- **Theory of relational invariance** [Jakobson+'79]
  - Also known as theory of distinctive feature
  - 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.

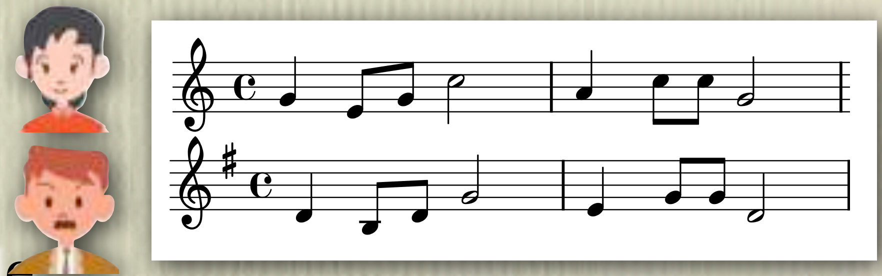
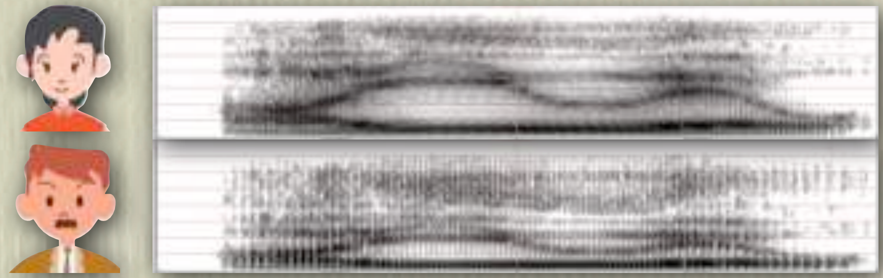




Menu of the last four lectures

- **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 perceive the equivalence very easily.**

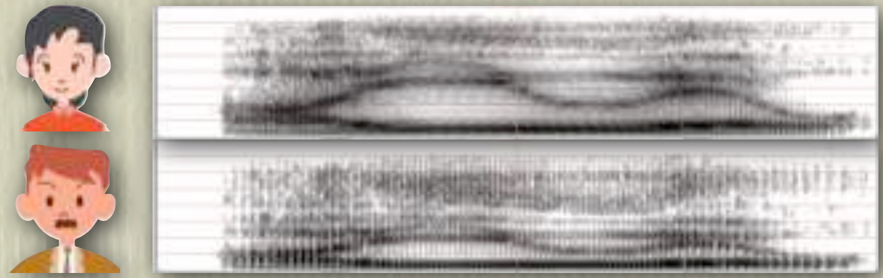
Physical variability and cognitive constancy

- **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

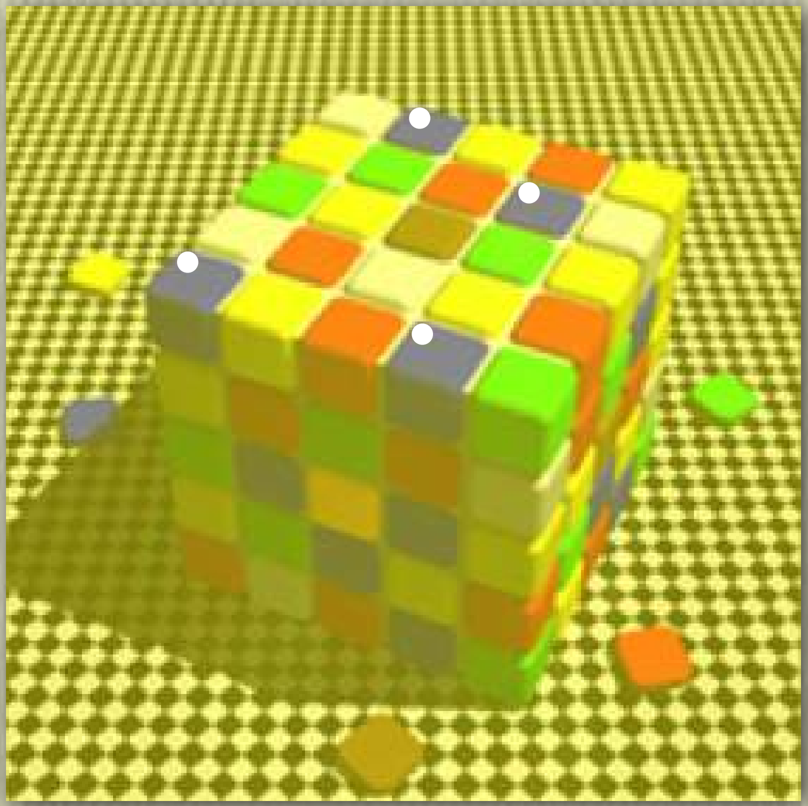
    - Key change (transposition) of a melody

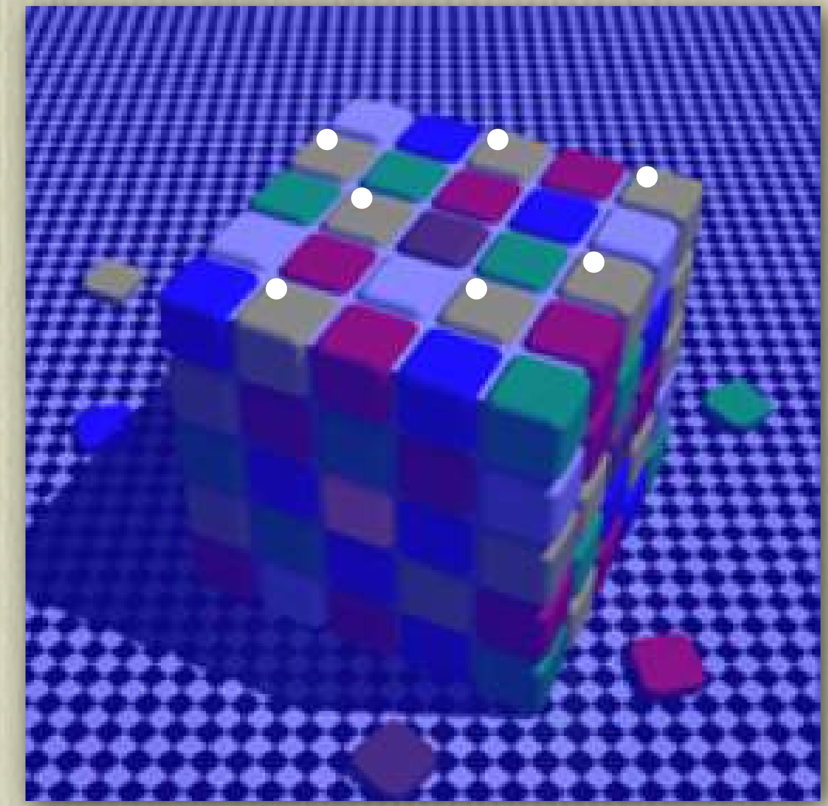


  - Variability in timbre
    - A male's "hello" and a female's
    - An adult's "hello" and a child's
- **But we can perceive the equivalence very easily.**

Invariant color perception against its bias

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

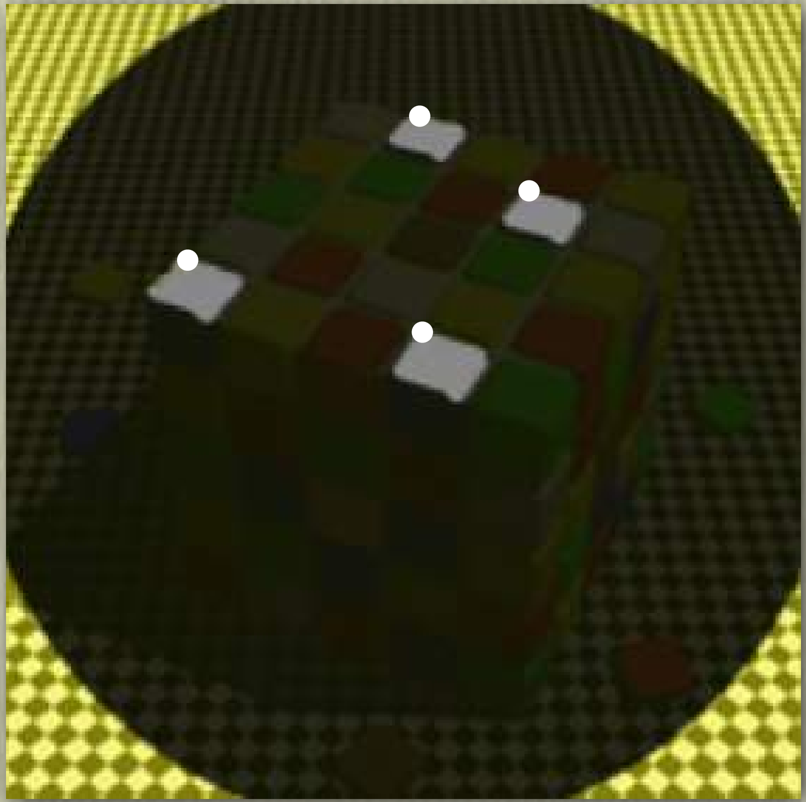


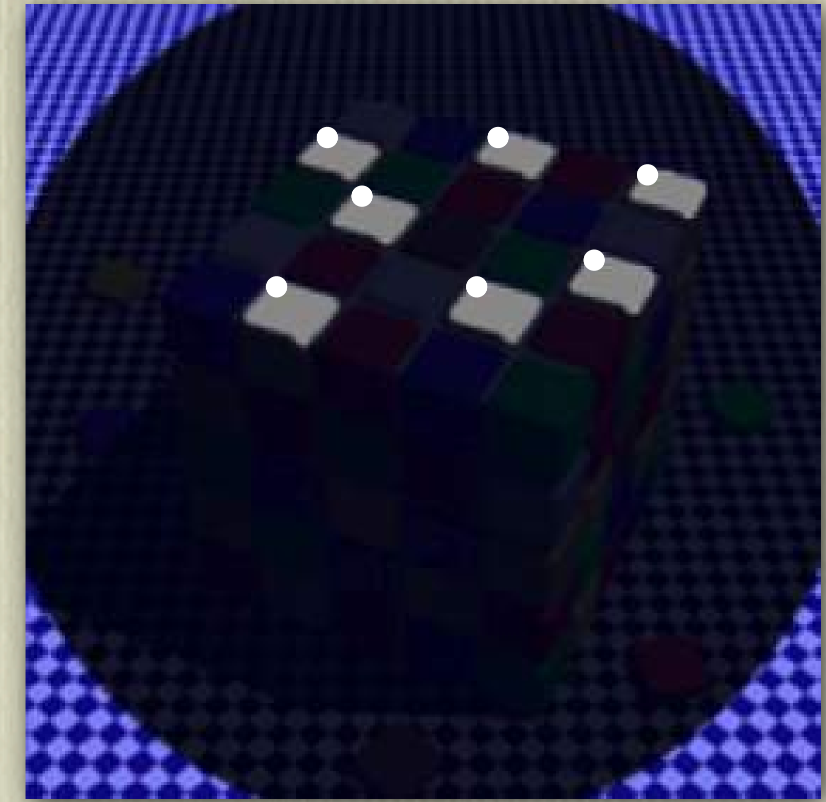


- We perceive that the two cubes are identical.
- Different / identical colors are claimed to be identical / different.
- Not only wavelength (absolute property) of each patch, but also it matters **what contrast each patch has to its surrounding patches.**

Invariant color perception against its bias

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

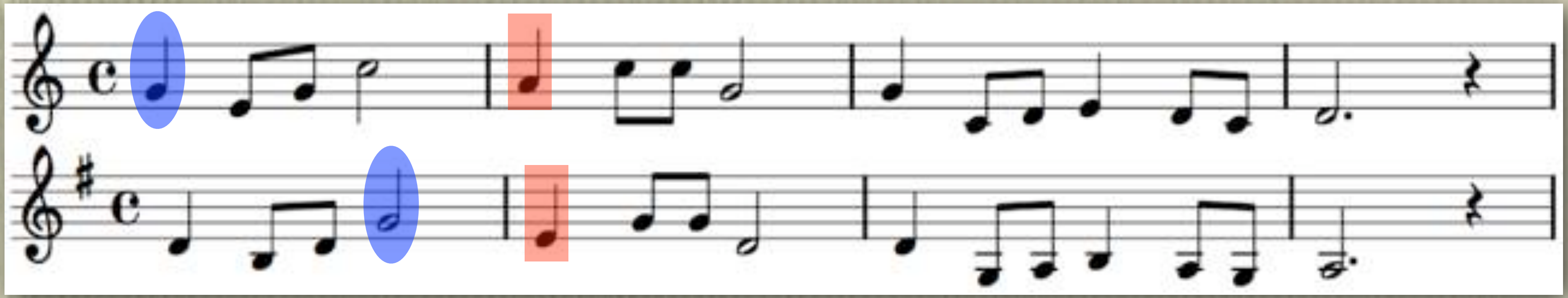




- We perceive that the two cubes are identical.
- Different / identical colors are claimed to be identical / different.
- Not only wavelength (absolute property) of each patch, but also it matters **what contrast each patch has to its surrounding patches.**

Invariant pitch perception against its bias

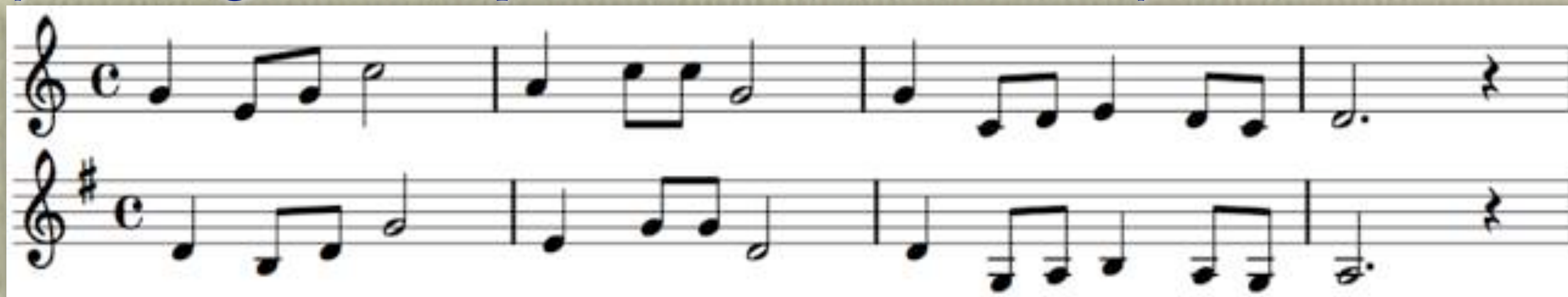
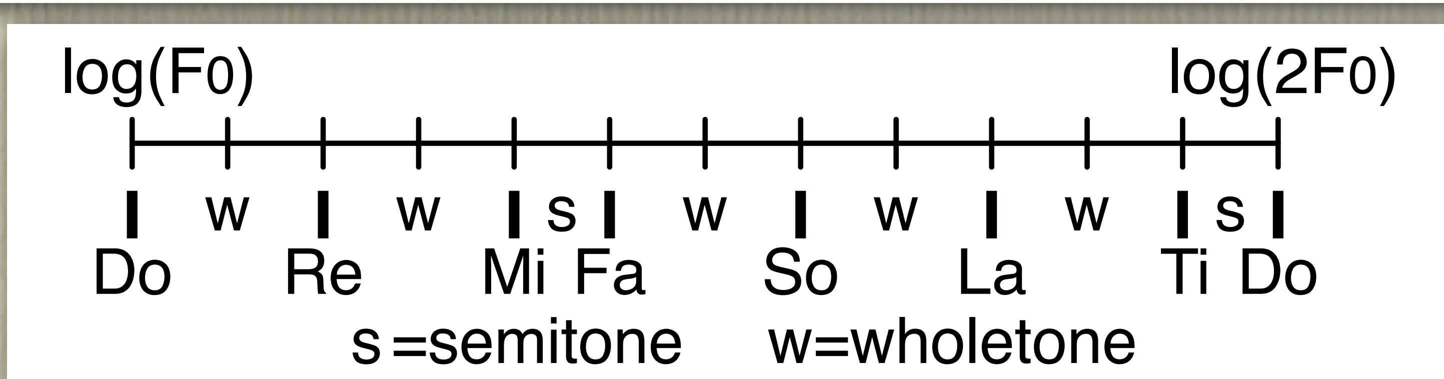
Key change (transposition) of a melody [Higashikawa'05]

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

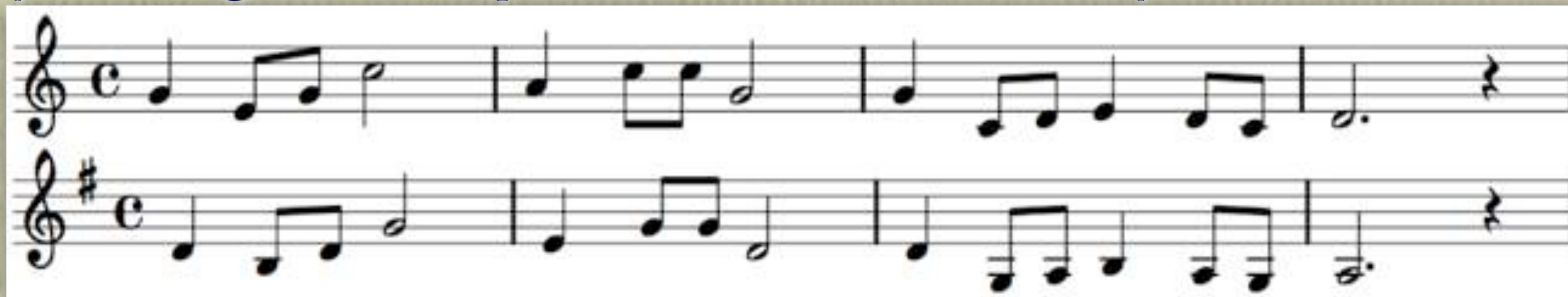
Key change (transposition) of a melody [Higashikawa'05]

1
2



- 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

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

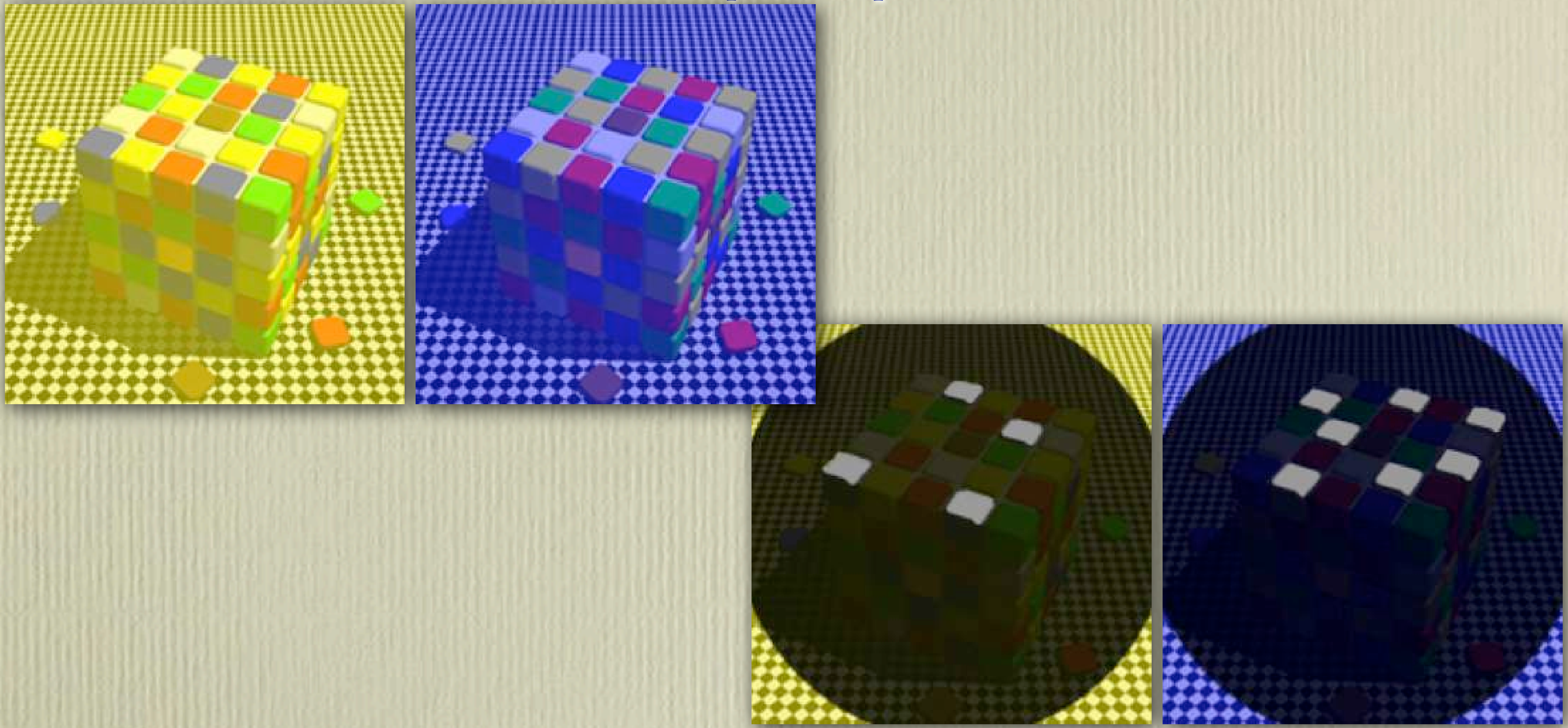
1
2


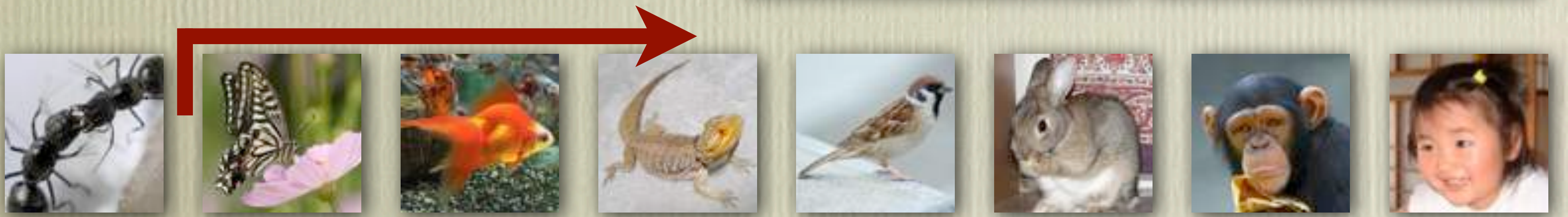
But it is very difficult to label a single tone because there is no contrast at all.

- Not fundamental frequency (absolute property) of each tone, but it only matters **what contrast each tone has to its surrounding tones.**

The nature's solution for static bias?

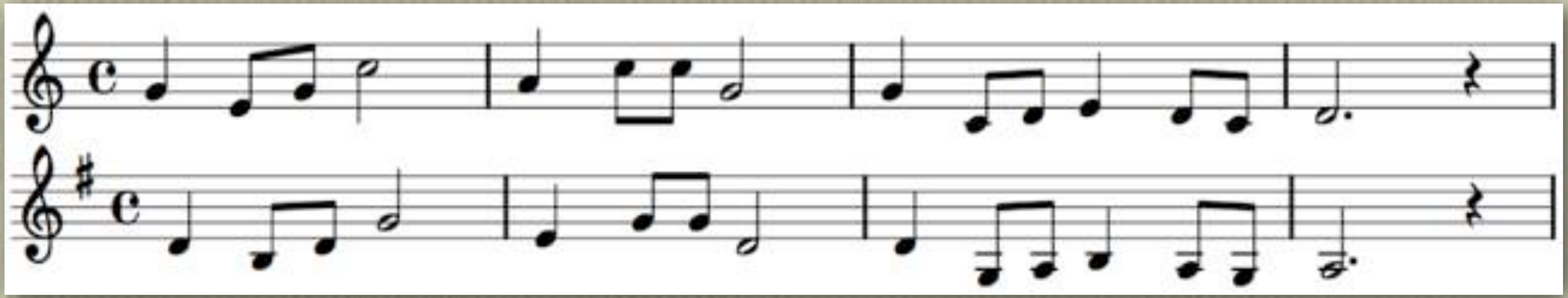
- **How old is the invariant perception in evolution?** [Briscoe'01]





The nature's solution for static bias?

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

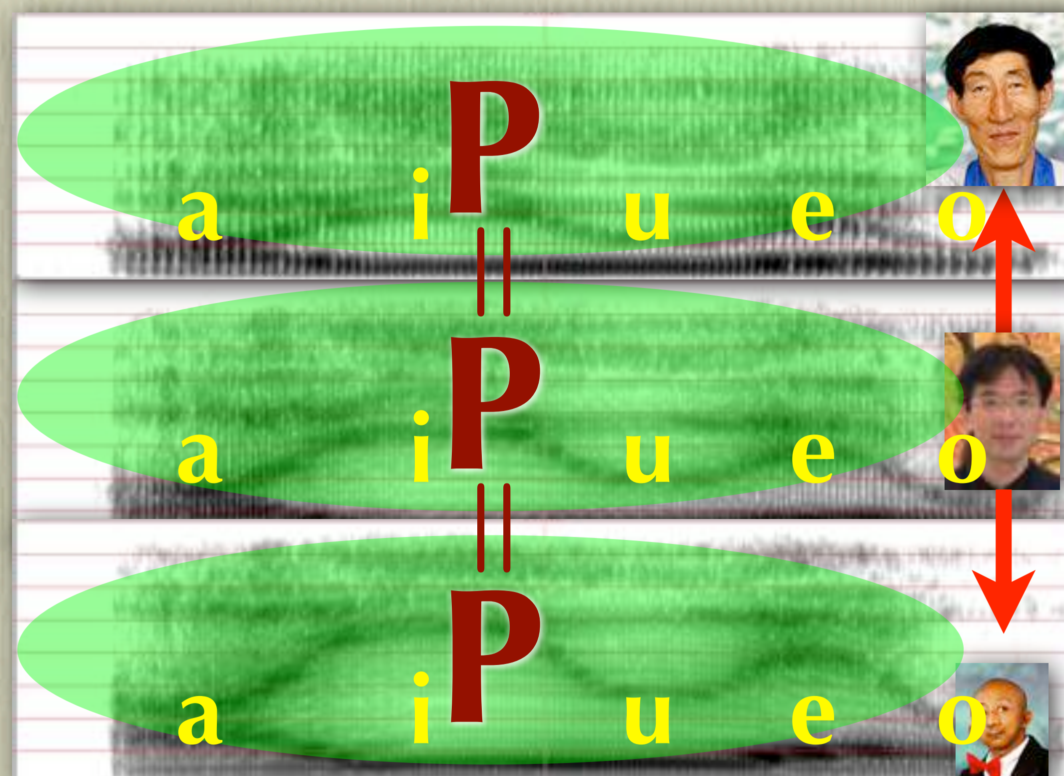
1
2

1 = 2

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



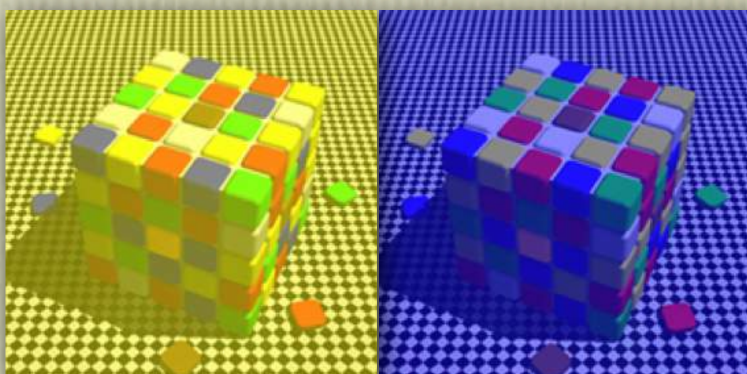
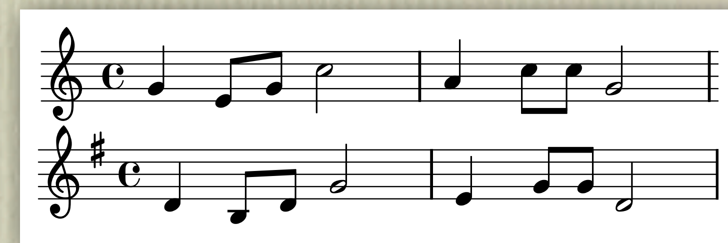
Tiniest high school girl!!

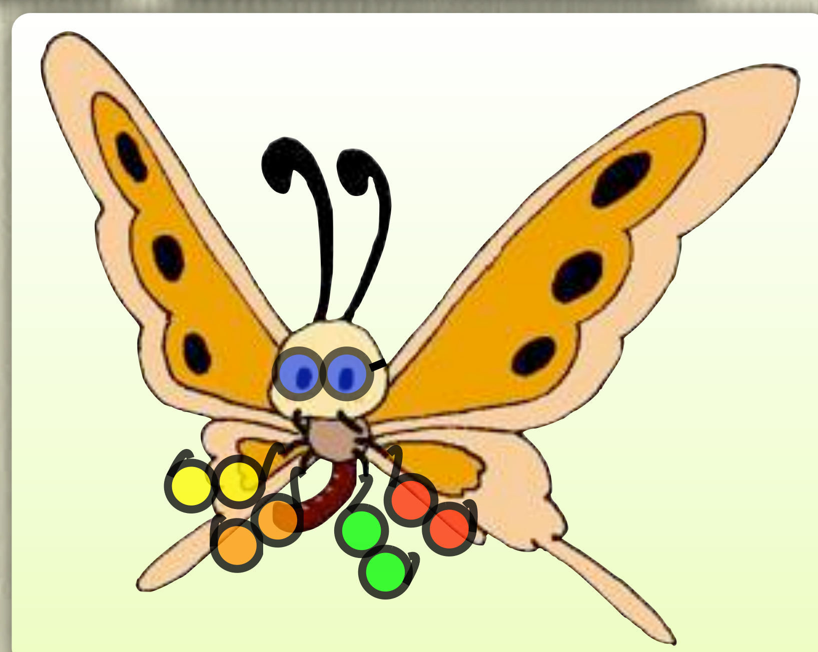
- **Linearly reduced individual!?**

Invariant timbre perception against its bias

Invariant and constant perception wrt. color and pitch

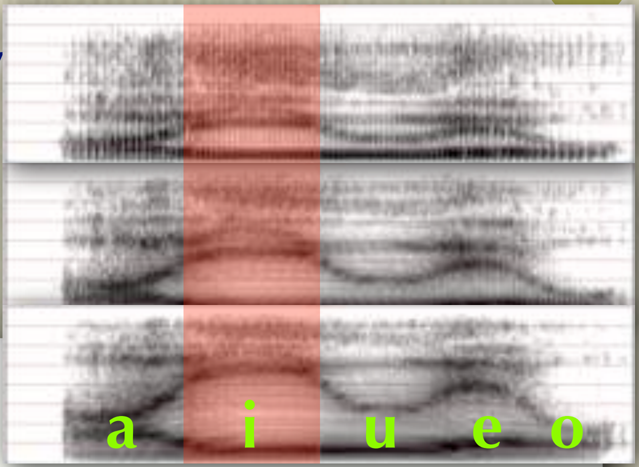
- Contrast-based information processing is important.
- Holistic & relational processing enables element identification.





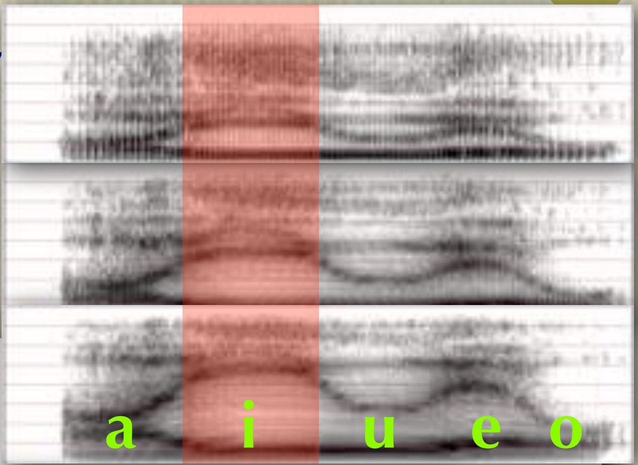
Invariant timbre perception against its bias

- **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

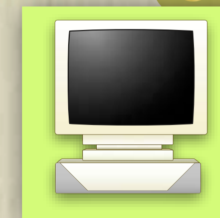


De facto standard for timbre variability

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

o
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.

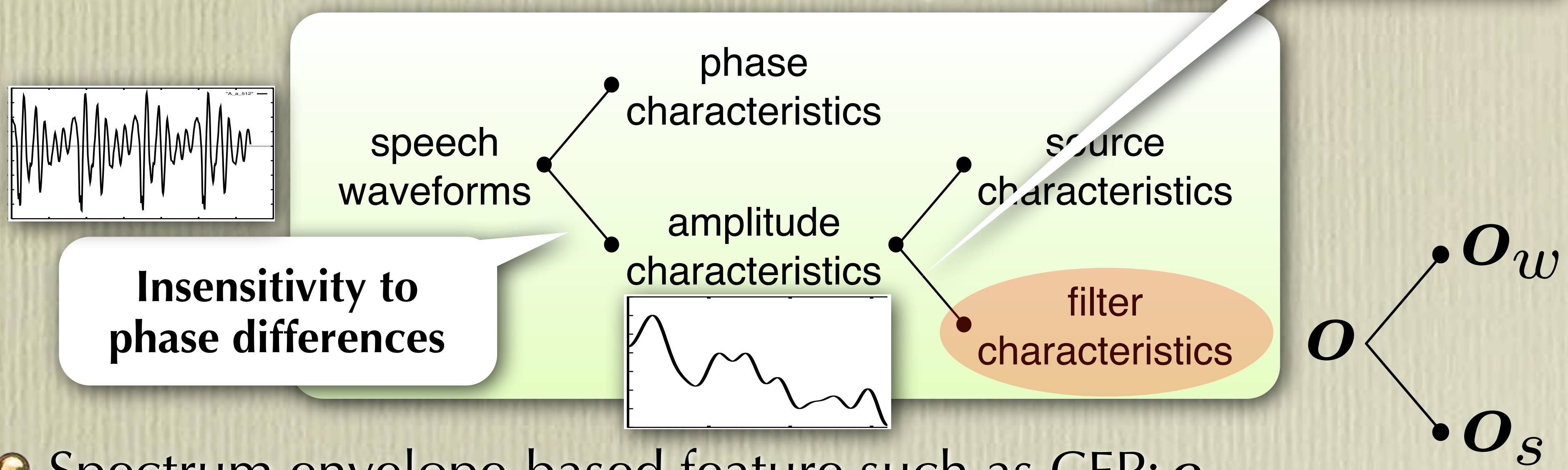
?

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: $\boldsymbol{o}$
  - But $\boldsymbol{o}$ depends on all the three kinds of info. (ling, para-ling, extra-ling).
- How to suppress extra-linguistic variation in $\boldsymbol{o}$?
  - Feature normalization: transforming $\boldsymbol{o}$ to that of the standard speaker
  - Model adaptation: modifying model parameters to fit to the input speaker
  - Statistical independence: hiding those variation through sample collection
  - Physical independence: pursuing features invariant to those 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 = ??

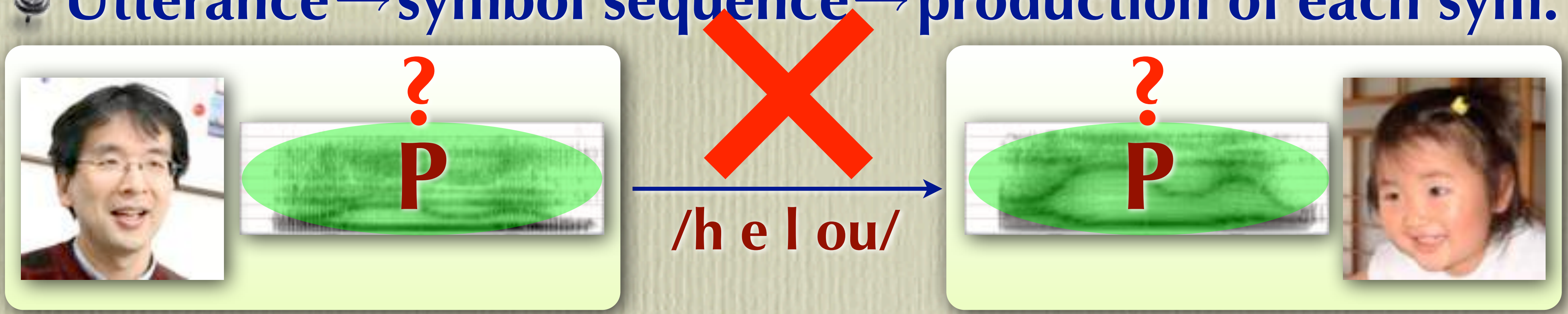
- Acoustic imitation performed by myna birds [Miyamoto'95]
  - They imitate the sounds of cars, doors, dogs, cats as well as human voices.
  - Hearing a very good myna bird say something, one can guess its owner.
- Beyond-scale imitation of utterances performed by children
  - No one can guess a parent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





Language acquisition through vocal imitation

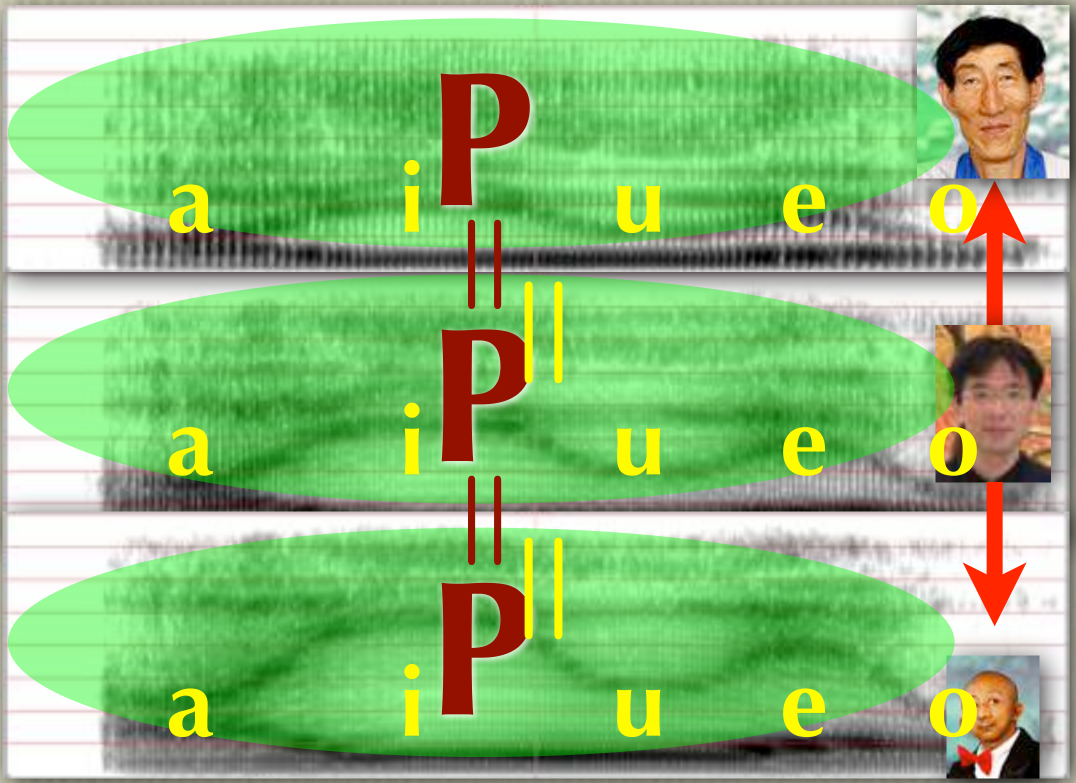
- **Utterance → symbol sequence → production of each sym.**



  - 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?

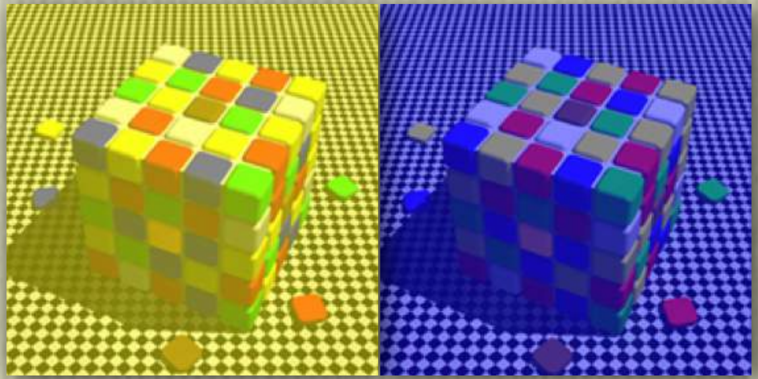
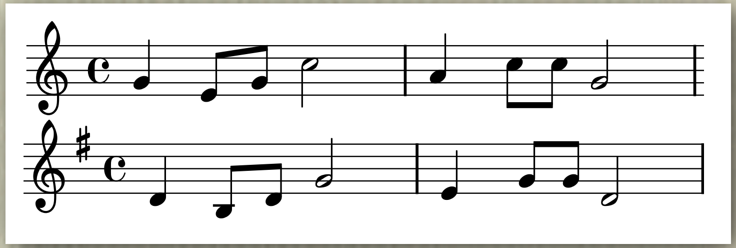
Invariant timbre perception against its bias

- **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.