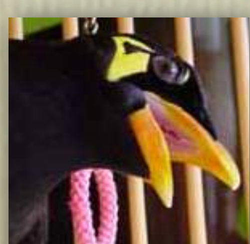
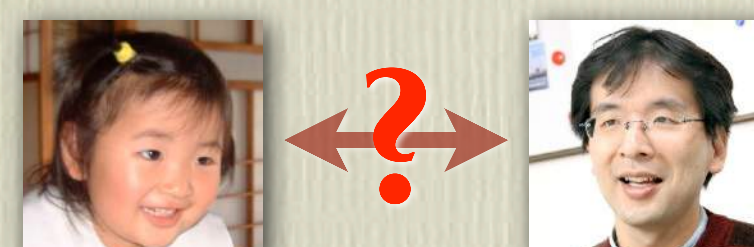
Cognitive Media Processing #11

Nobuaki Minematsu

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].

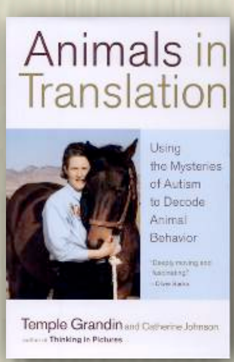
Claims from a professor of animal sciences

Dr. Temple Grandin @ Colorado State University

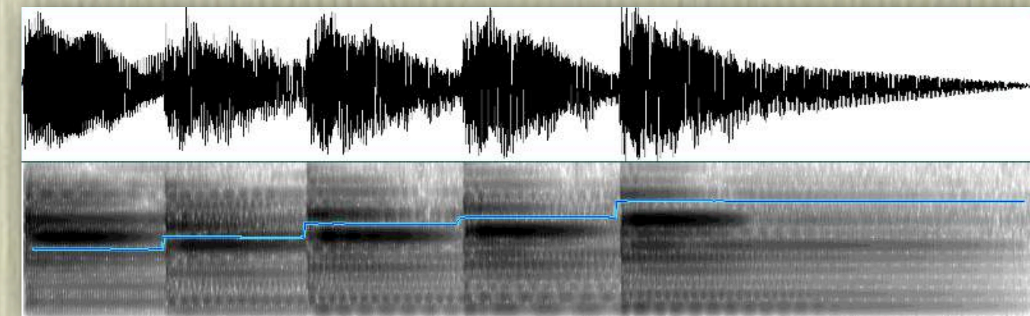
- She is herself autistic (Asperger syndrome).
- Autistics often imitate the utterances of TV/radio commercials.
  - TV/radio often gives "acoustically" identical utterances.
  - The utterances from family members change "acoustically" time to time.
- They often imitate the sounds of objects such as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

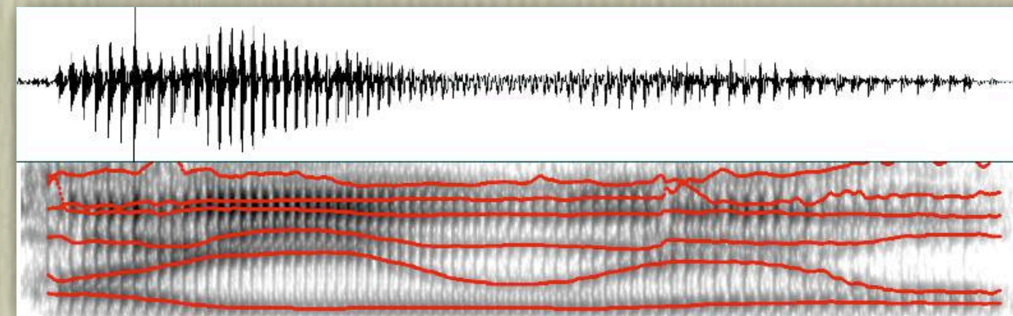
Interesting claims from her

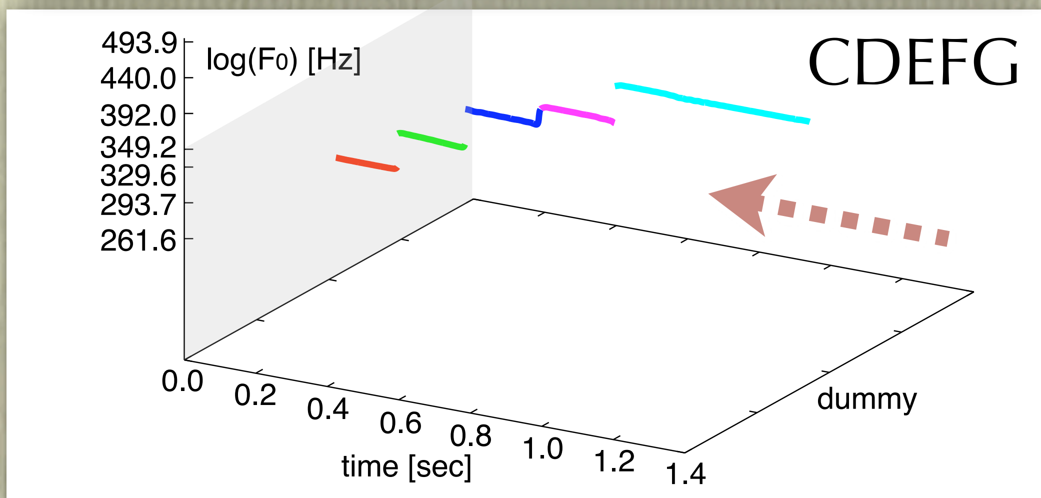
- Similarity of information processing between animals and autistics
- Storing the detailed aspects of input stimuli as they are in the brain
  - Animal : local / detail / absolute
  - Human : holistic / abstract / relative
    - Good ability to generalize

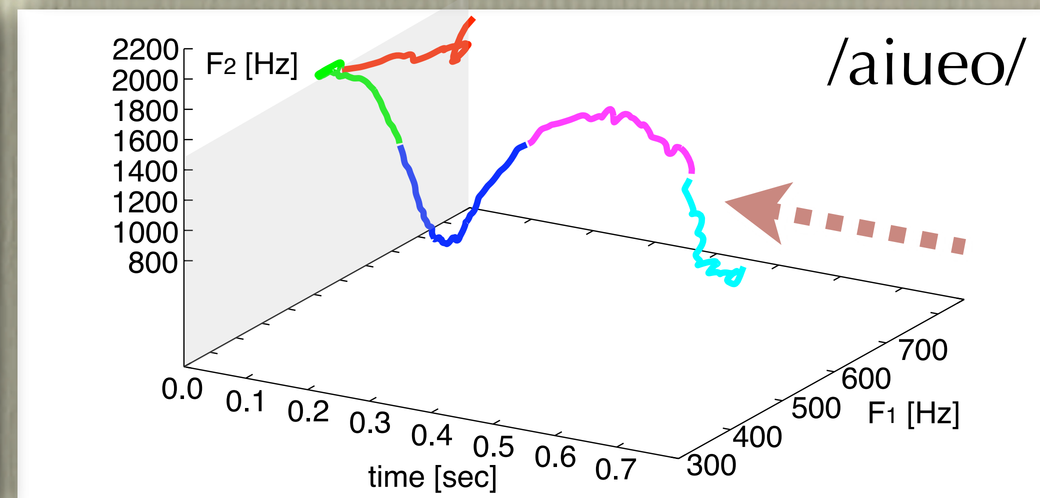


Relative pitch vs. relative timbre



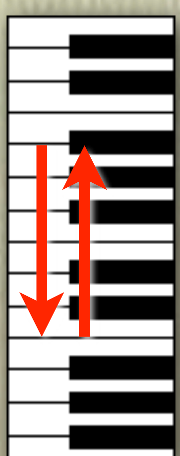




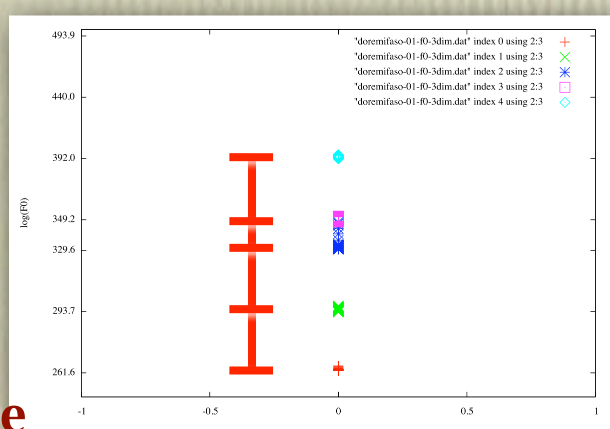


pitch modulation

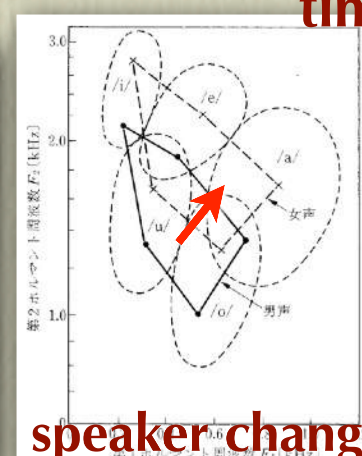
timbre modulation



log(F_0)

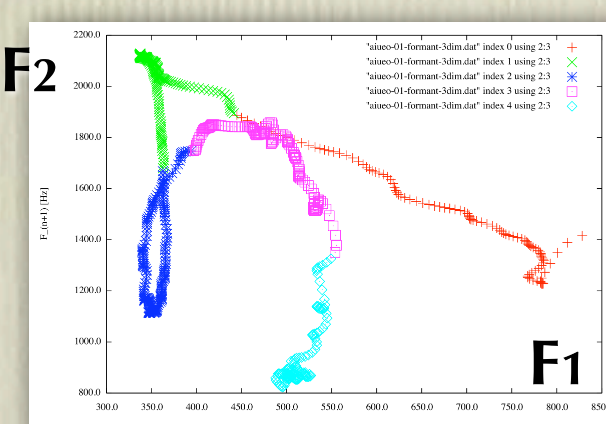


key change



speaker change

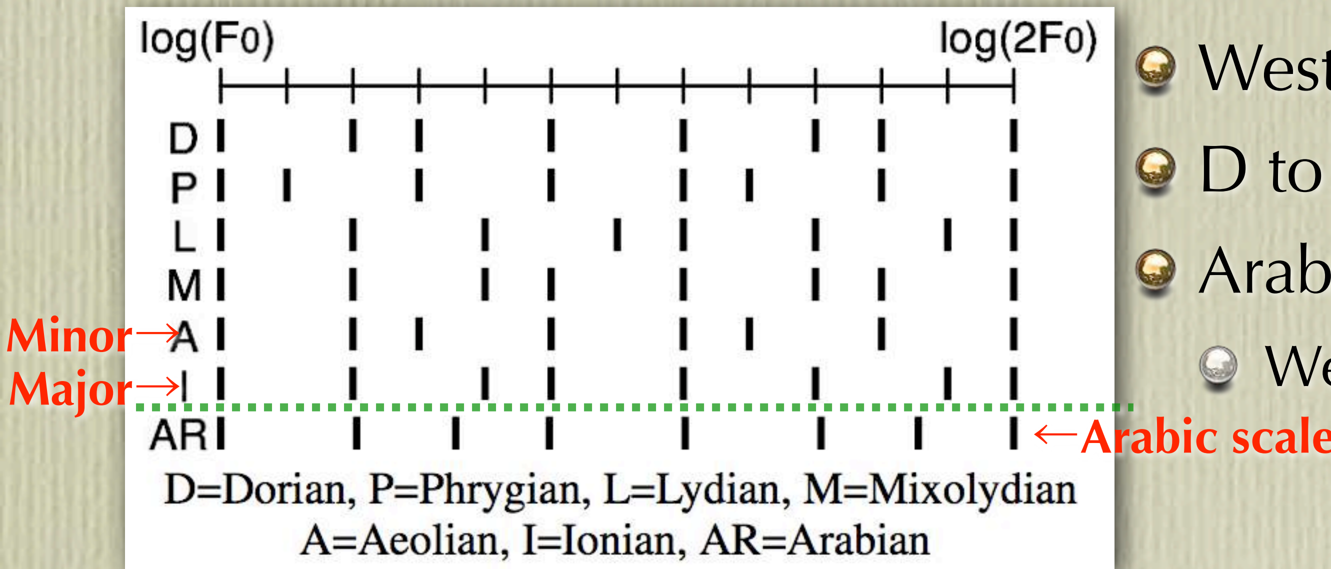
F_2



F_1

Relative pitch vs. relative timbre

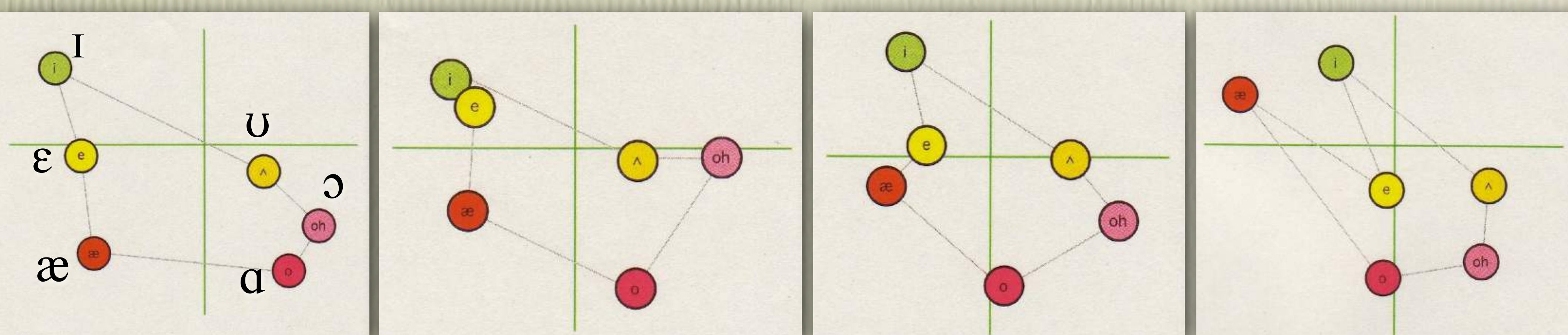
Key-invariant arrangement of tones and its variants



D=Dorian, P=Phrygian, L=Lydian, M=Mixolydian
A=Aeolian, I=Ionian, AR=Arabian

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

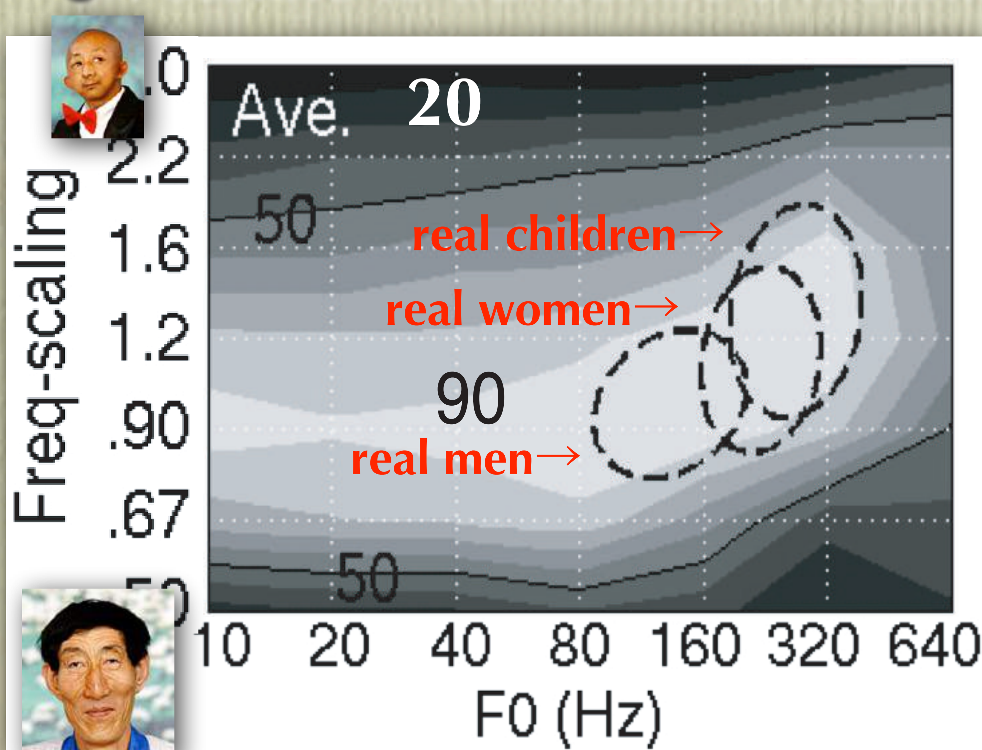
Spk-invariant arrangement of vowels and its variants

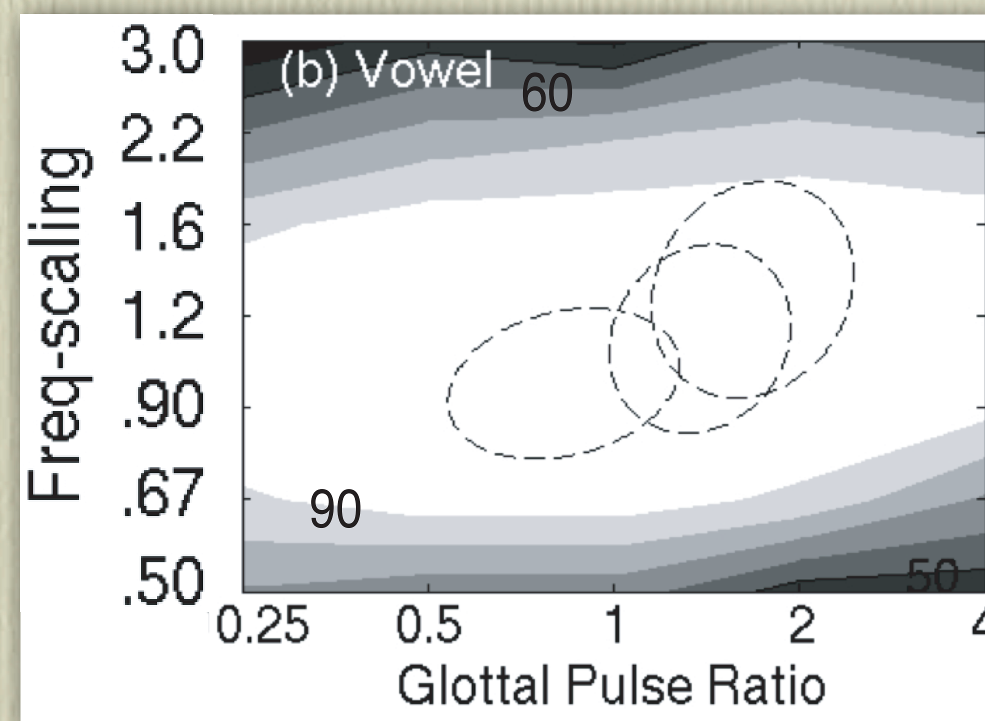


Williamsport, PA **Chicago, IL** **Ann Arbor, MI** **Rochester, NY**

What's difficult 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]





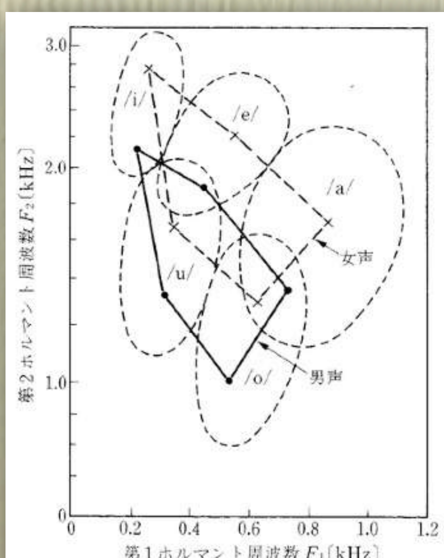
Another difficult task for RP listeners

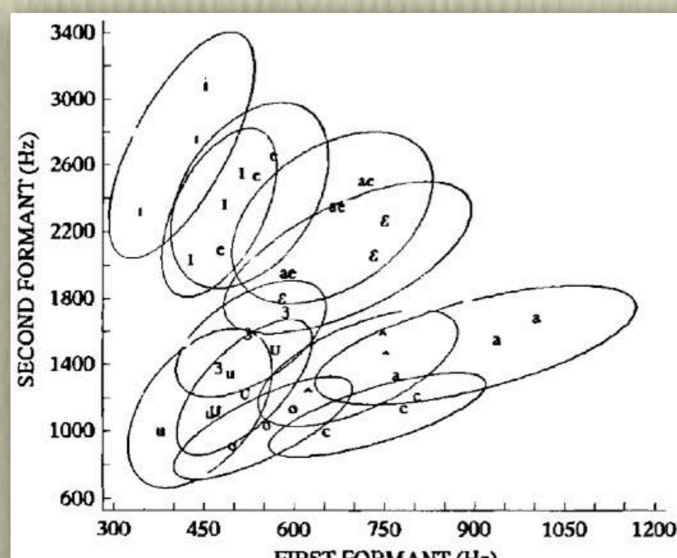
Difficult task for those who can**not** transcribe a melody

- 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 symbolic labeling is difficult, this task is very difficult.

Difficult 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 symbolic labeling is difficult, this task is very difficult.





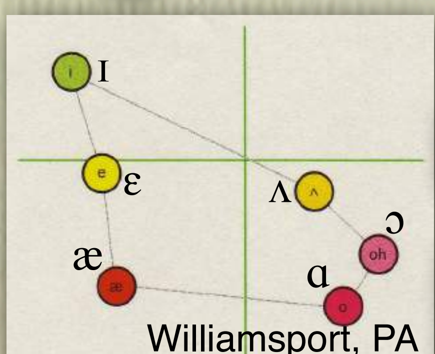
In E-speaking countries, there have to be people who have severe troubles in reading and writing?

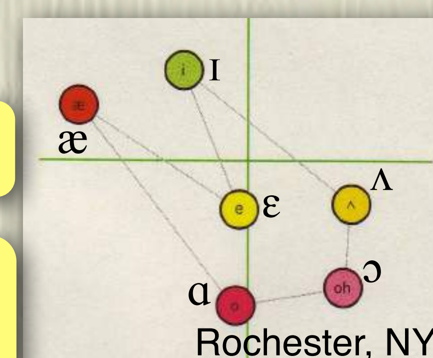
“Separately brought up identical twins”

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



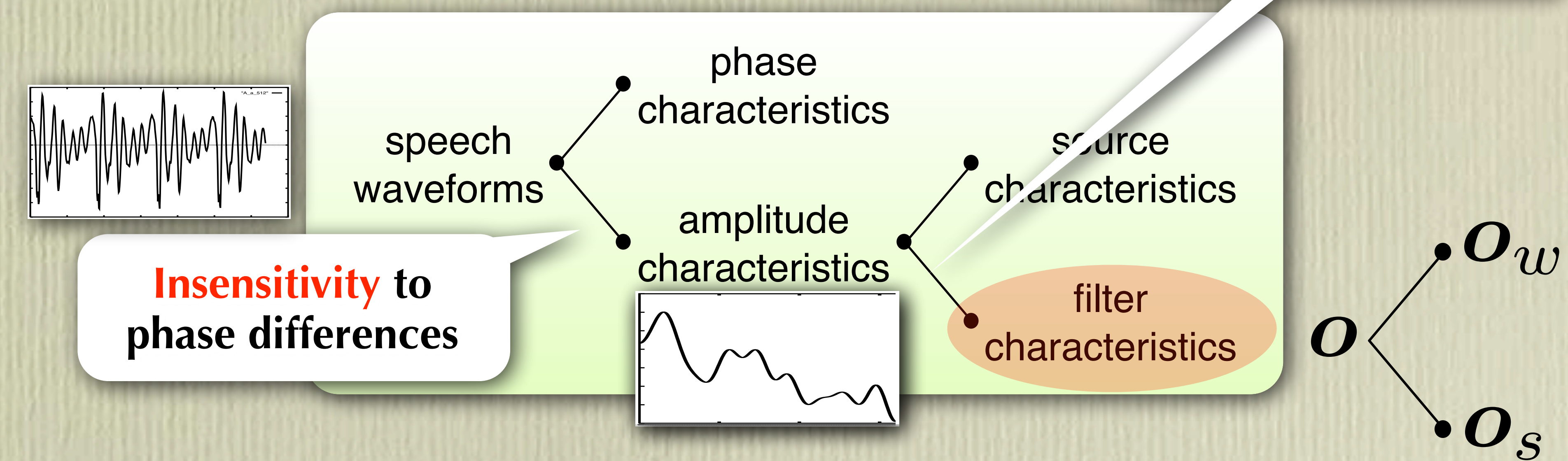


Diff. of regional accents = Diff. of timbre

The machines that don’t learn what infants don’t learn.

Feature separation to find specific info.

De facto standard acoustic analysis of s



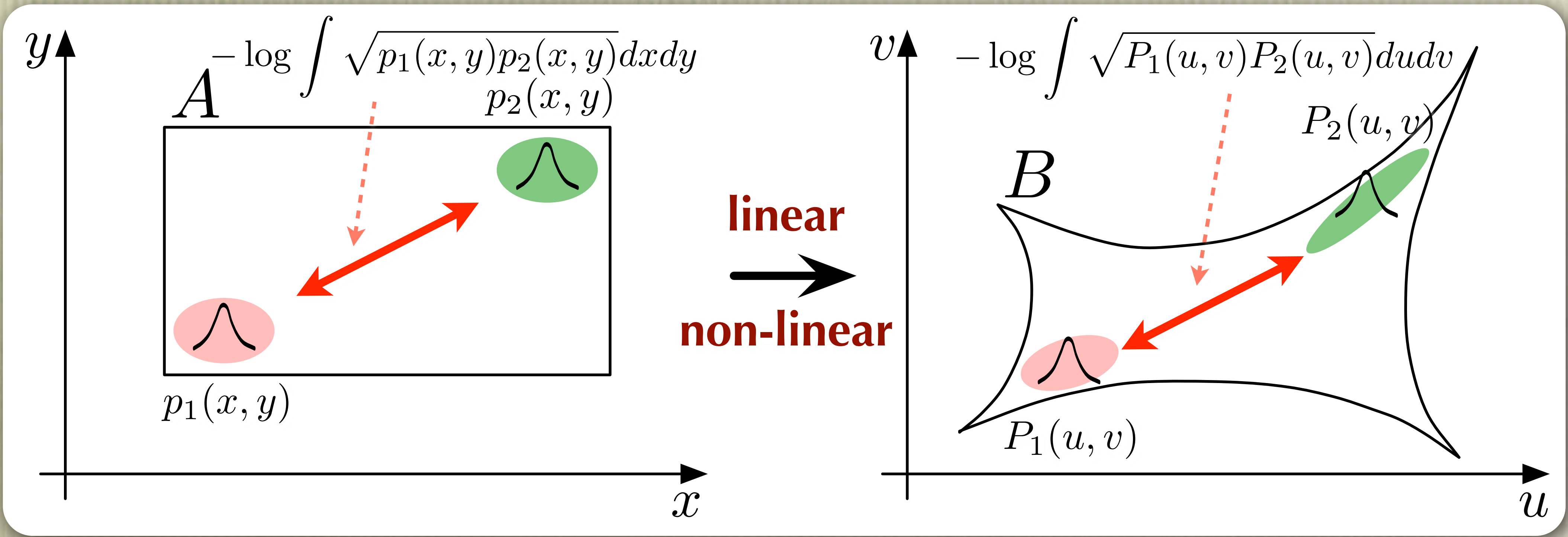
Two acoustic models for speech/speaker recognition

- 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?

Complete transform-invariance

Complete invariance between two spaces

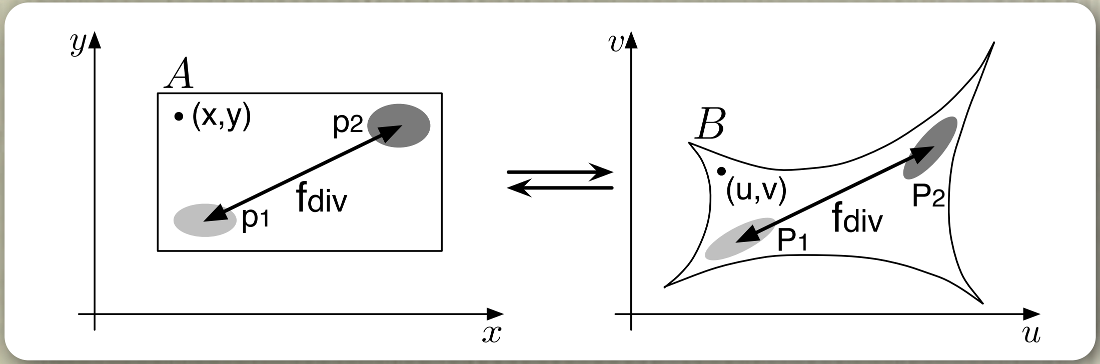
- An assumption
  - The transform is convertible and differentiable anywhere.
- An event in a space should be represented as distribution.
  - Event p in space A is transformed into event P in space B
  - p and P are physically different (/a/ of speaker A and /a/ of speaker B)



Complete transform-invariance

Any general expression for invariance? [Qiao'10]

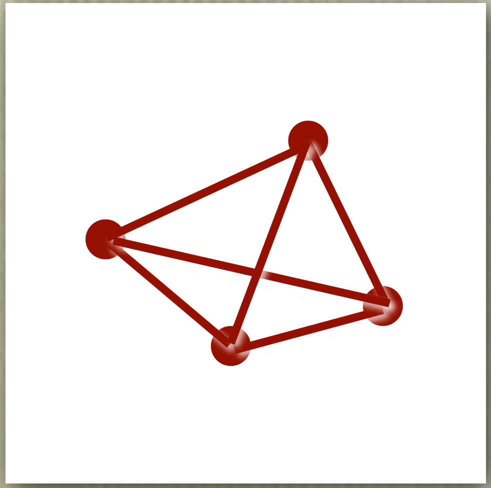
- BD is just one example of invariant contrasts.
- f-divergence is invariant with any kind of transformation.
  - $f_{div}(p_1, p_2) = \int p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}$
  - $g(t) = t\log(t) \to f_{div} = \mathrm{KL} - \mathrm{div}.$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

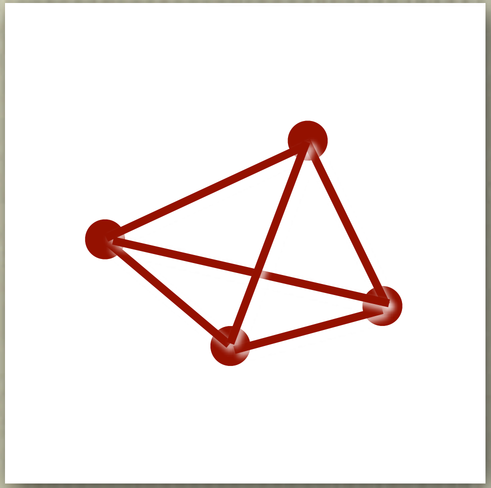


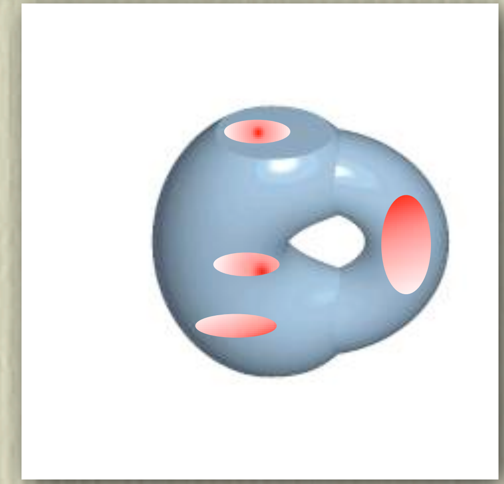
Invariance in variability

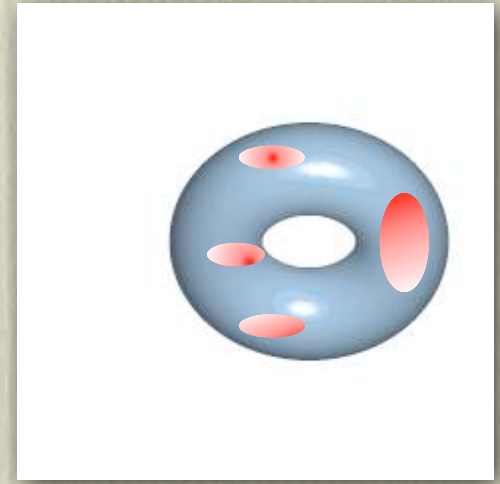
Topological invariance [Minematsu'09]

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



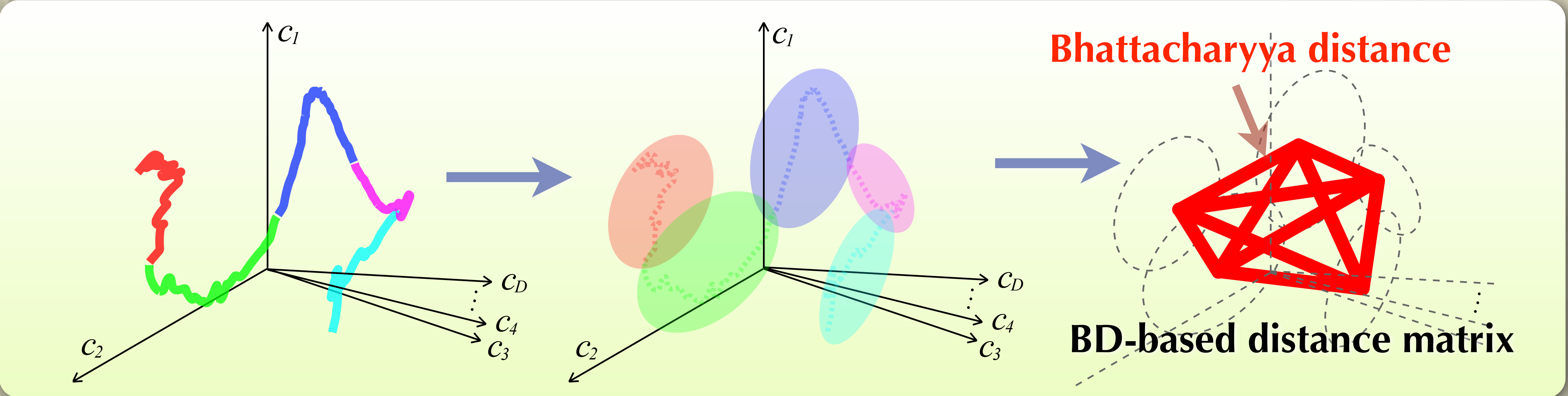


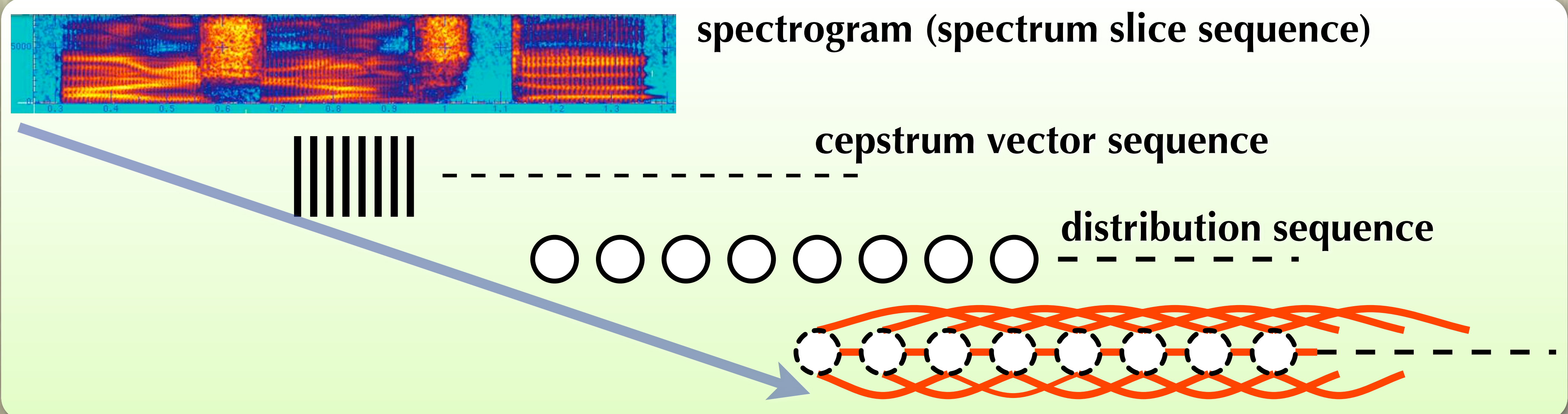




Invariant speech structure

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



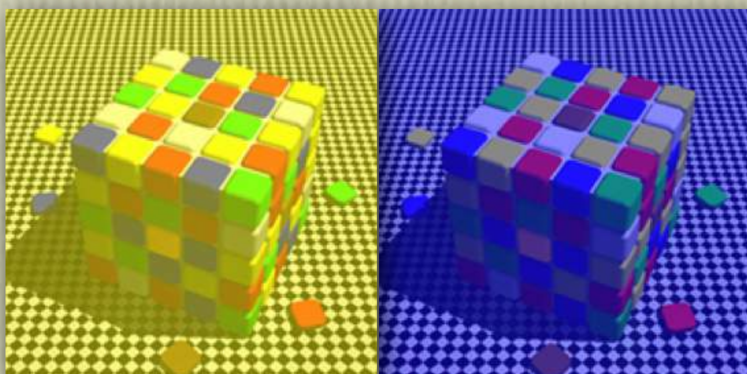
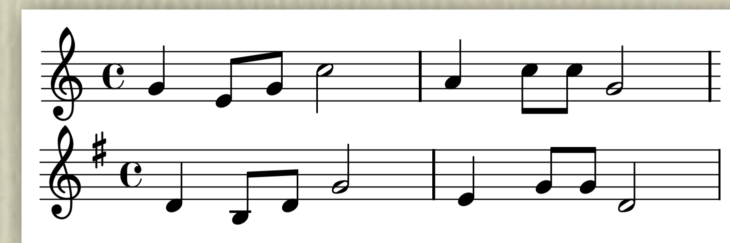


- An event (distribution) has to be much smaller than a phoneme.

Invariant timbre perception against its bias

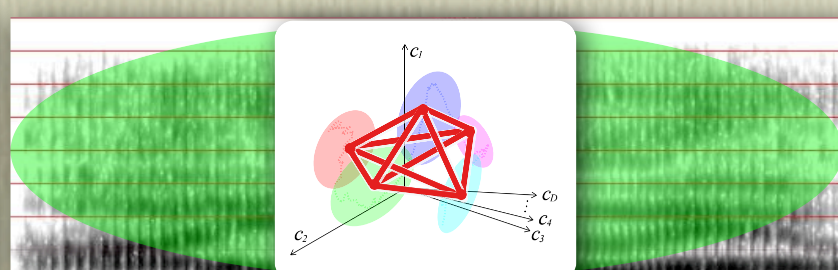
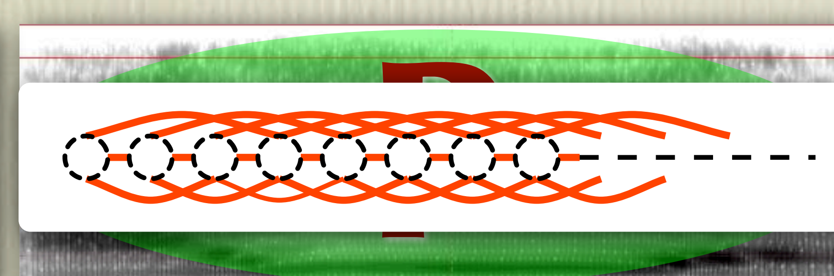
Invariant and constant perception wrt. color and pitch

- 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.

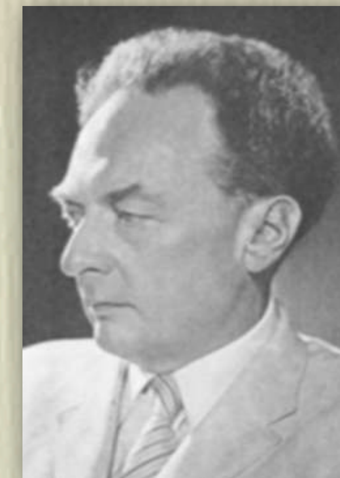
A claim found in classical linguistics

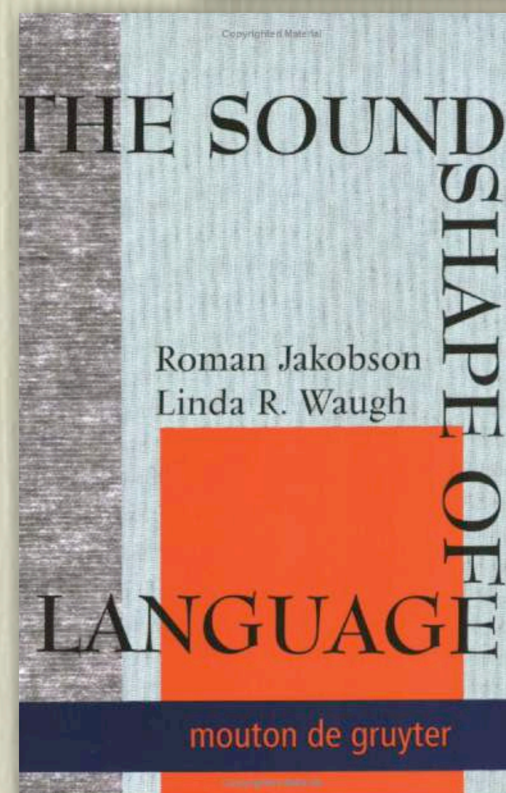
Theory of relational invariance [Jakobson+'79]

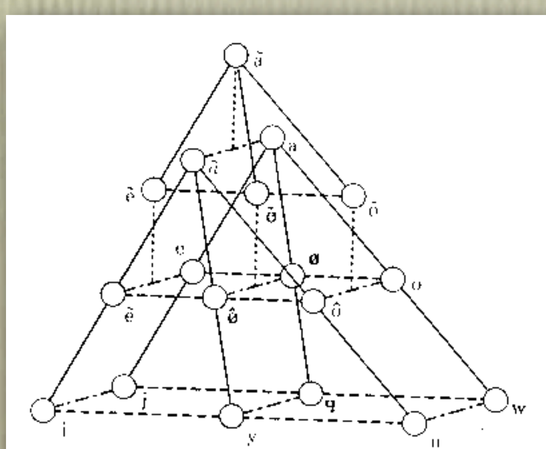
- Also known as theory of distinctive feature
- Proposed by R. Jakobson

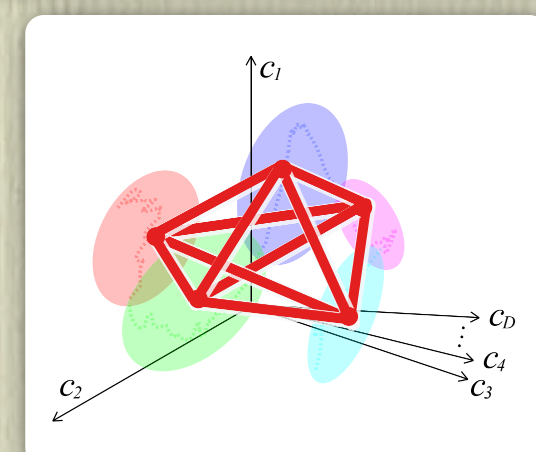
We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables.

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







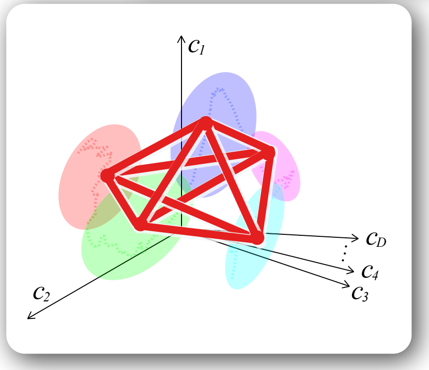


A new framework for "human-like" speech machines #3

Nobuaki Minematsu

Title of each lecture

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

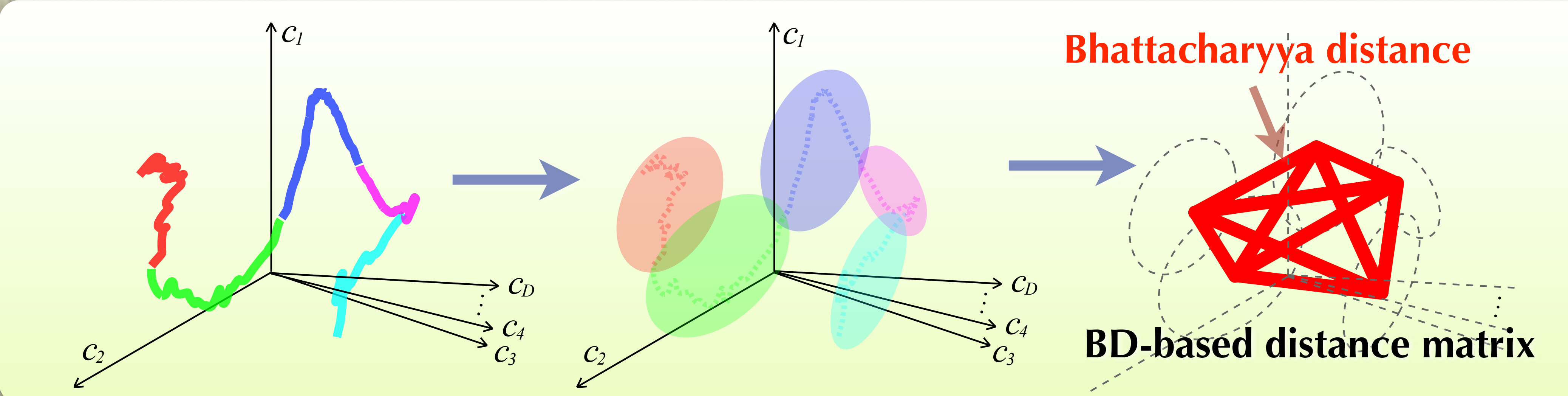


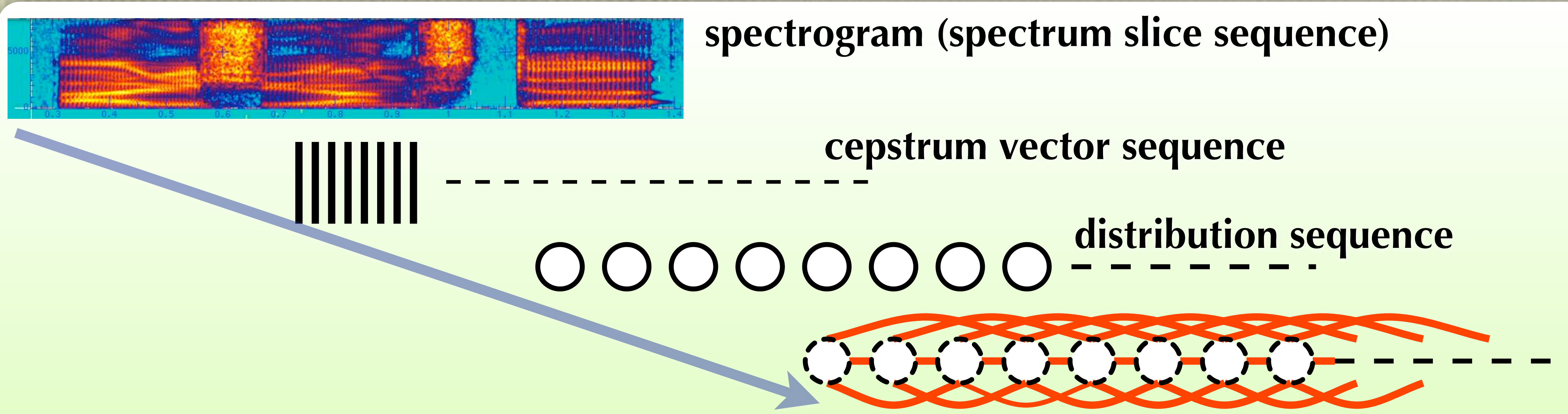
Menu of the last four lectures

- **Robust processing of easily changeable stimuli**
  - Robust processing of general sensory stimuli
  - Any difference in the processing between humans and animals?
- **Human development of spoken language**
  - Infants' vocal imitation of their parents' utterances
  - What acoustic aspect of the parents' voices do they imitate?
- **Speaker-invariant holistic pattern in an utterance**
  - Completely transform-invariant features -- *f*-divergence --
  - Implementation of word Gestalt as relative timbre perception
  - Application of speech structure to robust speech processing
- **Radical but interesting discussion**
  - A hypothesis on the origin and emergence of language
  - What is the definition of "human-like" robots?

Invariant speech structure

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

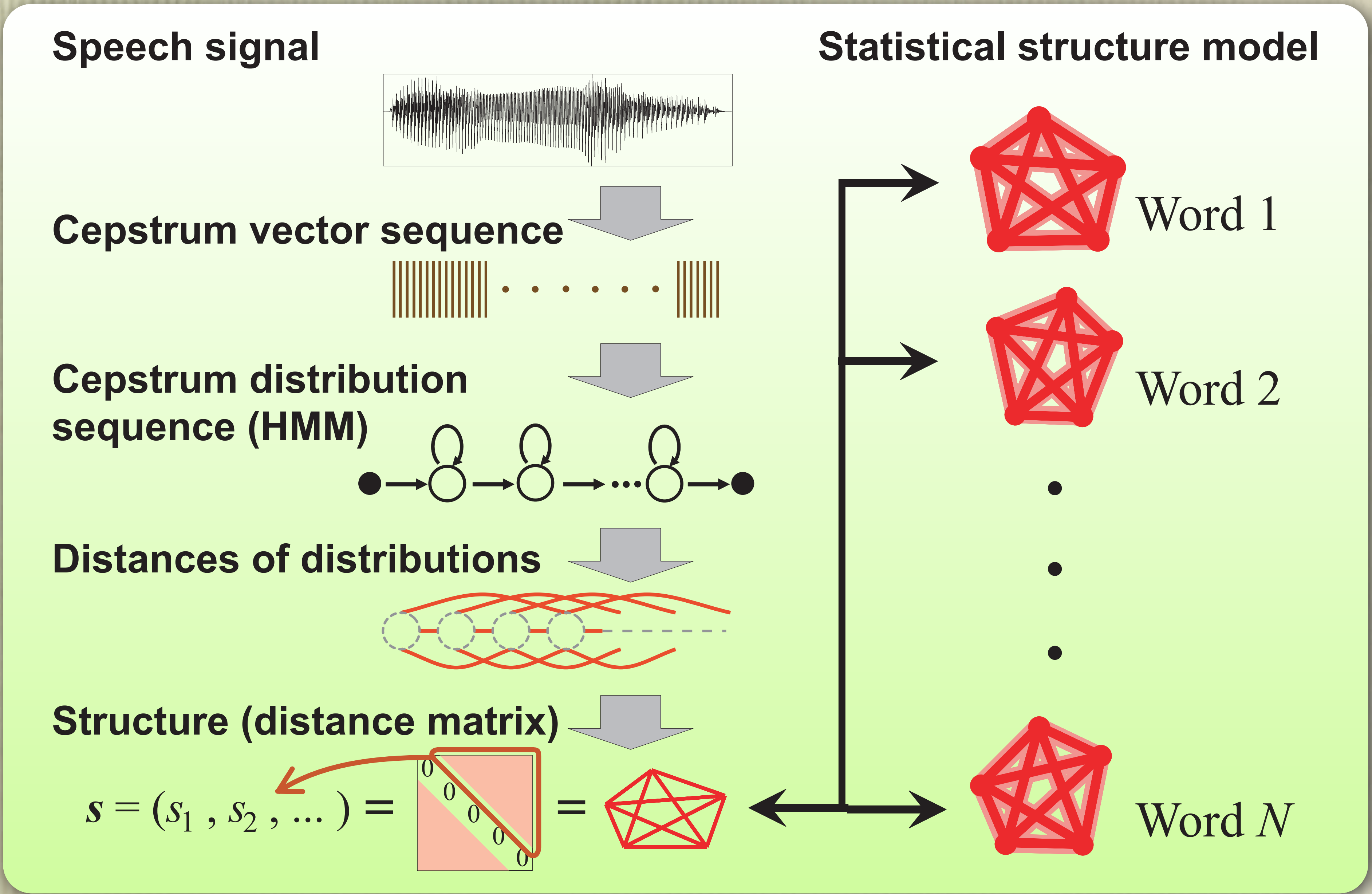




- An event (distribution) has to be much smaller than a phoneme.

Application of structures to ASR

A simple framework for isolated word recognition



Application of structures to ASR

Two big problems

- Too strong invariance (two different words can be the same.)
  - Multi-Stream Structuralization to constrain the invariance [Asakawa'08]
- Too high dimension (N events leads to an ${}_N C_2$ dimensional vector.)
  - 2-stage LDA to reduce the dimension effectively [Asakawa'08]

The invariance only wrt. speaker differences

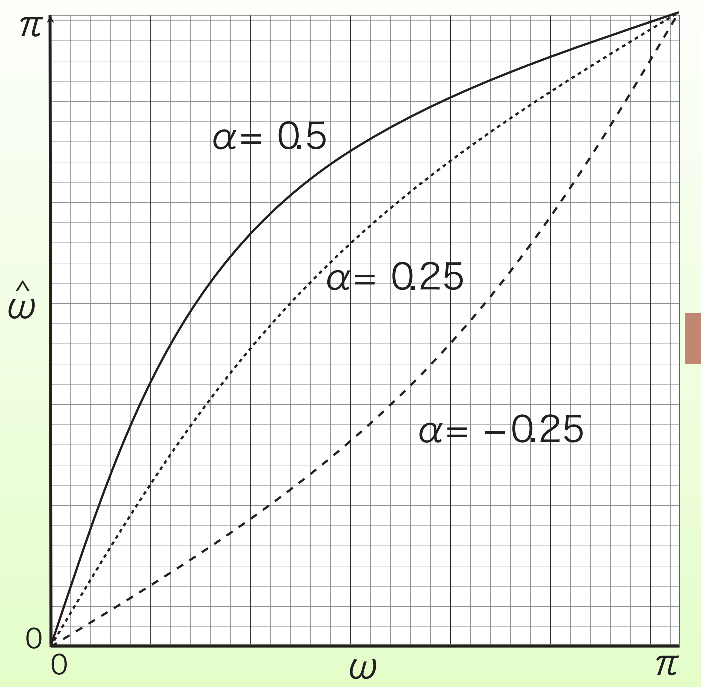
- A mathematical model for VTL differences [Pitz,05]
  - The invariance only wrt. any kind of band matrix ($\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c}$)

$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{c}'_{1,n} \\ \boldsymbol{c}'_{n+1,N} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{c}_{1,n} \\ \boldsymbol{c}_{n+1,N} \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_{1,n} \\ \boldsymbol{b}_{n+1,N} \end{pmatrix}$$

VTLD = × matrix A

- **Vocal tract length difference**
  - Can be approximated as multiplication of matrix A in cep. domain.
- **A is represented with warping parameter α.**



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

$$\hat{\boldsymbol{c}} = (\hat{c}_1 \ \hat{c}_2 \ \hat{c}_3 \ \hat{c}_4 \cdots)^t$$

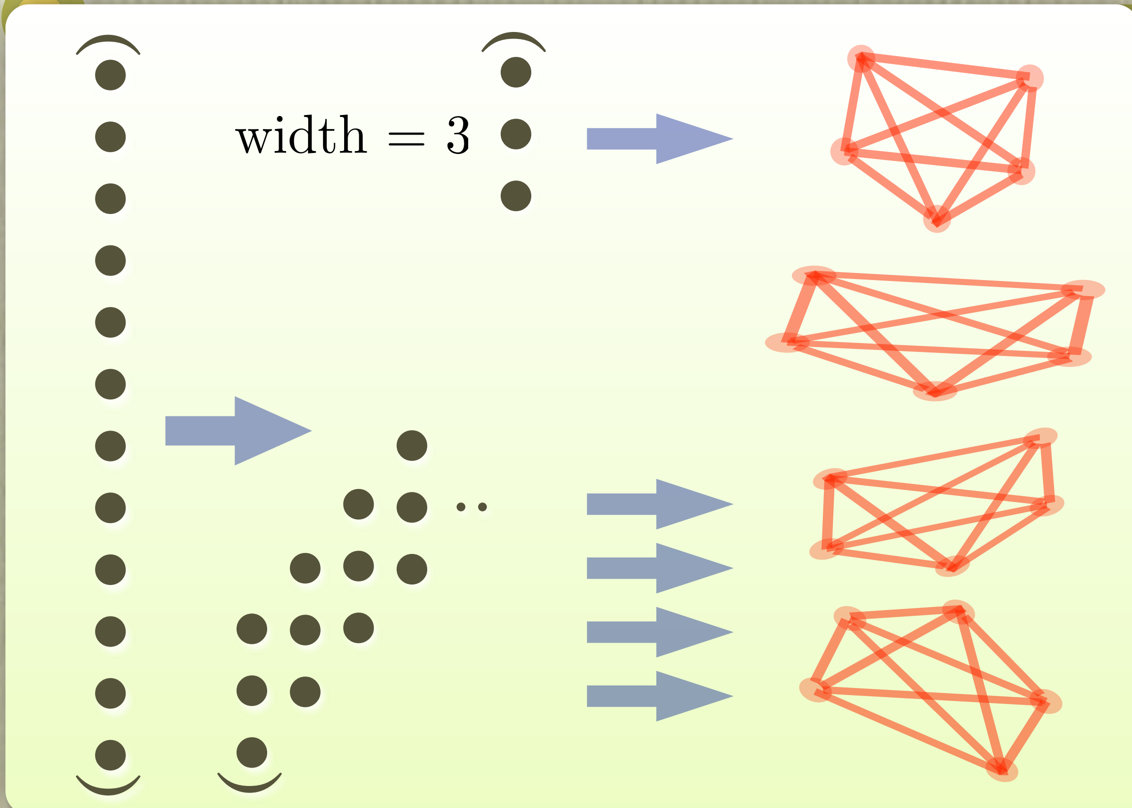
$$\boldsymbol{A} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

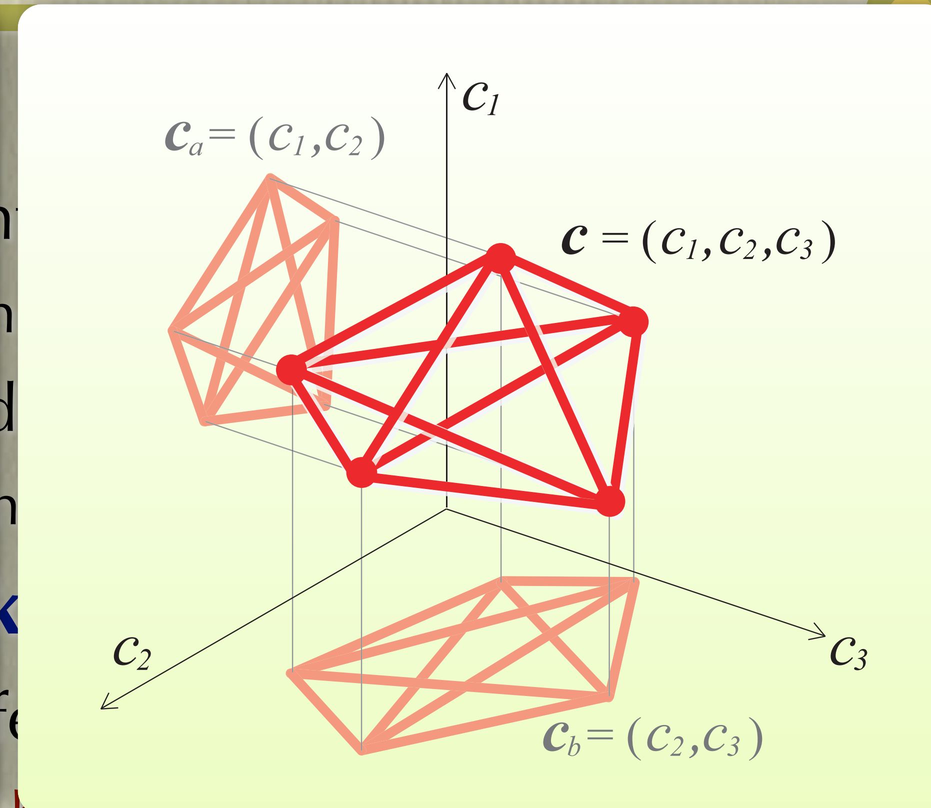
$$\boldsymbol{c} = (c_1 \ c_2 \ c_3 \ c_4 \cdots)^t.$$

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

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

Application of structures to ASR





The invariance only wrt. any kind of band matrix ($c' = Ac$)

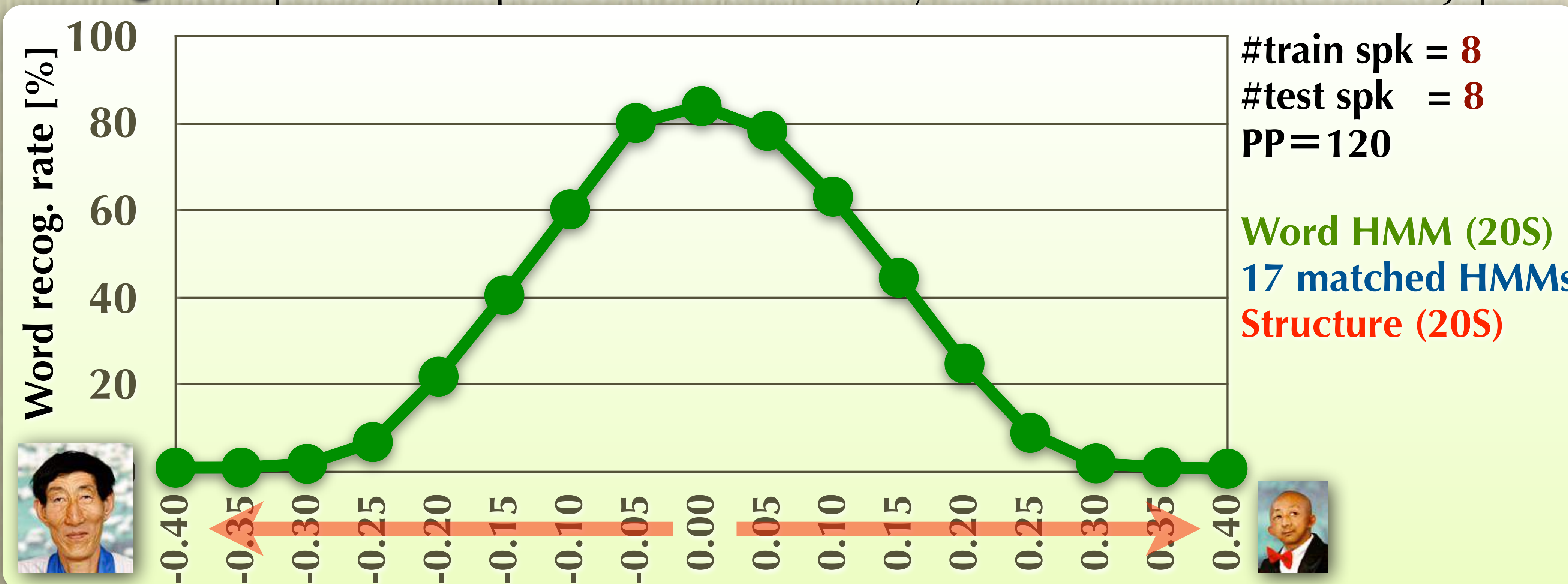
$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{c}'_{1,n} \\ \boldsymbol{c}'_{n+1,N} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{c}_{1,n} \\ \boldsymbol{c}_{n+1,N} \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_{1,n} \\ \boldsymbol{b}_{n+1,N} \end{pmatrix}$$

Application of structures to ASR

Isolated word recognition using warped utterances

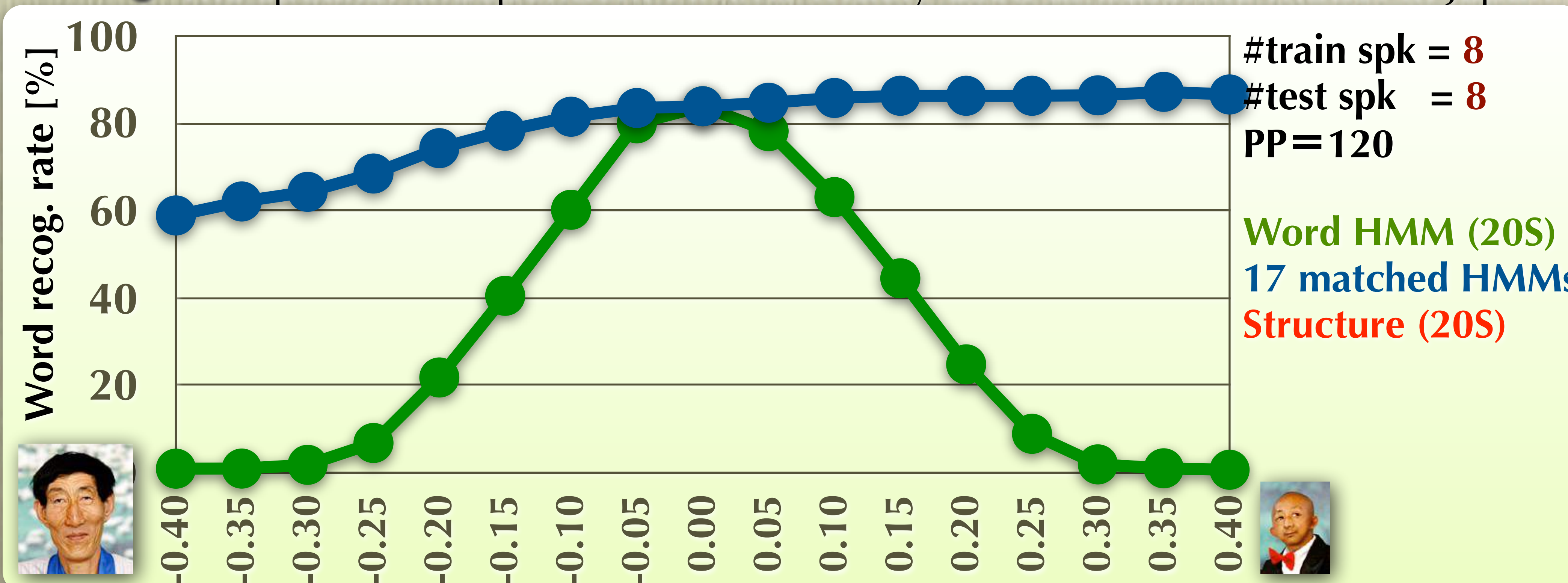
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used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



Application of structures to ASR

Isolated word recognition using warped utterances

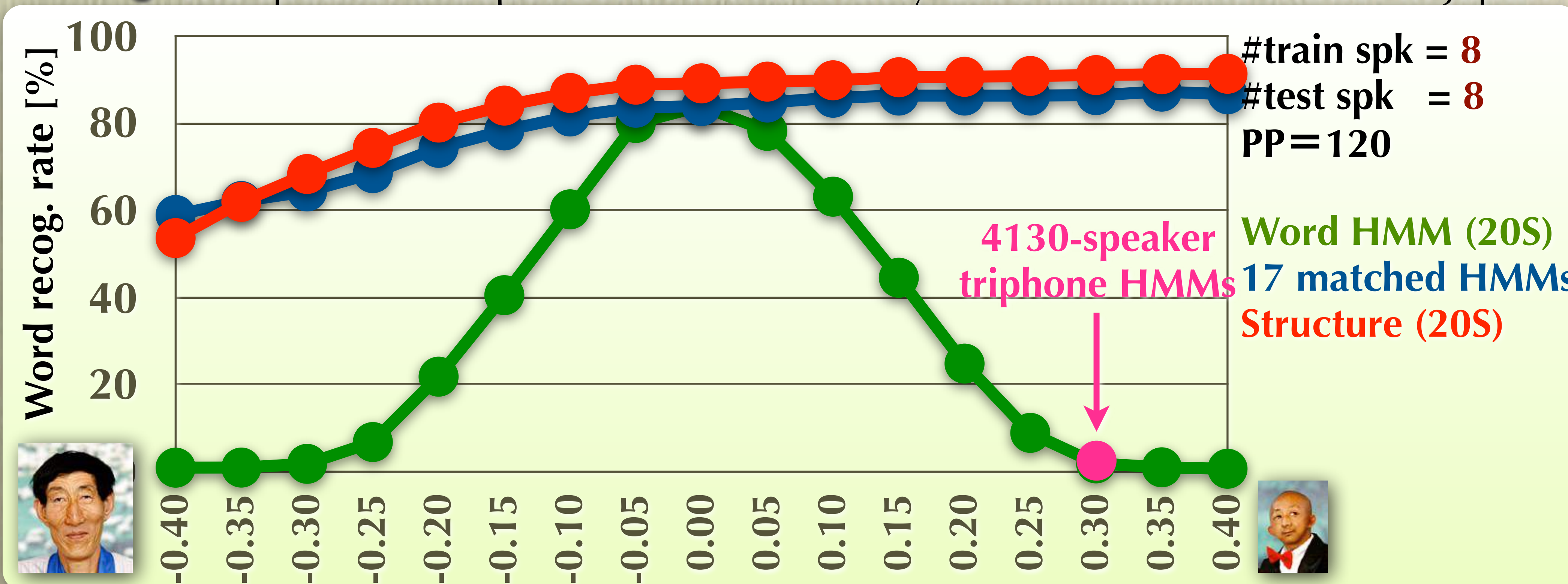
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used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



Application of structures to ASR

Isolated word recognition using warped utterances

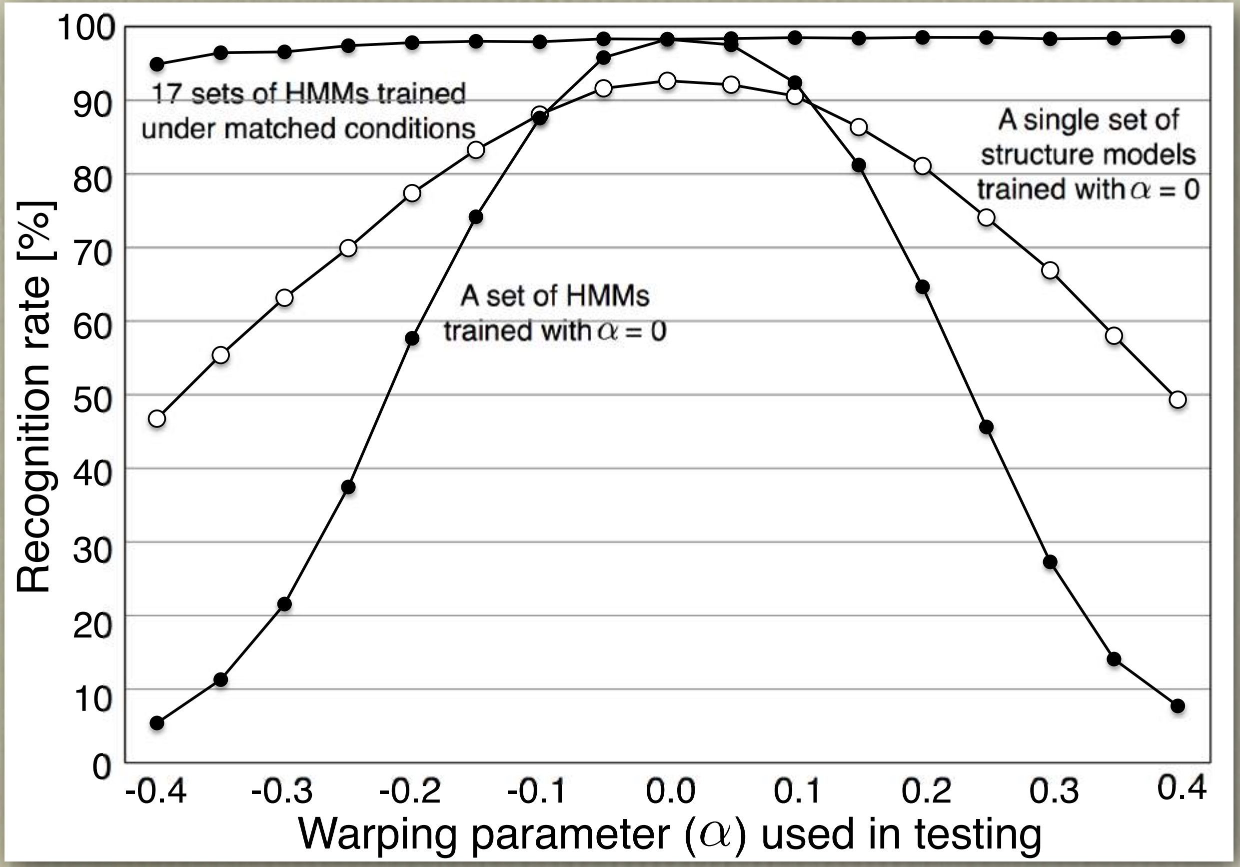
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used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



Application of structures to ASR

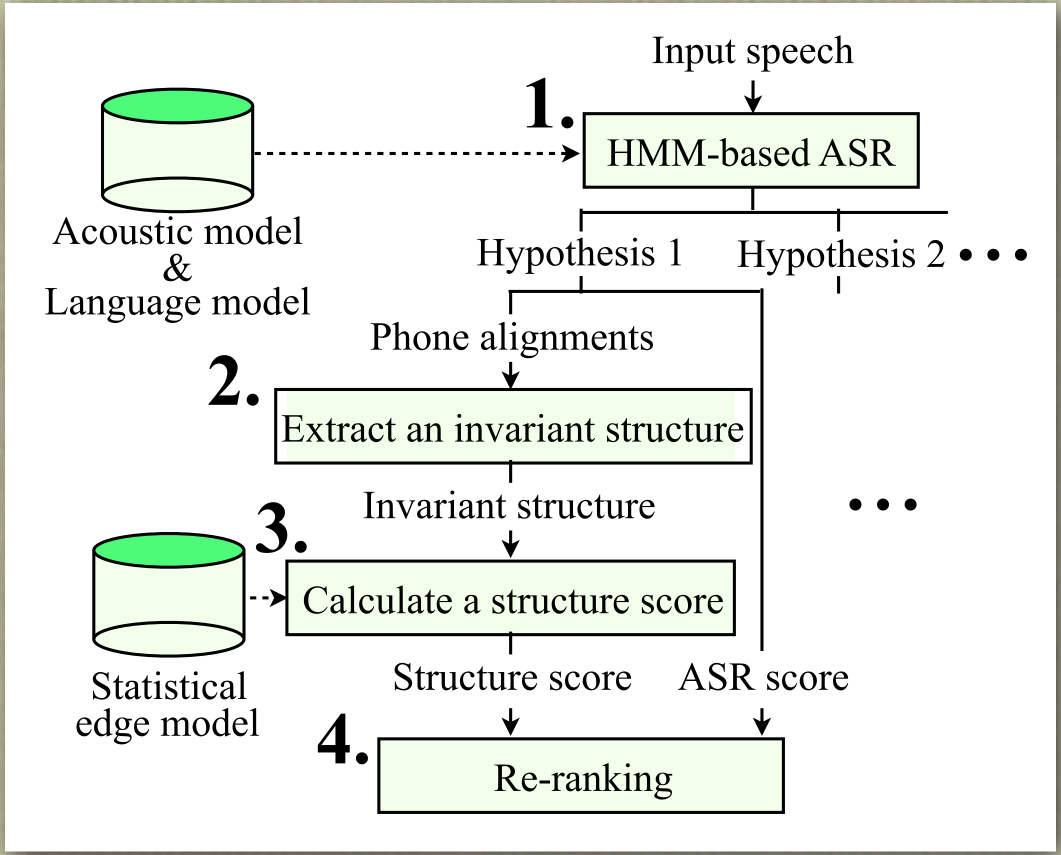
Isolated word recognition using warped utterances

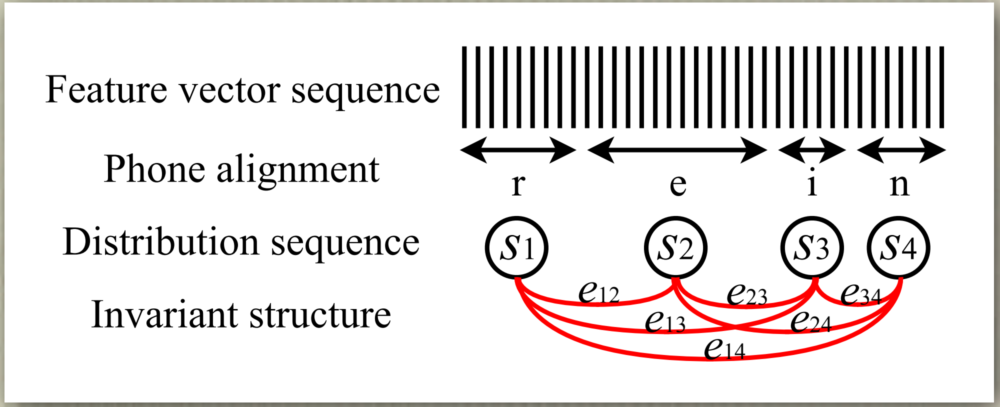
- Word = phoneme-balanced word, PP = 212
  - Mora-based length of words = 3 to 7
- Word-based HMMs (25 states) vs. word-based structures (25 events)
  - Training = 15M+15F adults, testing = other 15M+15F with various VTLs



Application of structures to LVCSR

- **Application to more realistic ASR tasks** [Suzuki+'15]
  - Digits recognition and LVCSR (dictation)
- **Use of structural features in discriminative reranking**
  - Str. scores and ASR scores are combined with average perceptron.

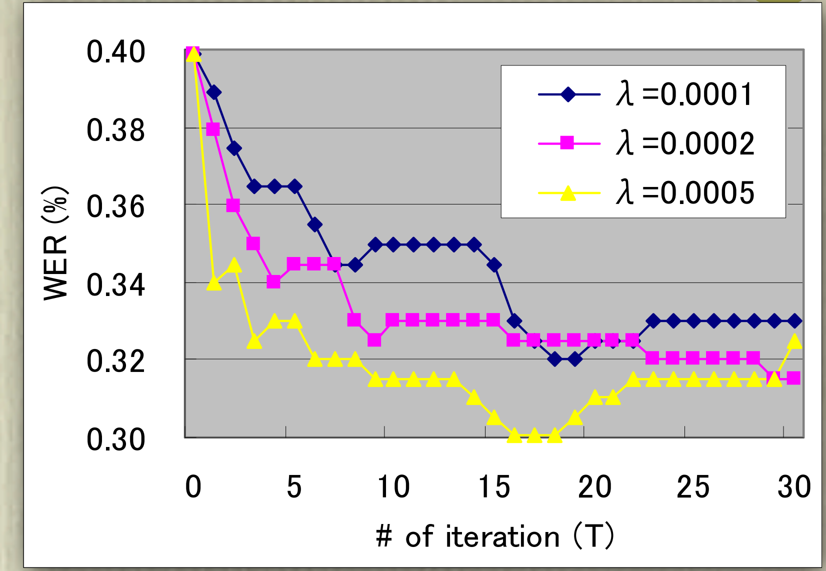




Application of structures to LVCSR

Continuous digits recognition

- Language = Japanese
- Baseline = GMM-HMM ASR
- Reranking = averaged perceptron
- Error reduction rate = 30%



Large vocabulary continuous speech recognition

- Language = Japanese
- Baseline = DNN-HMM ASR
- Reranking = averaged perceptron
- Error reduction rate = 5%

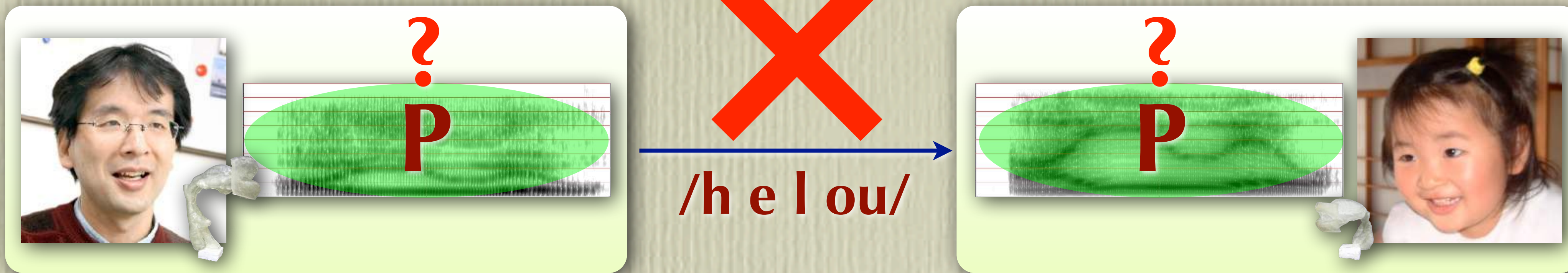
Many errors are due to a large number of homonyms in Japanese.

Table 6: CERs of the LVCSR experiment.

Baseline	Proposed	Relative improvement
2.67%	2.53%	5.24%

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]

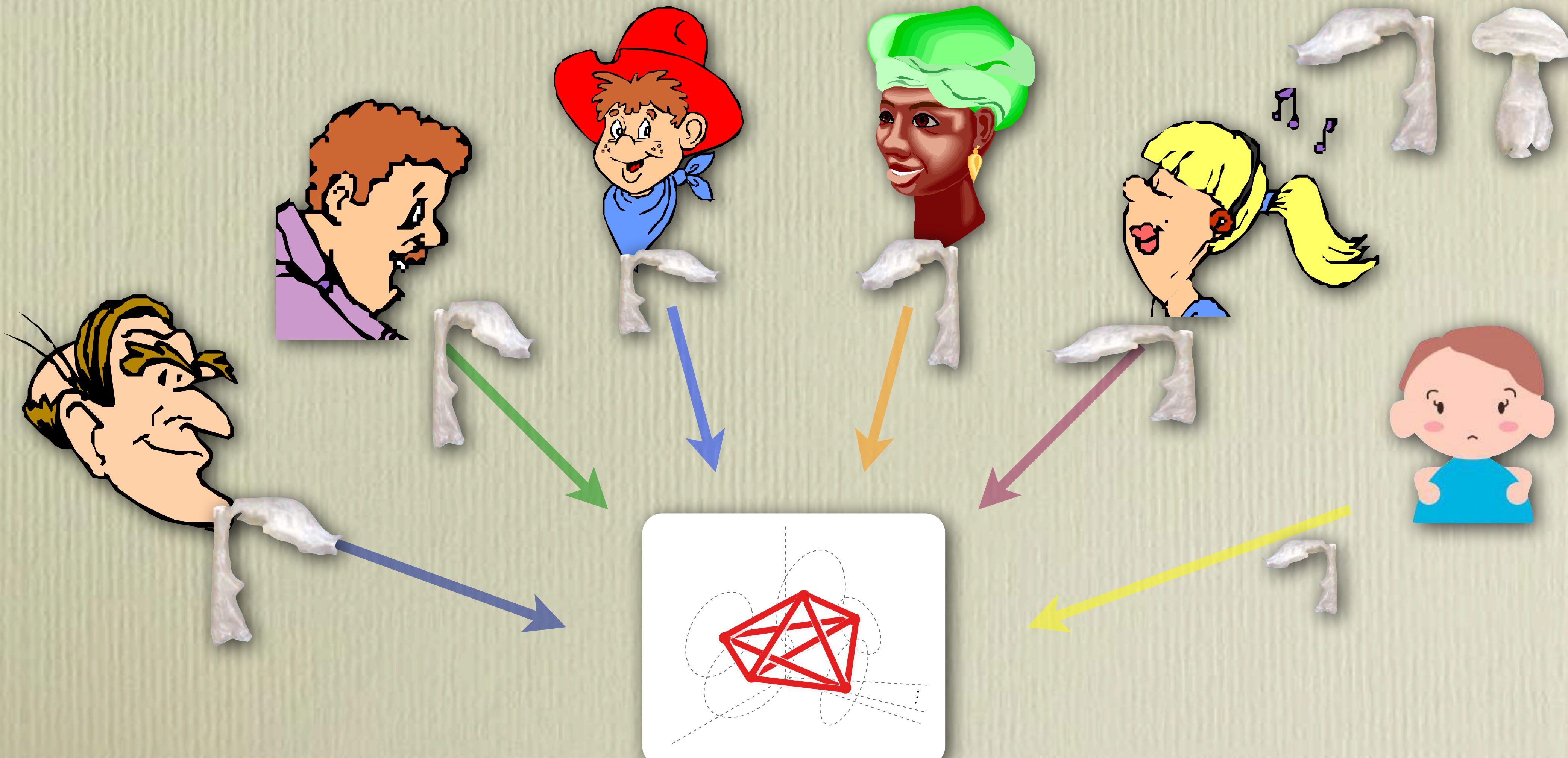
  → **No mathematical formulation**

  - 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?

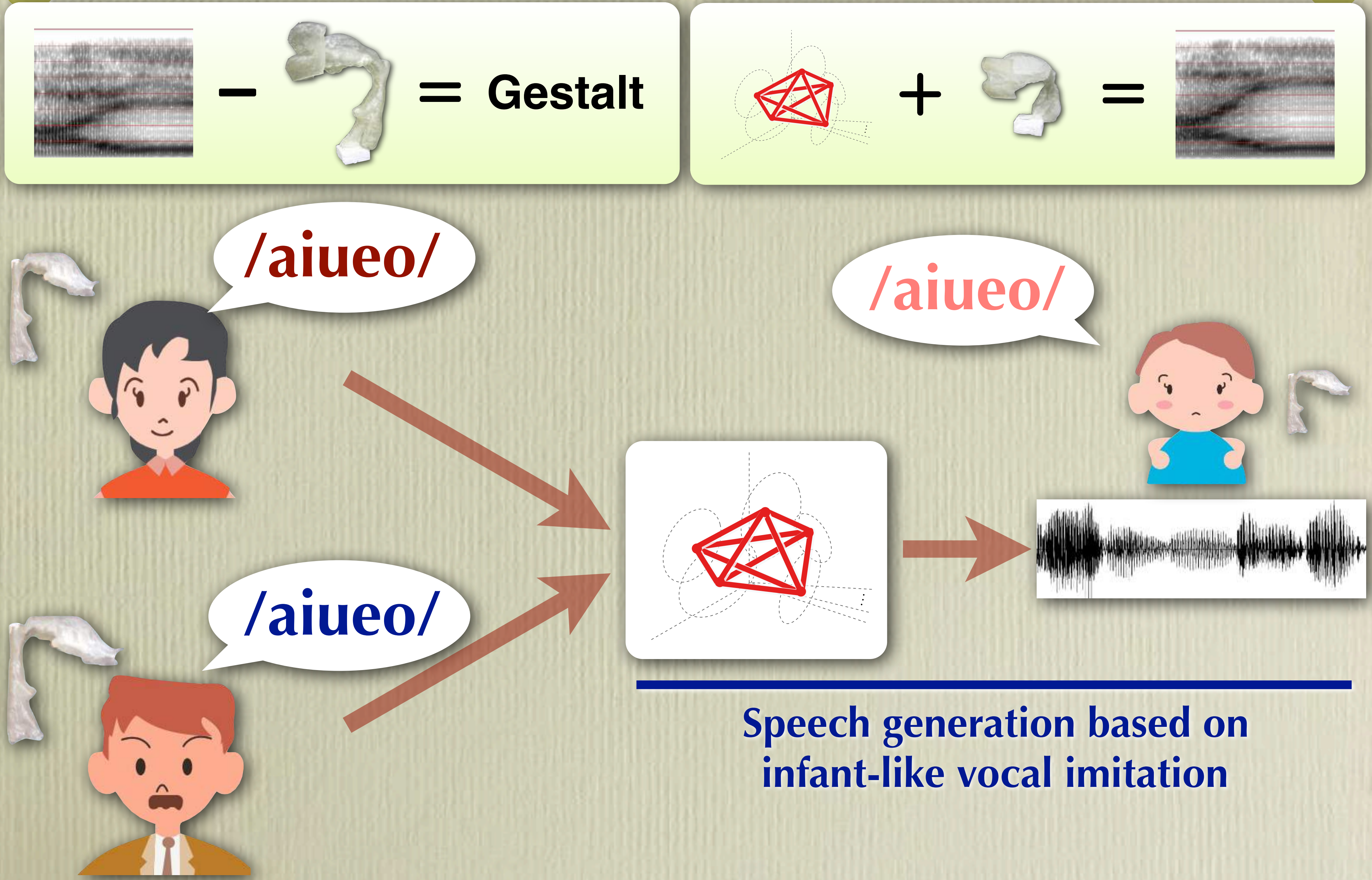
Structure-to-speech conversion

Speech representation with extra-ling. features removed

- Speaker-specific vocal tract features are removed.
- With them, we can identify speakers by hearing voices.



Structure-to-speech conversion

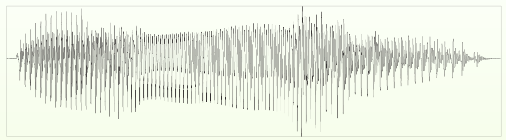


Speech generation based on infant-like vocal imitation

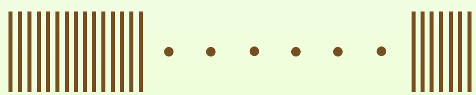
How to implement the vocal imitation?

Extraction of a structure through training of an HMM

1. Speech waveforms

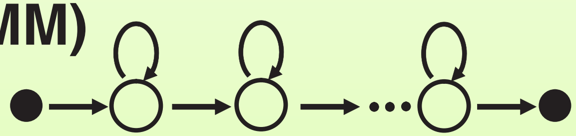


2. Cepstrum vector sequence

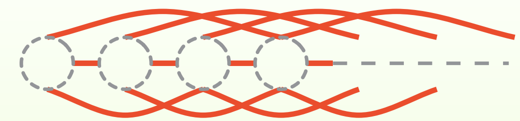


3. Cepstrum distribution sequence (HMM)

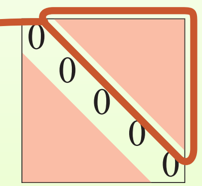
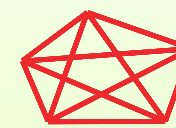
MAP estimation



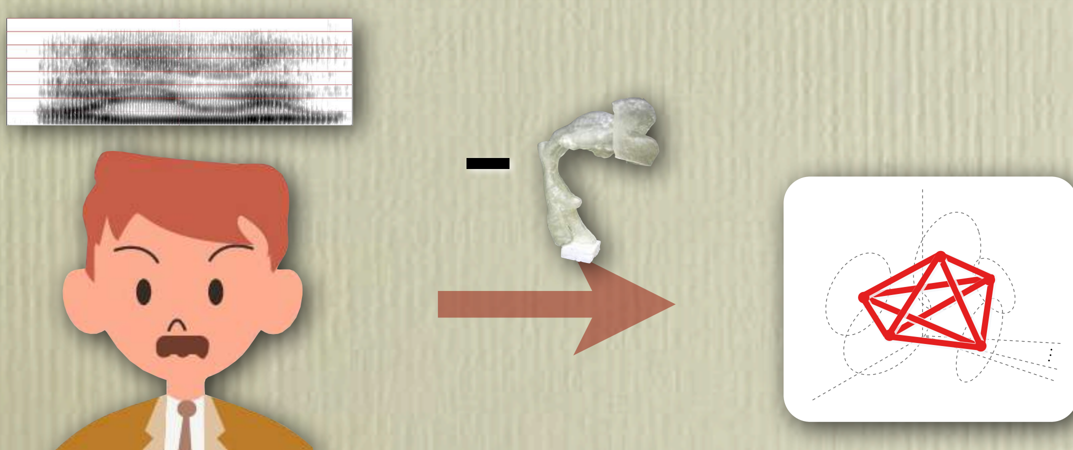
4. Bhattacharyya distances



5. Structure (distance matrix)

$\boldsymbol{s} = (s_1 , s_2 , \ldots) =$  $=$ 

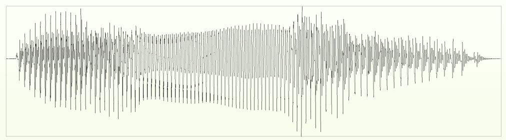
structure vector



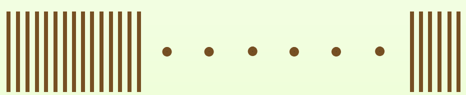
How to implement the vocal imitation?

Extraction of a structure through training of an HMM

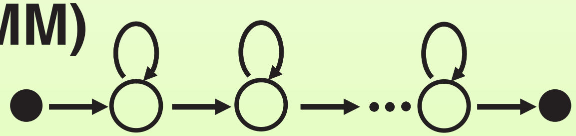
1. Speech waveforms



2. Cepstrum vector sequence

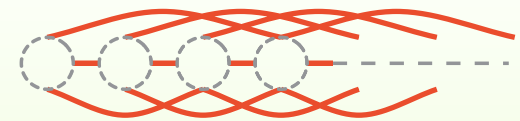


3. Cepstrum distribution sequence (HMM)

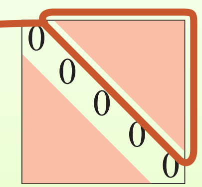
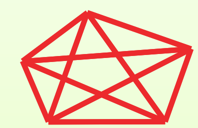


MAP estimation

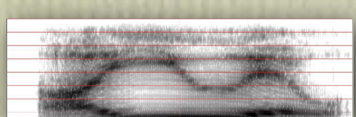
4. Bhattacharyya distances

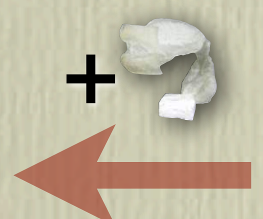


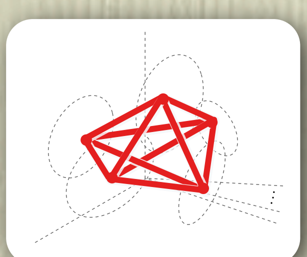
5. Structure (distance matrix)

$\boldsymbol{s} = (s_1 , s_2 , \ldots) =$  $=$ 

structure vector

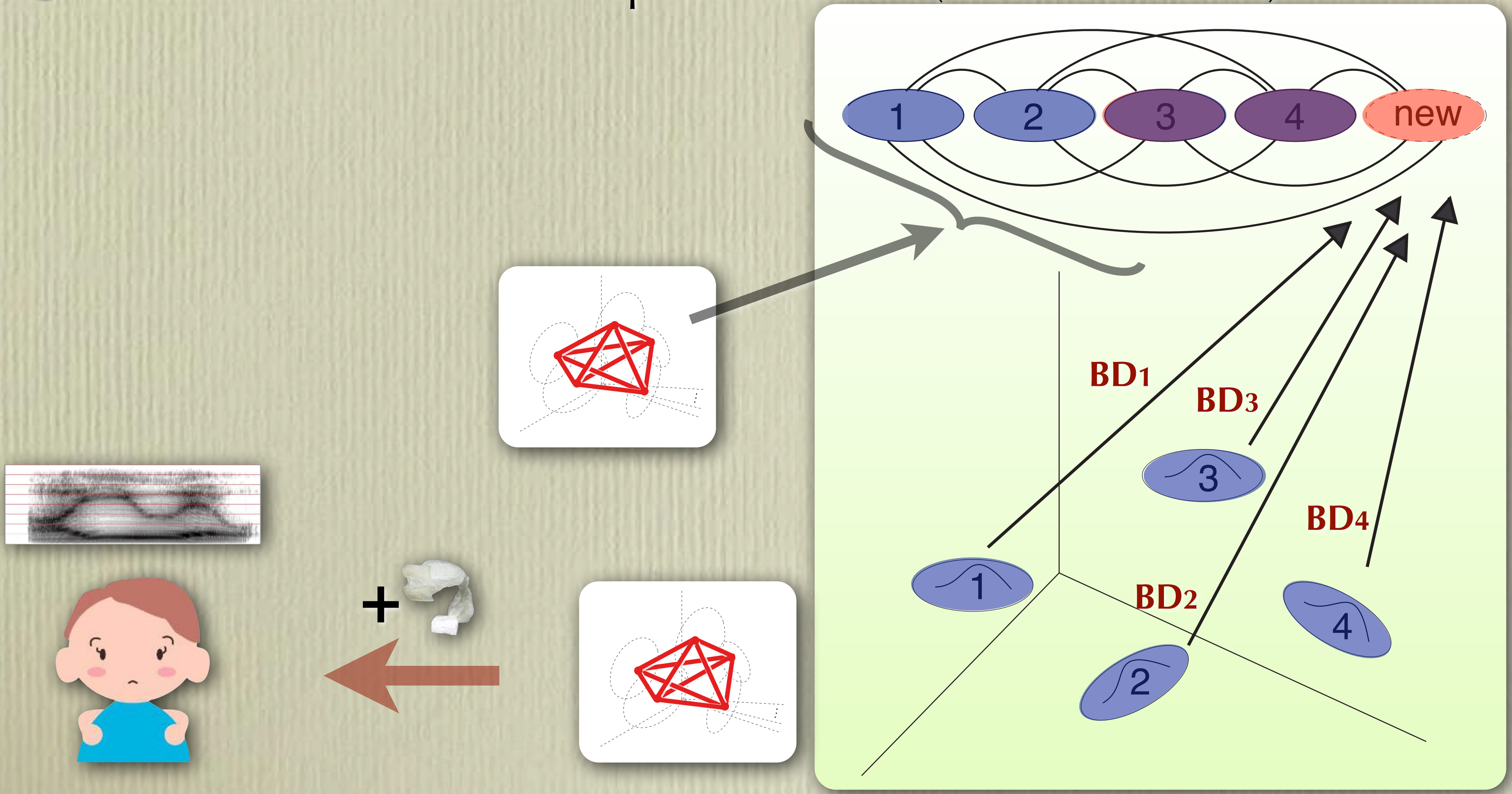






How to implement the vocal imitation?

- **Acoustic instances are searched for in the voice space.**
  - Initial conditions : a few acoustic instances given from an infant
  - Constrained conditions : speech Gestalt (distance matrix)

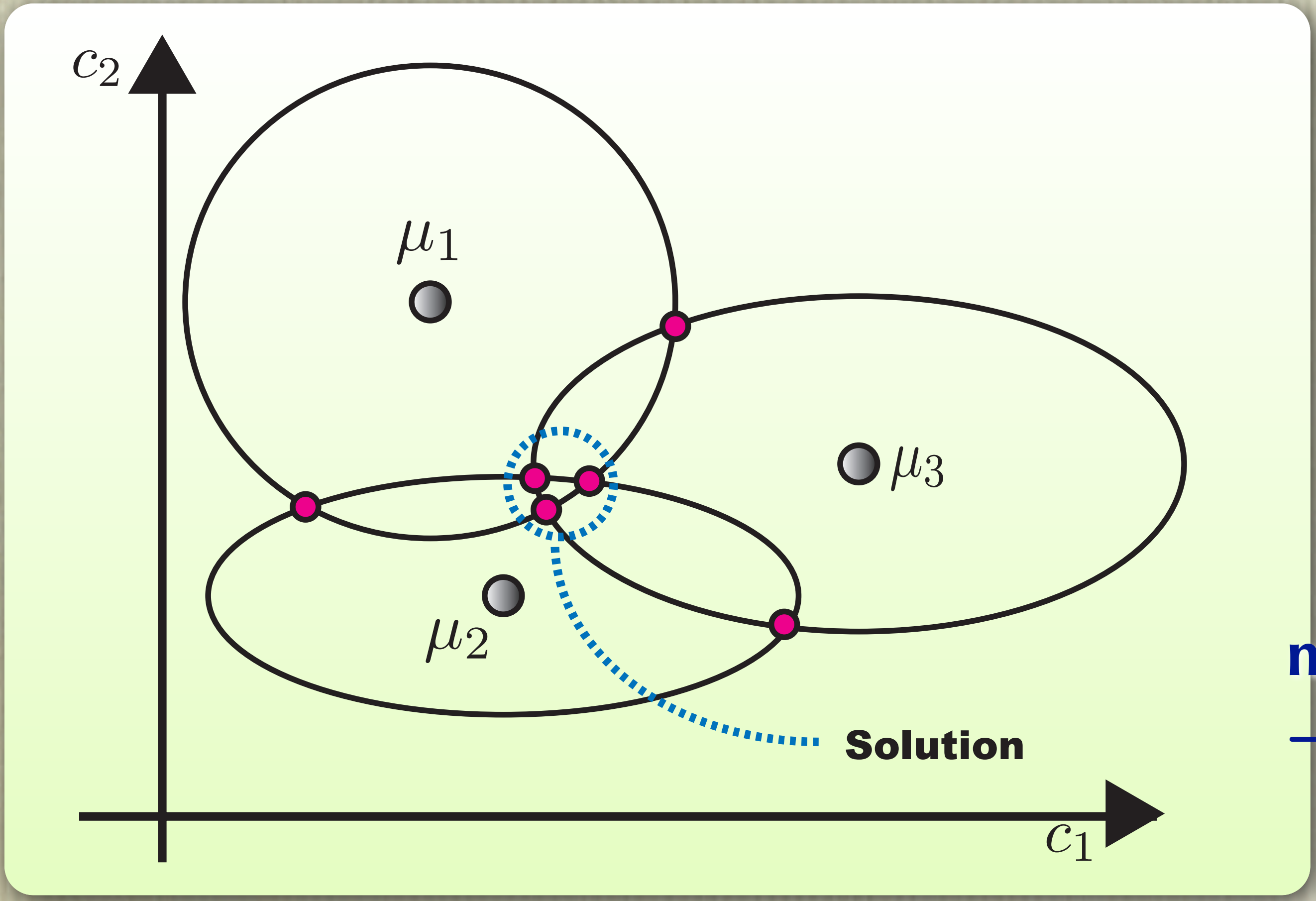


How to implement the vocal imitation?

Geometrical interpretation of BD-based constraints

$$BD(p_1(x), p_2(x)) = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma_{12}^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\frac{|\Sigma_{12}|}{|\Sigma_1||\Sigma_2|}$$

- Search for a new target using BD(1,new), BD(2,new), BD(3,new)...
  - Σ_{new} is given. Only μ_{new} is searched for in the current paper.

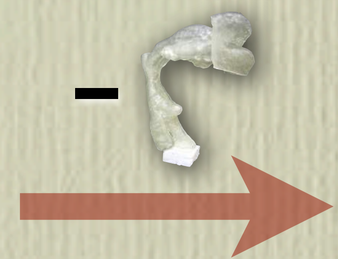


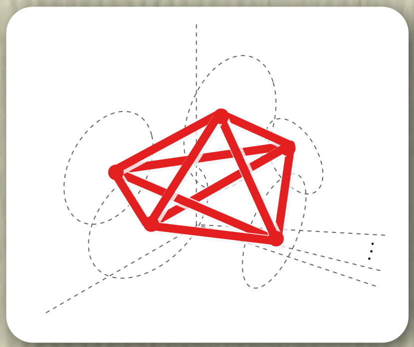
multiple solutions $\longrightarrow$ averaging

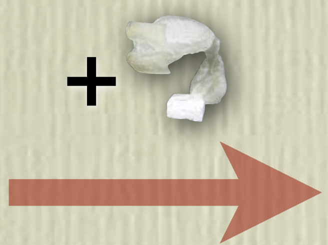
An experiment with real vocal imitation

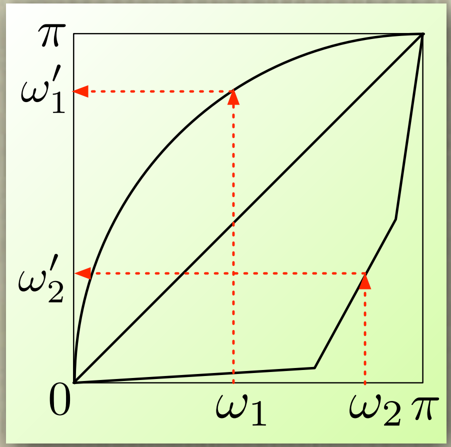
Demonstration with my wife and daughter

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





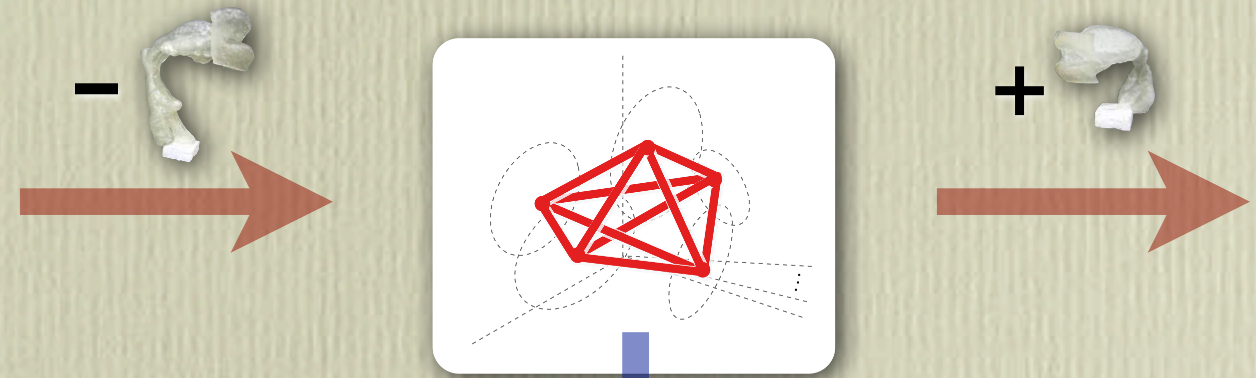


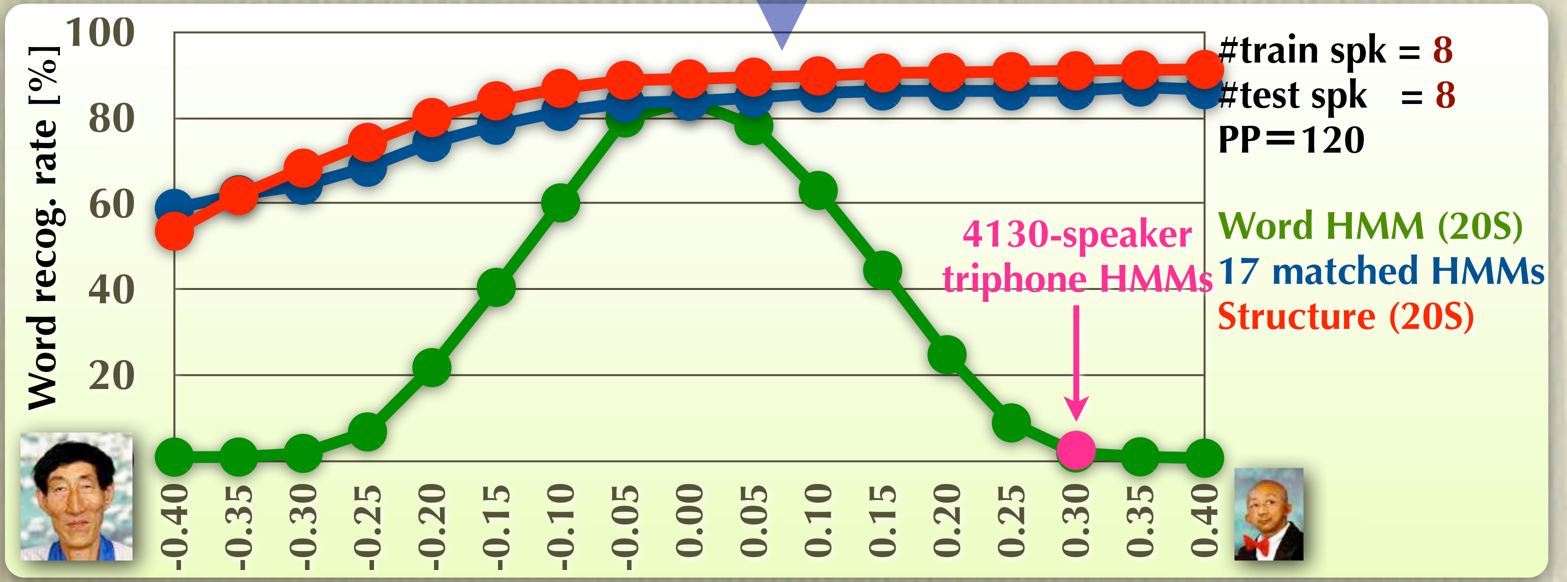


An experiment with real vocal imitation

Demonstration with my wife and daughter

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

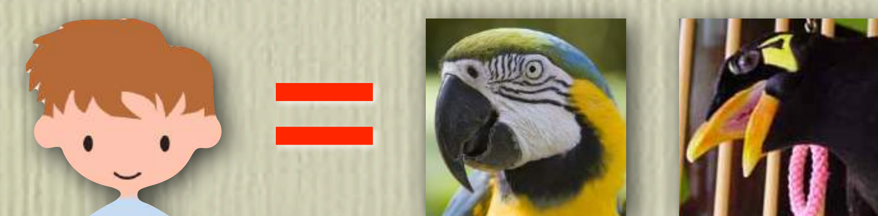




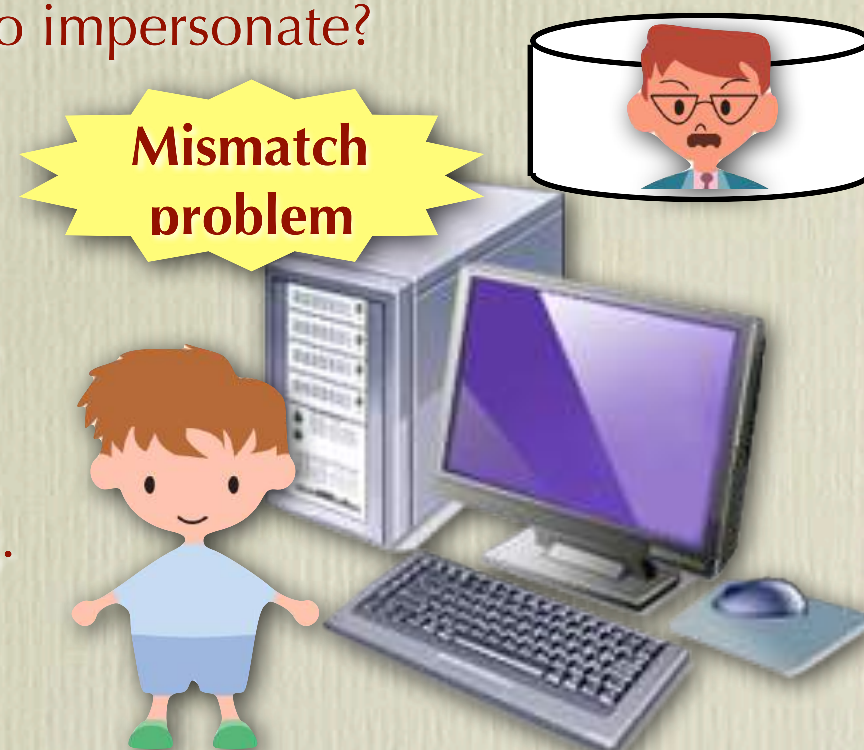
A big problem in CALL development

A very important and requisite function for CALL systems

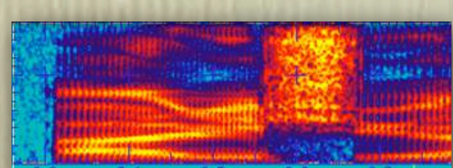
- The system has to be able to ignore speaker differences.
  - Age and gender (the size and length of the vocal tube)
  - But no current system can ignore speaker differences well enough.
- Requirement of "acoustic matchedness" bet. HMMs and learners
  - Collection of children's speech or speaker adaptation of adult HMMs
  - Q : Learning to pronounce is learning to impersonate?



?



- Speech model for another separation
  - Separation between source and filter
  - Separation between ling. and extra-ling.



-

= ?

A big solution for CALL development

To which does Minematsu's normal English sound closer ?

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

(Minematsu@ICSLP 2004)

A big solution for CALL development

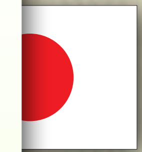
- Proficiency estimation based on P(o | M)

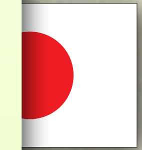
(Minematsu@ICSLP 2004)

A big solution for CALL development

- **Proficiency estimation based on P(M | o) = GOP**

$$
\begin{aligned}
P(M|o) &= P(p_1, ..., p_N|o) \\
&= \frac{P(o|p_1, ..., p_N)P(p_1, ..., p_N)}{\sum_{p_i} P(o|p_1, ..., p_N)P(p_1, ..., p_N)} \\
&\approx \frac{P(o|p_1, ..., p_N)}{\sum_{p_i} P(o|p_1, ..., p_N)} \\
&\approx \frac{P(o|p_1, ..., p_N)}{\max_{p_i} P(o|p_1, ..., p_N)} \\
&= \frac{P(o|M)}{\max_M P(o|M)} \\
&= \text{GOP (Goodness Of Pronunciation)}
\end{aligned}
$$

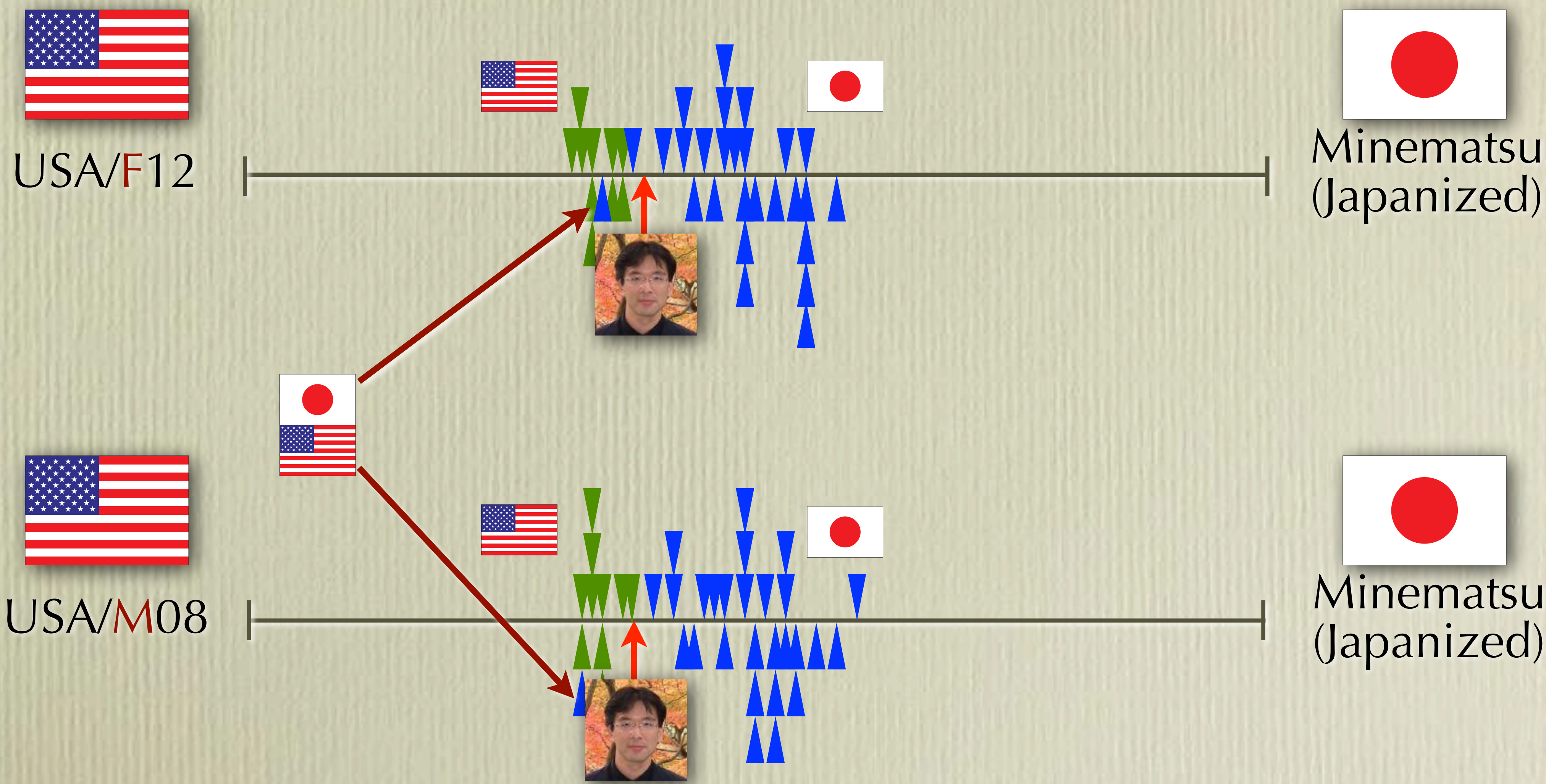




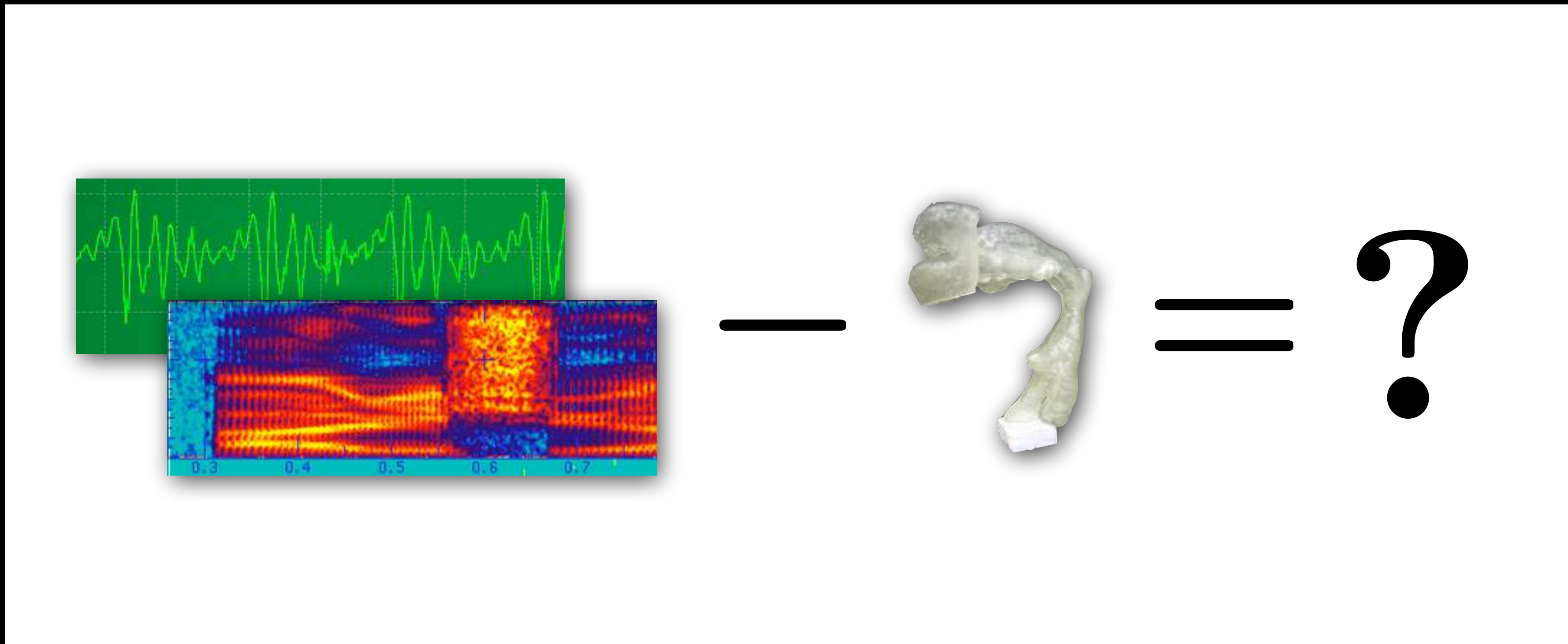
(Minematsu@ICSLP 2004)

A big solution for CALL development

Proficiency estimation based on structural distance



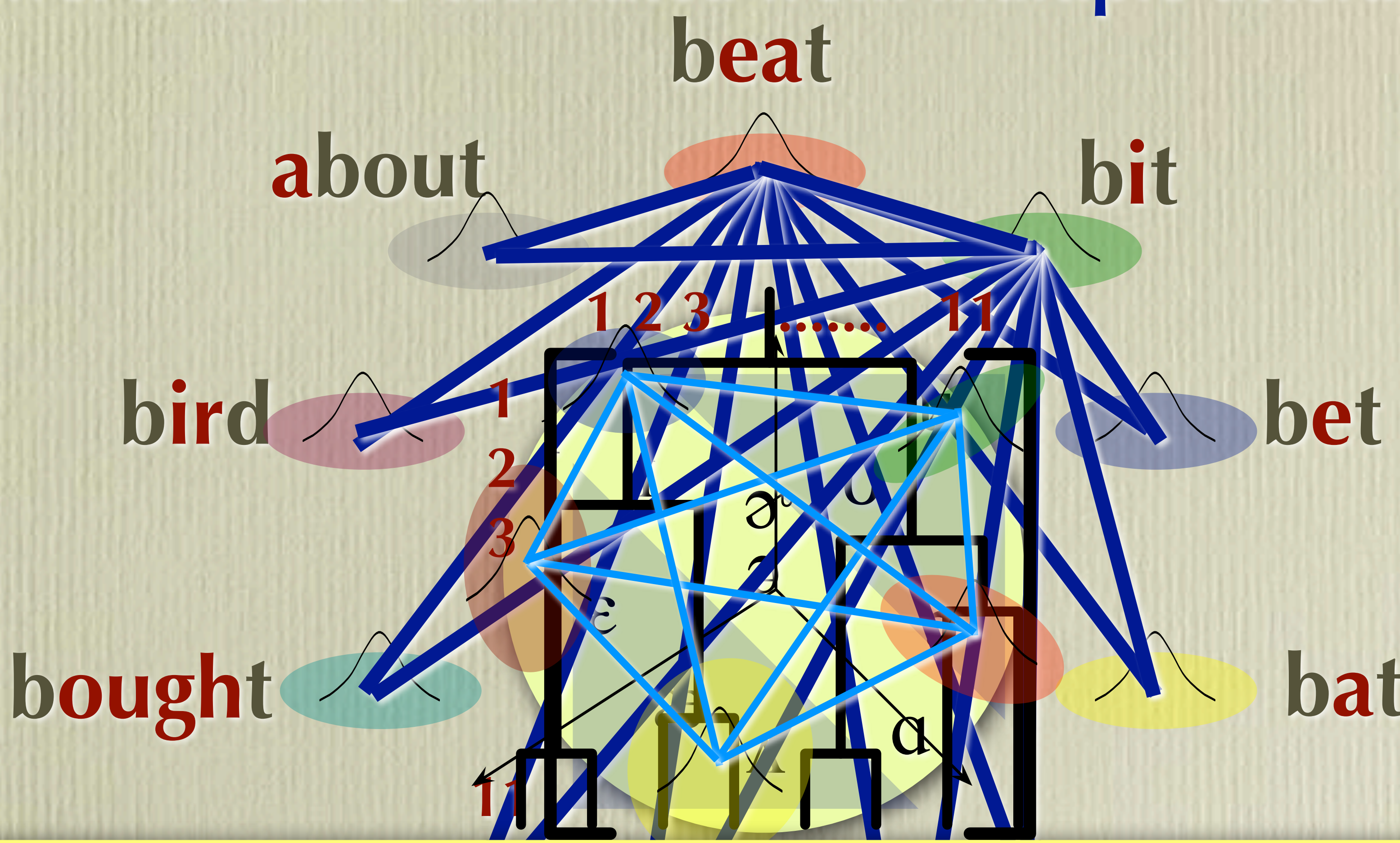
(Minematsu@ICSLP 2004)



proficiency	perfect △	good ✕	Japanized

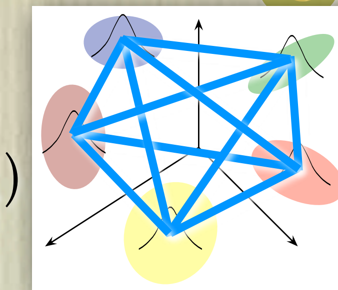
Application of structures to CALL

- Vowel structure estimated from multiple utterances



Evaluation is done not based on whether each vowel sound has adequate acoustic property independently of others but based on whether a good vowel system underlies a learner's pronunciation.

Clustering of learners



Preparation of data -- **96** simulated learners --

- ***12*** Japanese students who are returnees from US (***A*** to ***L***)
- English words of /b-V-t/ and Japanese words of /b-V-to/
  - AE vowels : 1 word utterance per vowel
  - J vowels : 5 word utterances per vowel
  - Vowel segments are extracted automatically to estimate a vowel system.

Replacement of some AE vowels with J vowels

- **12** speakers **[A-L]** x **8** pronunciations **[1-8]** = **96** learners

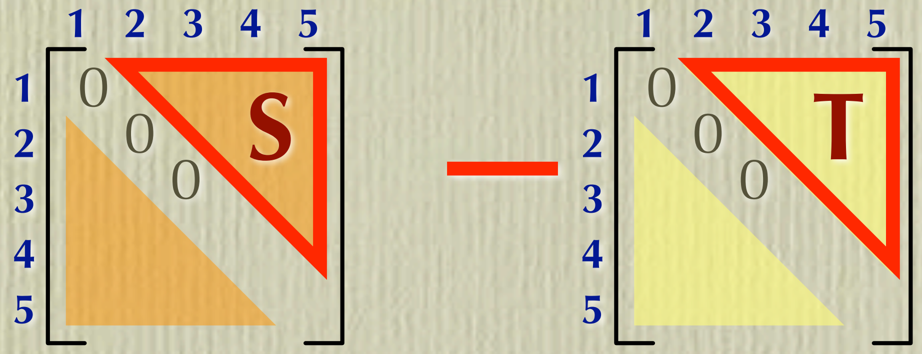
	ɑ	æ	ʌ	ə	ɚ	ɪ	i	ʊ	u	ɛ	ɔ
S1	J	J	J	J	J	J	J	J	J	J	J
S2	E	E	E	E	E	J	J	J	J	J	J
S3	J	J	J	J	J	E	E	E	E	E	E
S4	E	E	J	J	J	E	E	J	J	E	E
S5	J	J	E	E	E	J	J	E	E	J	J
S6	E	J	E	J	E	J	J	J	J	E	E
S7	J	E	J	E	J	E	E	E	E	J	J
S8	E	E	E	E	E	E	E	E	E	E	E

ɑ, æ, ʌ, ə, ɚ	a
ɪ, i	i
ʊ, u	u
ɛ	e
ɔ	o

Clustering of learners

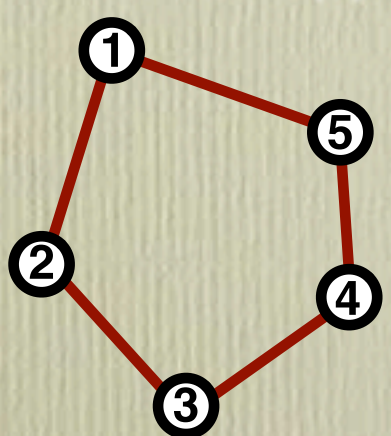
Structure-to-structure distance measure

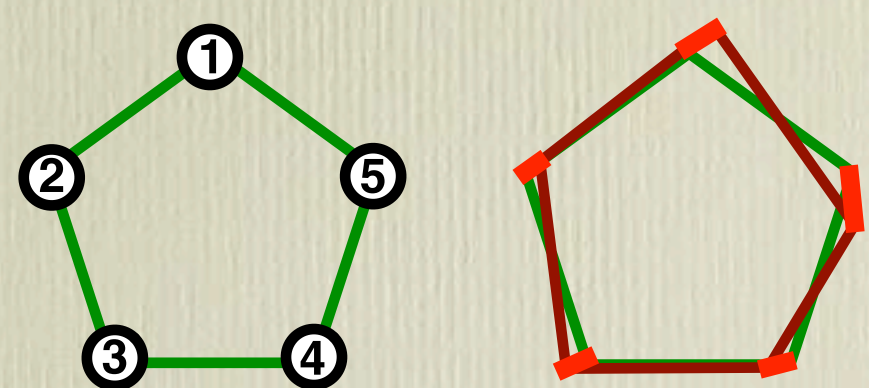
- Euclidian distance between two distance matrices



$$\sqrt{\frac{1}{M}\sum_{i<j}(S_{ij}-T_{ij})^2}$$

- Can approximate the structural distance after shift and rotation



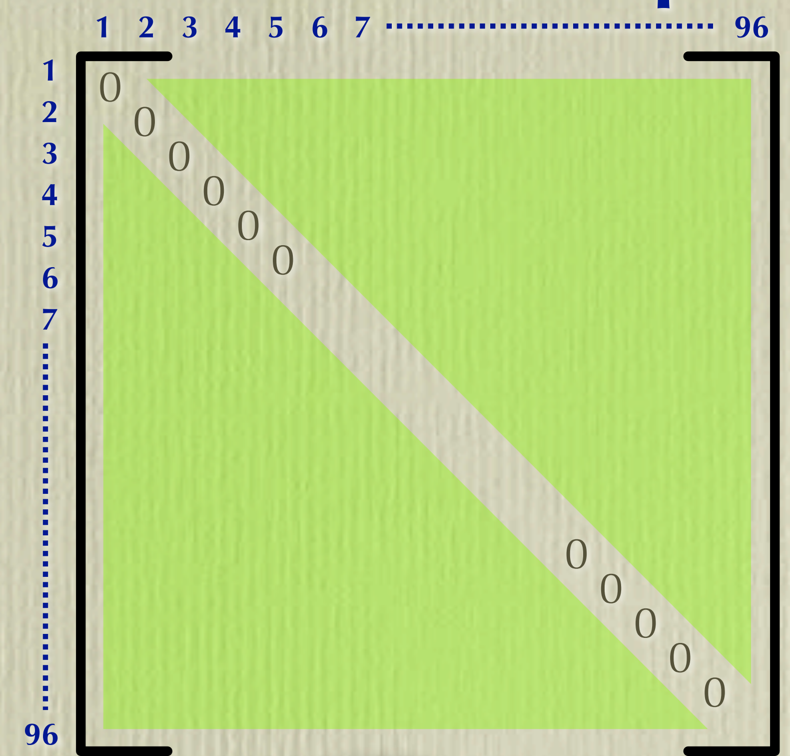


Minimum of the total distances between corresponding points

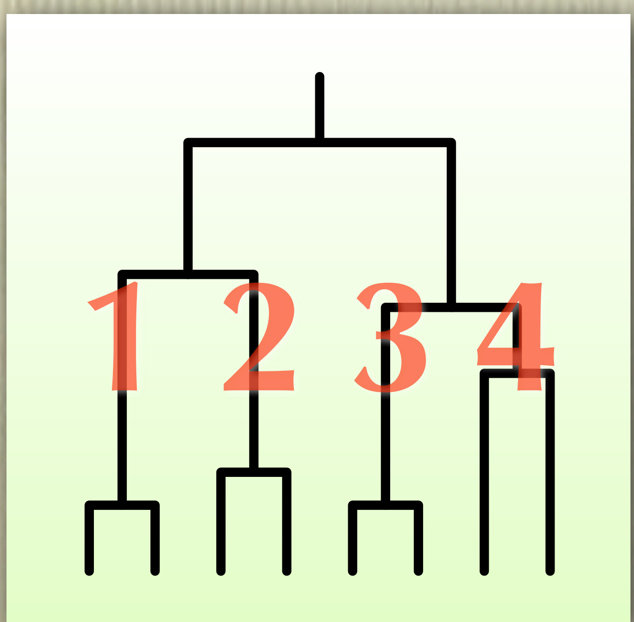
Clustering of learners

96 x 96 large distance matrix (12 spk. x 8 pron.)

- Speakers: A to L
- Prons: 1 to 8

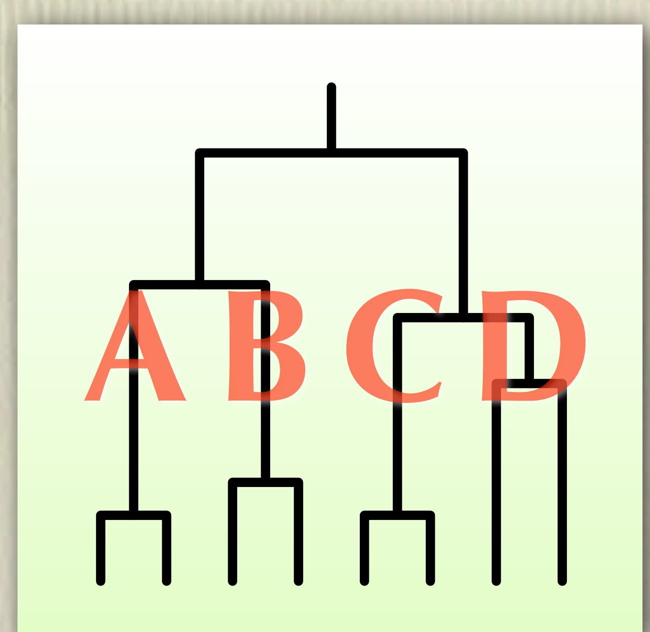


Pronunciation classification



?

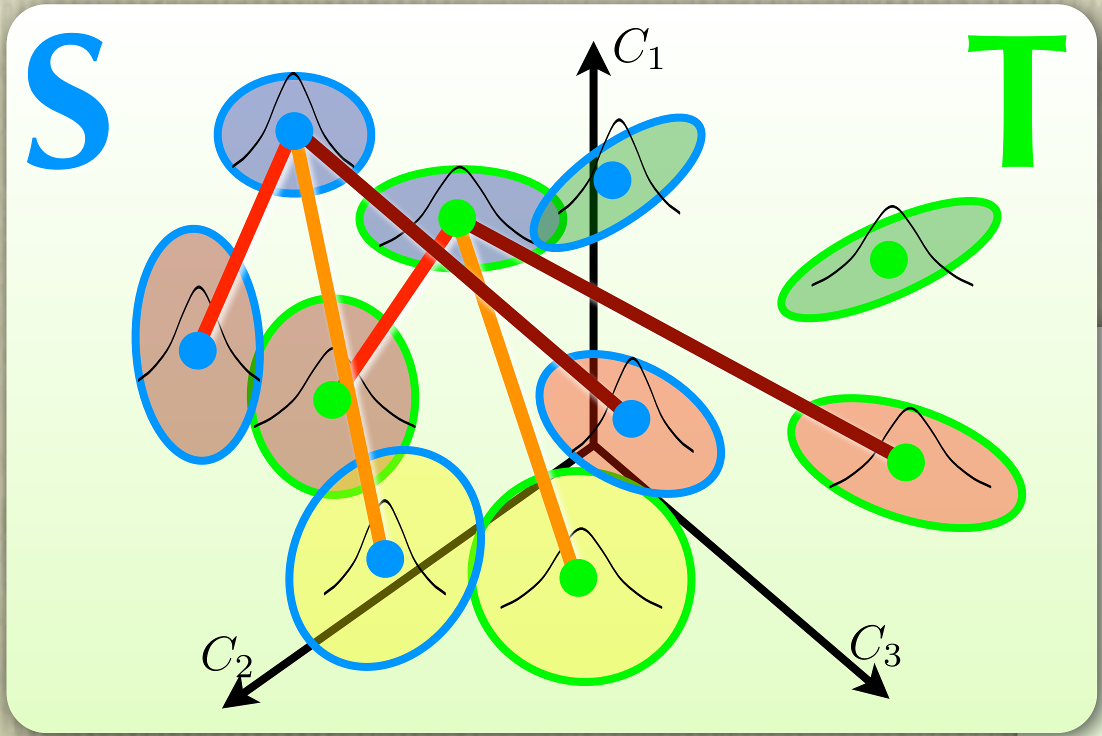
Speaker classification



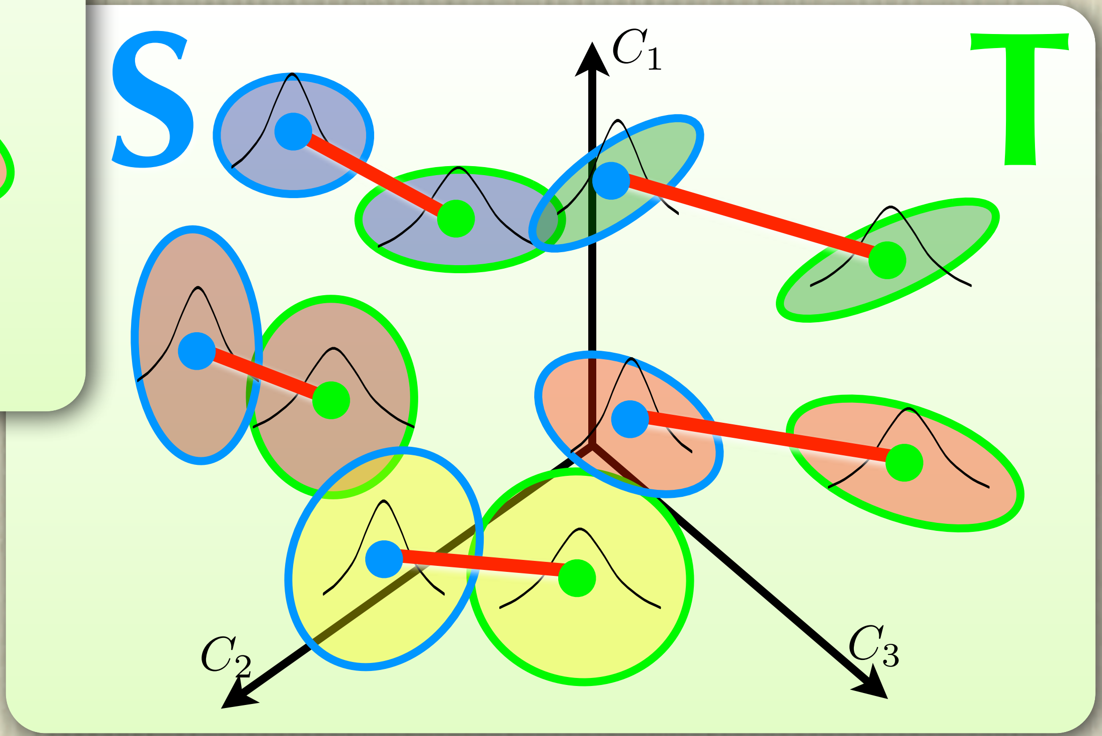
Clustering of learners

Another distance measure between two structures

- Contrast-based comparison
- Substance-based comparison



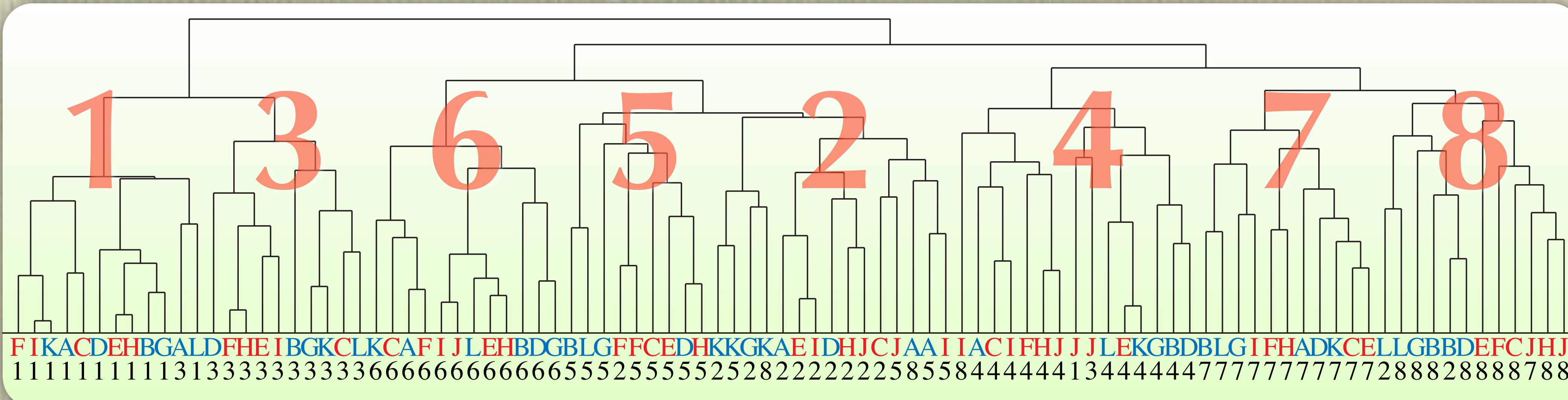
$$\sqrt{\frac{1}{M}\sum_{i<j}(S_{ij}-T_{ij})^2}$$



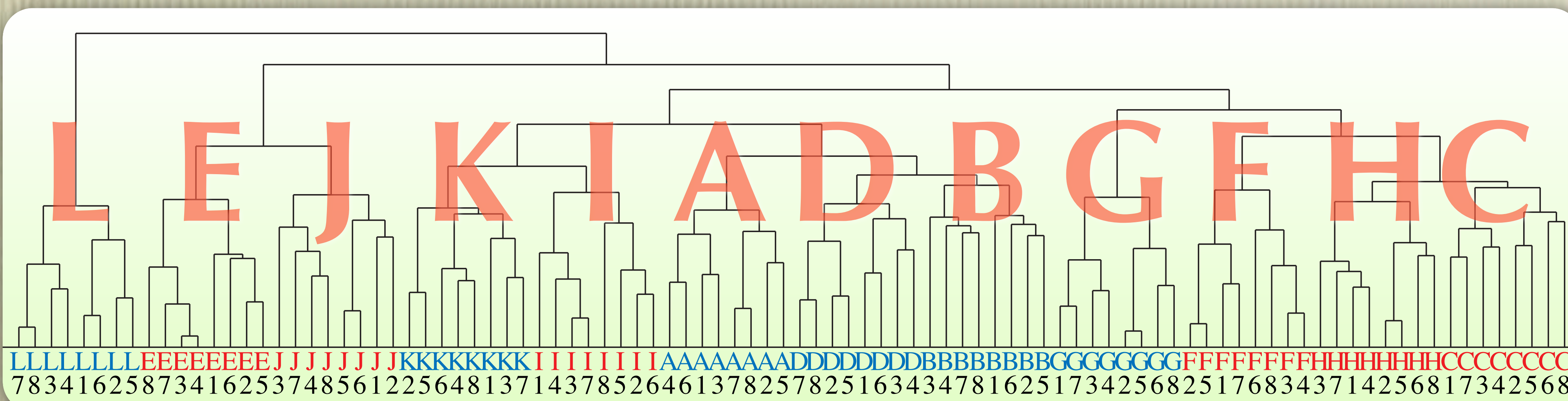
$$\sqrt{\frac{1}{M}\sum_{i} BD(v_i^S, v_i^T)}$$

Clustering of learners

Contrast-based comparison

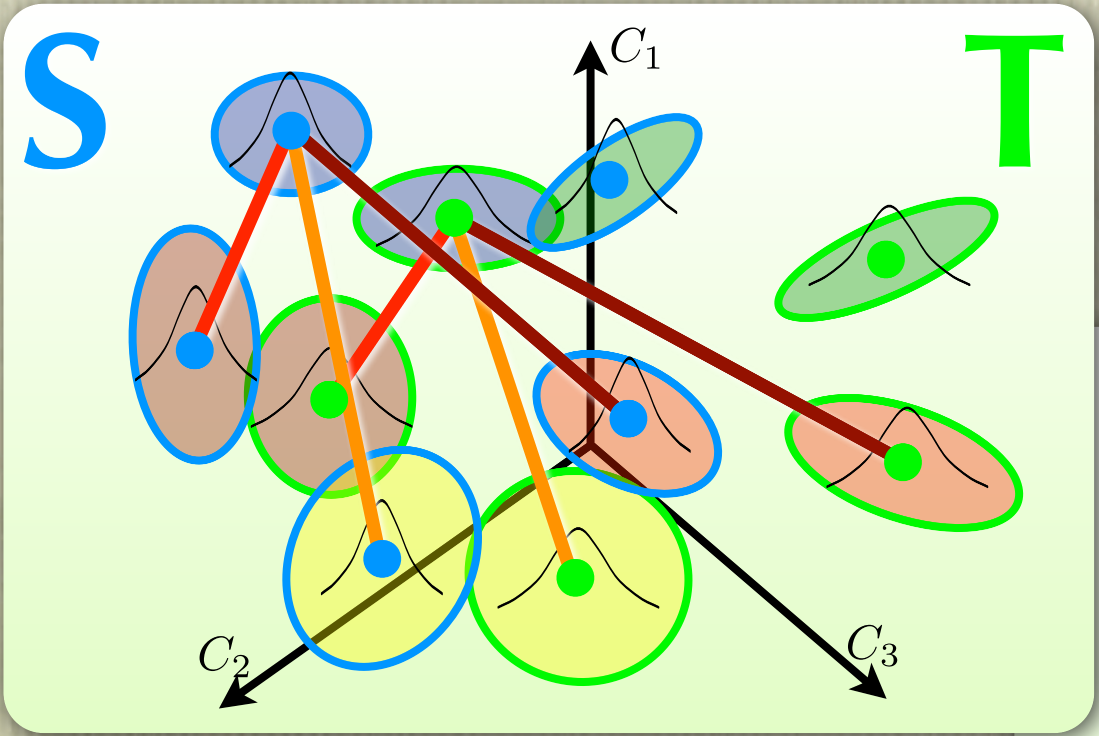


Substance-based comparison

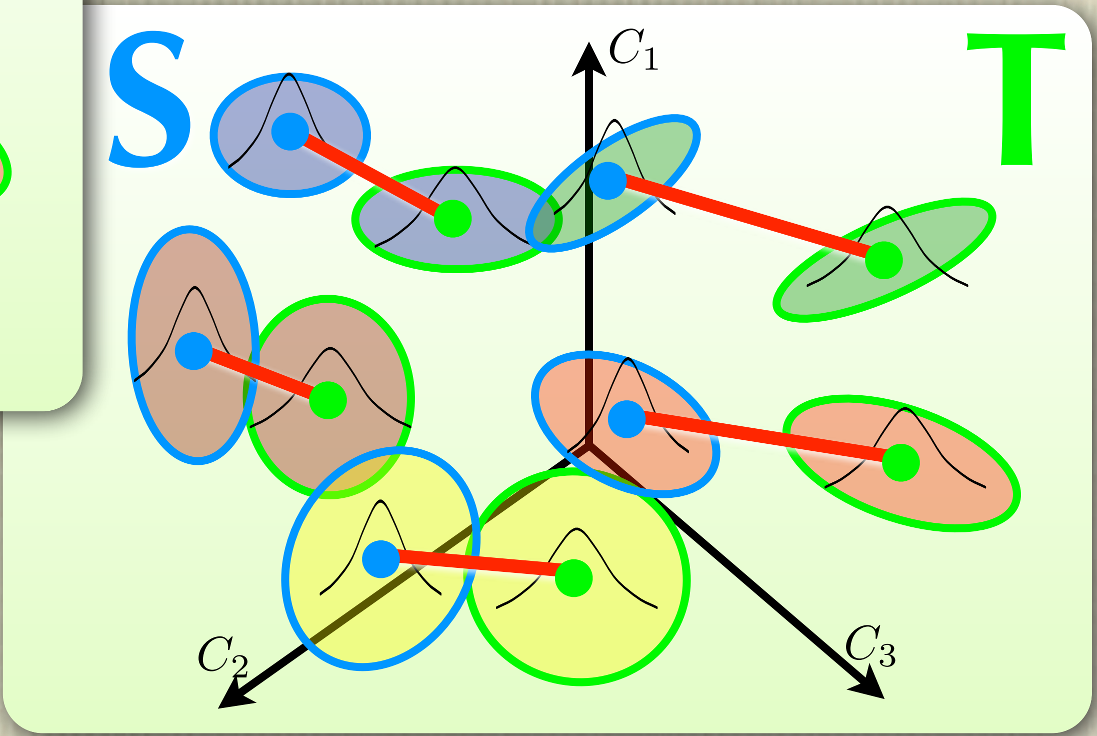


Clustering of learners

- **Another distance measure between two structures**
  - Contrast-based comparison
  - Substance-based comparison



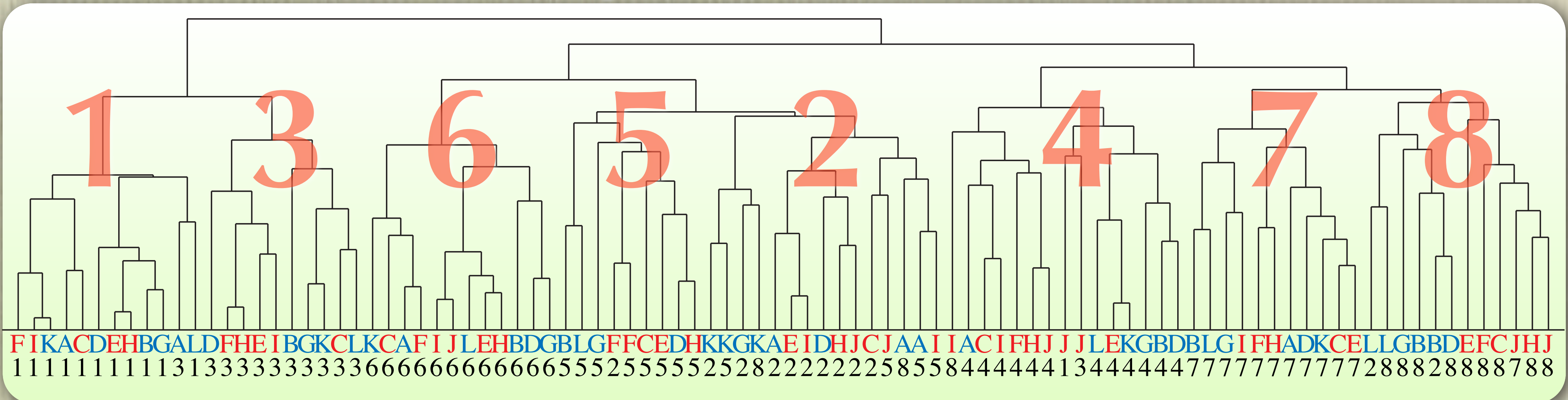
$$\sqrt{\frac{1}{M}\sum_{i<j}(S_{ij}-T_{ij})^2}$$



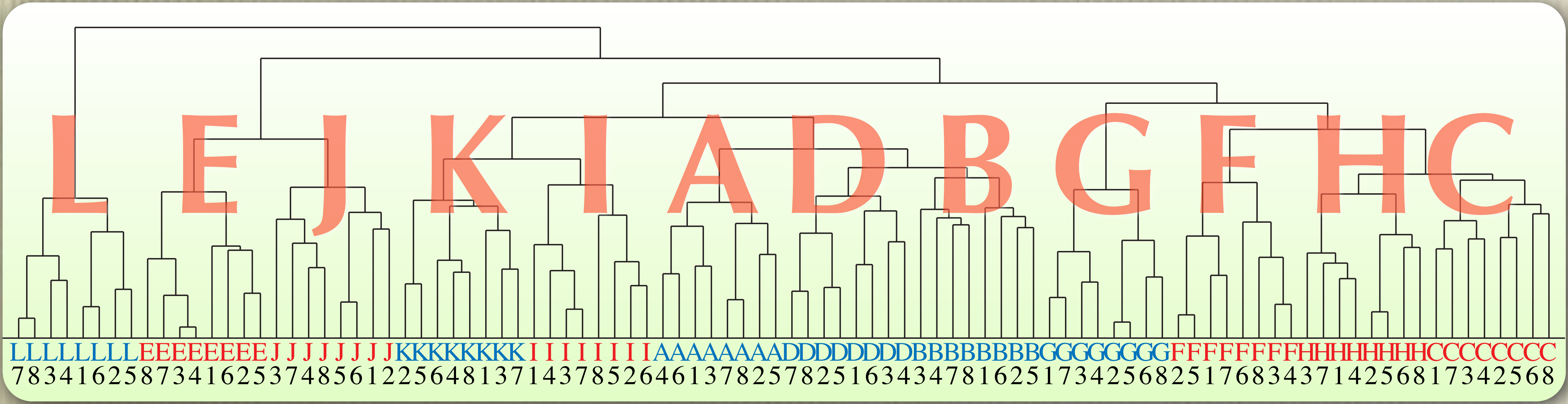
$$\sqrt{\frac{1}{M}\sum_{i}BD(v_i^S, v_i^T)}$$

Clustering of learners

Contrast-based comparison

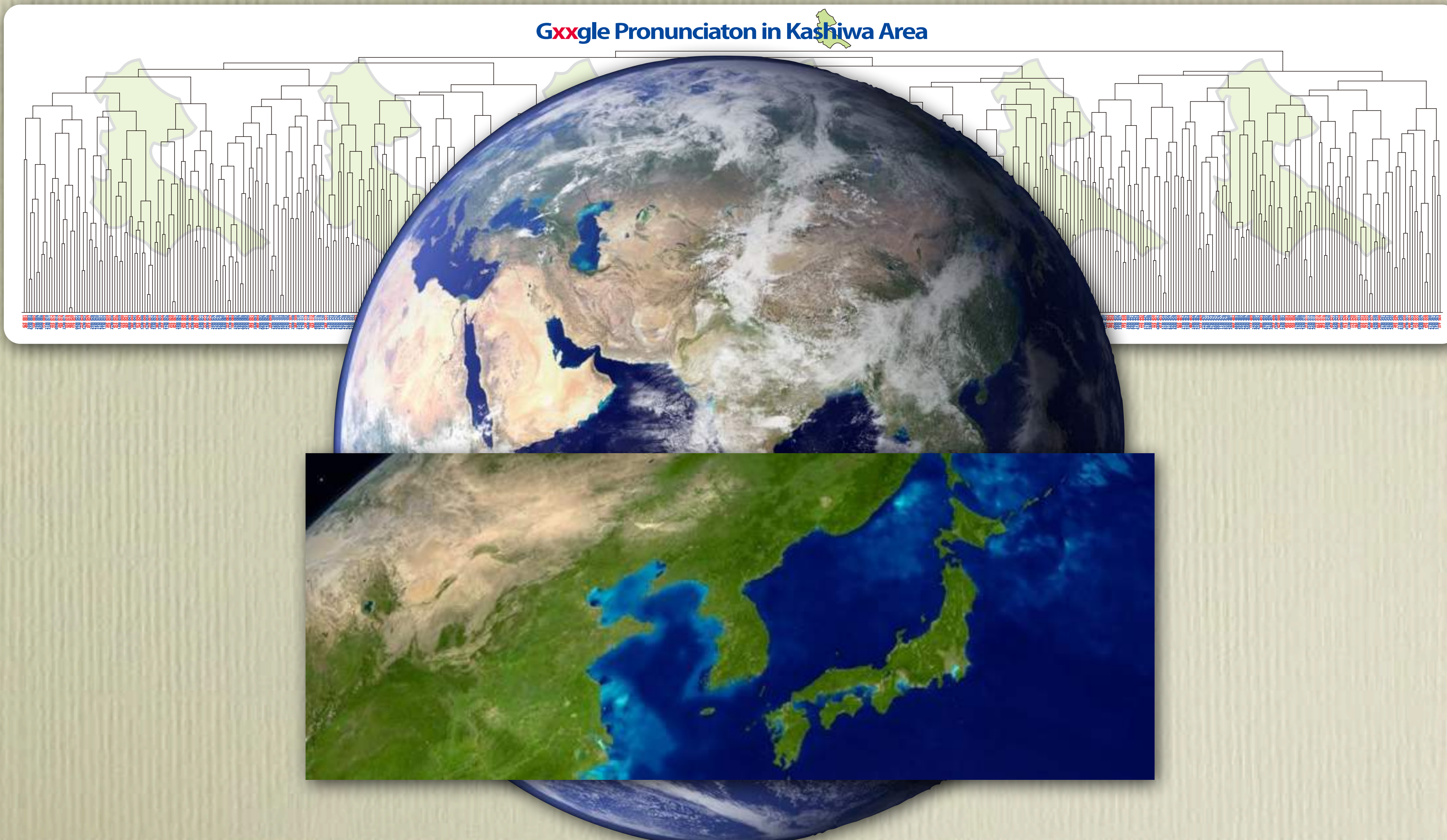


Substance-based comparison



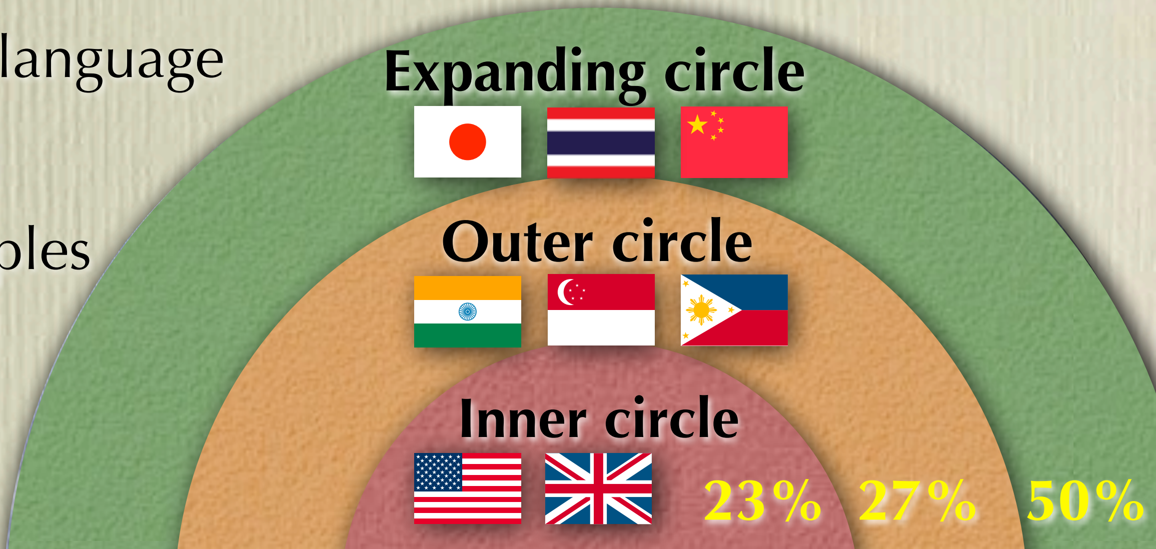
Clustering of "Kashiwa" Englishes

Classification of 600 citizens living in Kashiwa city



The current state of English

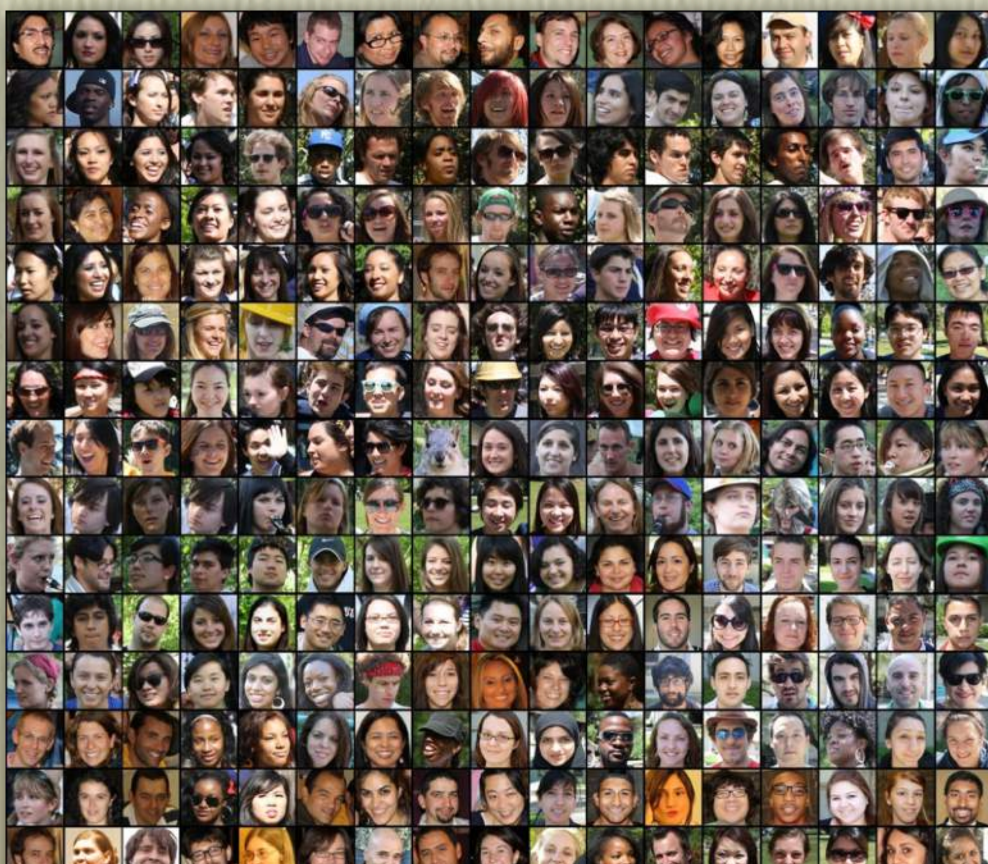
- **It is the only language used for global communication.**
  - About **1.5 billion** users on earth
- **It has the largest diversity in its form.**
  - Internationalization of a thing inevitably alters its form.
  - English is not exceptional.
    - Syntax, pragmatics, lexical choice, spelling, **pronunciation**, etc
- **World Englishes (WE)**
  - Three circles model [Kachru1992]
    - E as native / official / foreign language
  - No standard pronunciation
    - AE and BE are just two examples of **accented** Engilshes.



Pronunciation diversity of WE

Is English a useful tool or a troublesome tool?

- A useful tool for global communication
  - The same language can be shared by all.
- A troublesome tool for global communication
  - Its pronunciation diversity can cause miscommunications.



http://academicaffairs.ucdavis.edu/diversity/

Diversity of pronunciation in WE

What is the minimal unit and how many units?

Country?

↓

Region / State / Prefecture?

↓

City / Town / Village?

↓

Individual!

↓

1.5 billions!

[Kachru 1992]

Huge pron. diversity in World Englishes

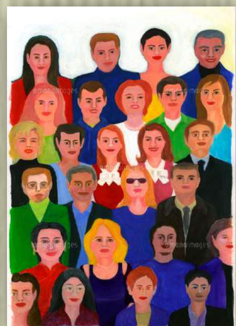
1) native language, 2) official language, and 3) foreign language

Huge pron. diversity in World Englishes

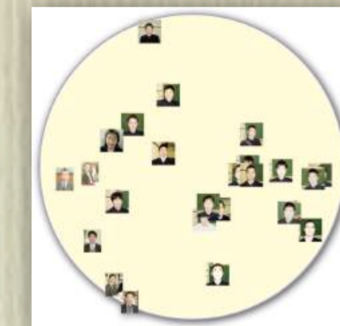
1) native language, 2) official langua

Speaker-basis pronunciation clustering

Requires a speaker-basis pronunciation distance matrix

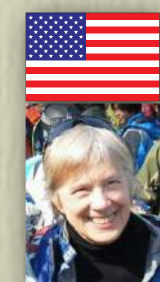


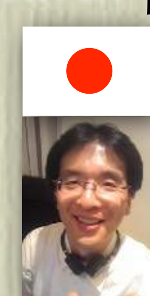
$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ d_{31} & & & \\ \vdots & & & \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix}$$



What is technically challenging?

- To which is Minematsu's natural pronunciation closer?



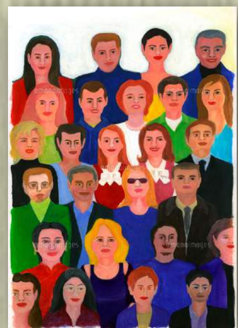


"Those answers will be straightforward if you think them through carefully first."

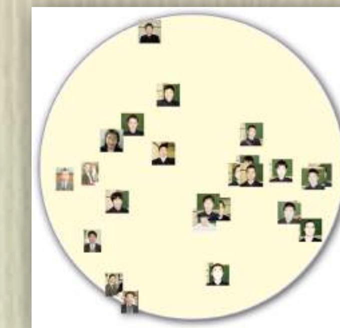
- Pronunciation distance = phonetic distance between speakers
  ≠ acoustic distance between speakers
  ≠ spectral distance between speakers

Speaker-basis pronunciation clustering

Requires a speaker-basis pronunciation distance matrix

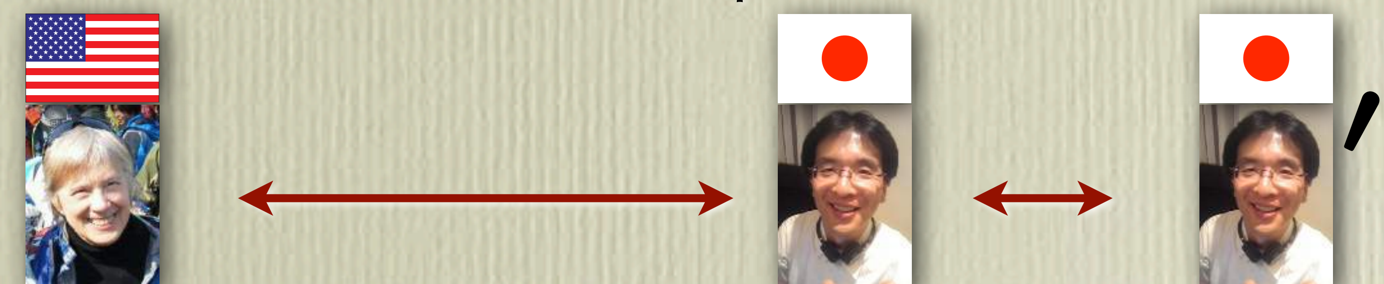


$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ d_{31} & & & \\ \vdots & & & \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix}$$



What is technically challenging?

- To which is Minematsu's natural pronunciation closer?

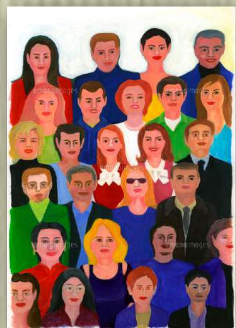


"Those answers will be straightforward if you think them through carefully first."

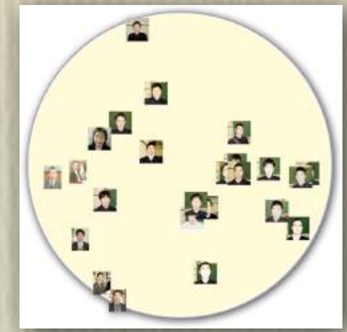
- Pronunciation distance = phonetic distance between speakers
  ≠ acoustic distance between speakers
  ≠ spectral distance between speakers

Speaker-basis pronunciation clustering

Requires a speaker-basis pronunciation distance matrix

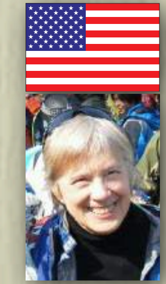


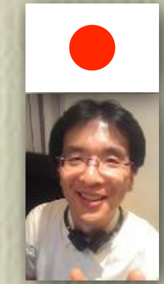
$$
\begin{bmatrix}
d_{11} & d_{12} & \dots & d_{1N} \\
d_{21} & d_{22} & \dots & d_{2N} \\
d_{31} & & & \\
\vdots & & & \\
d_{N1} & d_{N2} & \dots & d_{NN}
\end{bmatrix}
$$

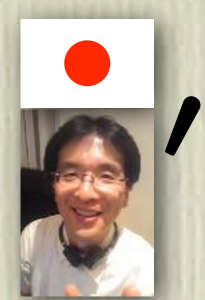


What is technically challenging?

- To which is Minematsu's natural pronunciation closer?

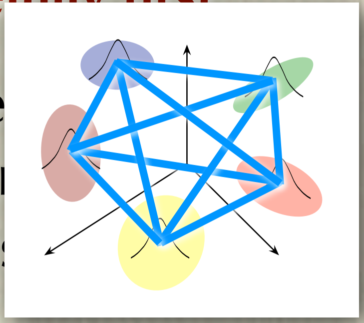






"Those answers will be straightforward if you think them through carefully first."

- Pronunciation distance = phonetic distance between speakers
  ≠ acoustic distance between speakers
  ≠ spectral distance between speakers



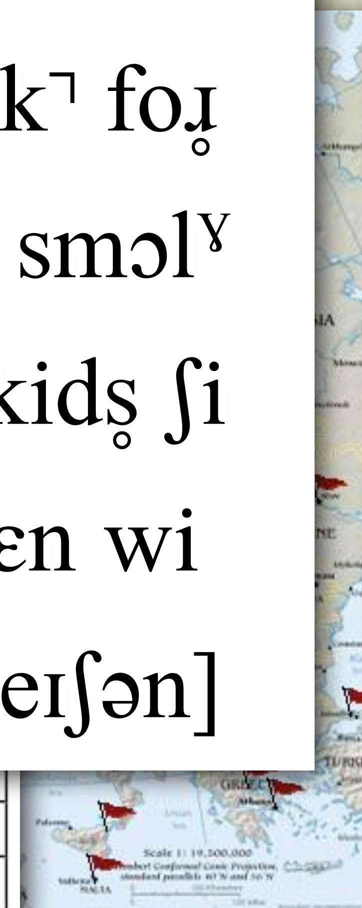
Pron. ... of WE

- Speech Ac...
  - A commo... ...ral speakers
  - ...

Vowels and Consonants used in Acoustic Analysis					
1. i	2. ĭ	3. iː	4. i̥	5. ï	6. ĩ
7. y	8. ɪ̞	9. ɪ	10. ɪː	11. ɪ̥	12. ɪ̃
13. e	14. ë	15. ẽ	16. ɛ	17. ɛ̈	18. ɛ̃
19. æ̞	20. æ	21. æː	22. æ̃	23. a	24. ã
139. ʔ	140. h	141. ɦ	142. w	143. ɥ	144. pɸ
145. tθ	146. dð	147. ts	148. dz	149. tɕ	150. dʑ
151. tʃ	152. dʒ	153. kx			

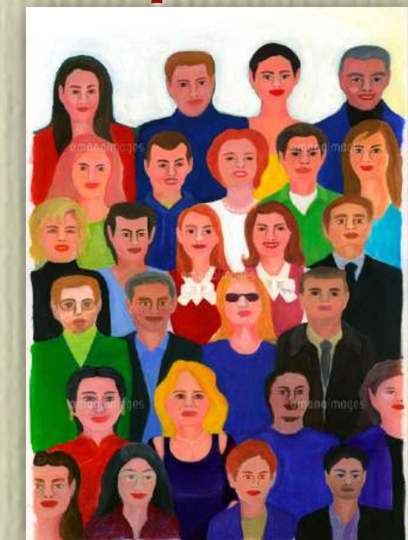
Plea...
the s...
chee...
smal...
scoo...
Wed...

[pʰlis kɔl stɛːlʌ ɑsk˺ ɜ tə bɹĩŋ ðiz θĩŋz̥ wɪf hɜ
fɹʌ̃m ðə stɔɹ siks spunz̥ əv̥ fɹɪʃ sn̥oʊ piːs faɪf
θɪk slæb̥s əv blu ʧiːz ɛn măɪbi ɜ snæk˺ foɹ̥
hɜ bɹɑɾə̆ ʔə brʌðə bɑp wi ɔlˠsŏ nid ə smɔlˠ
plæstɪk sneɪk ɛn ə bɪk tʊi fɹɔg̥ fɛ̃ ðə kids̥ ʃi
kɛ̆n skøp ðiz θĩŋs ɪntu fɹi ɹɛd bægz̥ ɛn wi
wɪl goʊ miːd ɜ̆ wɛ̆nz̥d̥eɪ ɛt d̪ə tɹeɪn steɪʃən]

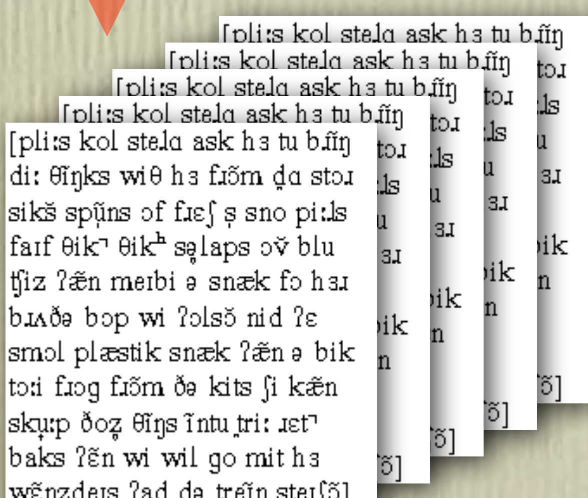


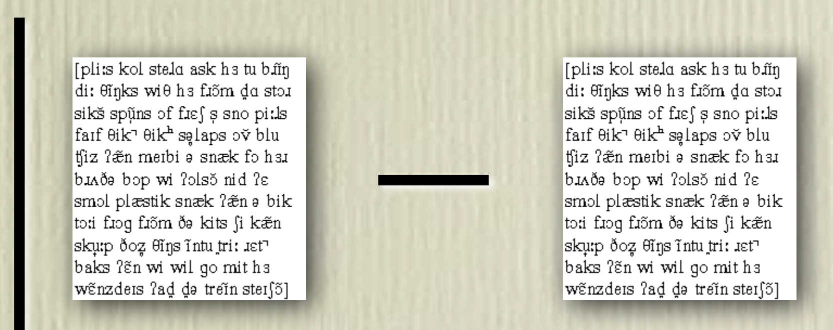
Pron. clustering only based on SAA

N speakers



$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ N \end{array}
\begin{array}{c}
\begin{array}{cccc} 1 & 2 & \dots & N \end{array} \\
\begin{bmatrix}
d_{11} & d_{12} & \dots & d_{1N} \\
d_{21} & d_{22} & \dots & d_{2N} \\
d_{31} & & & \\
\vdots & & & \\
d_{N1} & d_{N2} & \dots & d_{NN}
\end{bmatrix}
\end{array}
$$

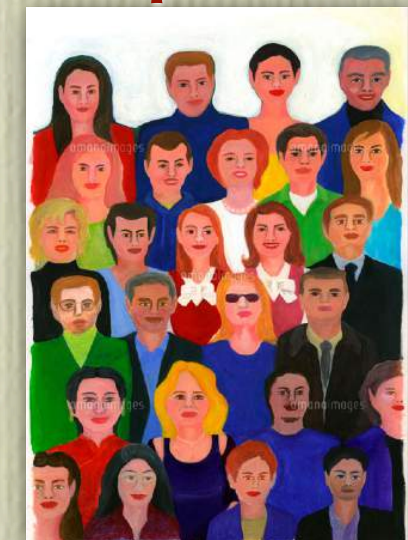

[pliːs kol stelə ask hɜ tu bɹ̃iŋ
diː θiŋks wiθ hɜ fɹɔ̃m ɖa stoɹ
sikš spũns of fɹɛʃ ʂ sno piːls
faɪf θik˺ θikʰ sɐlaps oṽ blu
ʧiz ʔæ̃n meɪbi ə snæk fo hɜɹ
bɹʌðə bop wi ʔolsõ nid ʔɛ
smol plæstik snæk ʔæ̃n ə bik
toːi fɹog fɹɔ̃m ðə kits ʃi kæ̃n
skuːp ðoz θiŋs ĩntu ɹriː ɹɛt˺
baks ʔɛ̃n wi wil go mit hɜ
wɛ̃nzdeɪs ʔad ɖə tɹeĩn steɪʃɔ̃]

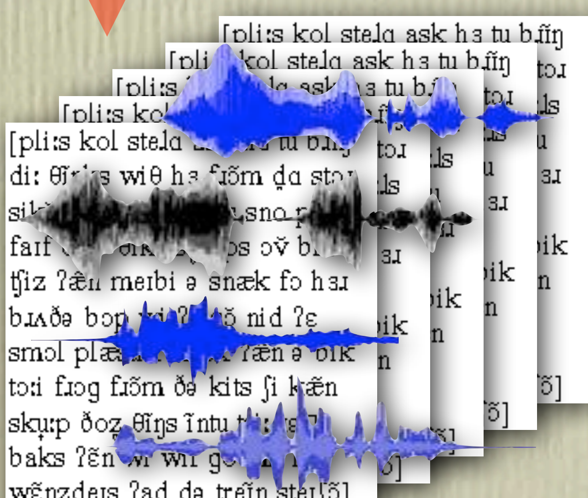


[Miller et al.'95, Bailey et al.'05, Wieling et al.'12]

Pron. clustering only based on SAA

N speakers





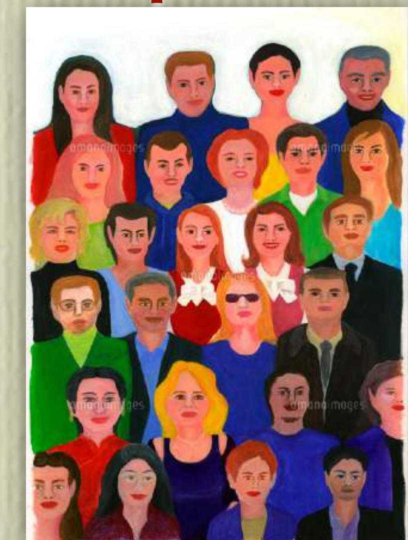
	1	2	...	N
1	d_{11}	d_{12}	$\cdots$	d_{1N}
2	d_{21}	d_{22}	$\cdots$	d_{2N}
3	d_{31}			
:	$\vdots$			
N	d_{N1}	d_{N2}	$\cdots$	d_{NN}

Pron. Structure Analysis

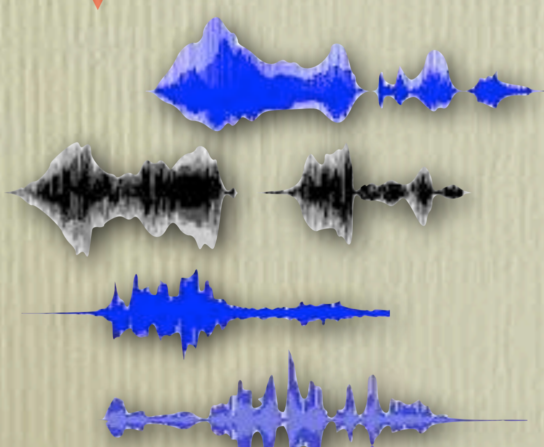
[Miller et al.'95, Bailey et al.'05, Wieling et al.'12]

Pron. clustering only based on SAA

N speakers



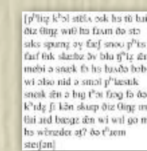
$$\{d_{mn}\} \approx \{p_{mn}\} \text{ ?}$$



Pron. Structure Analysis

IPA-b… …tance

- **Optimal al… …et al.,'13]**
  - Dynamic T…



- = ?
  - DTW can…

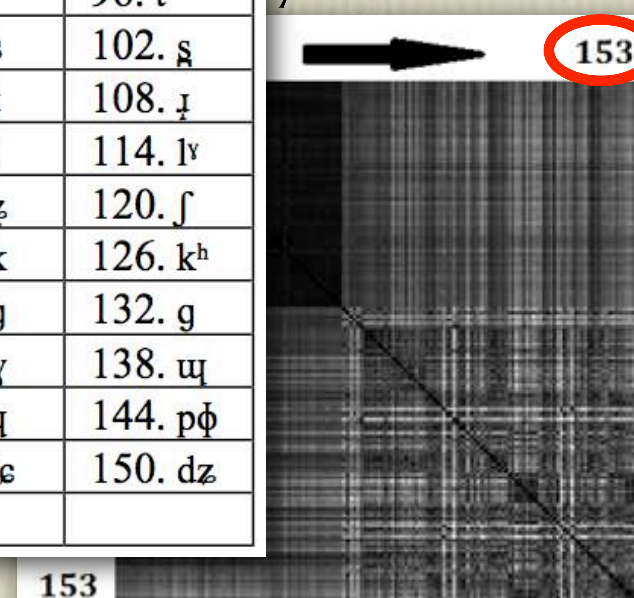
p_h l i z … 0 l s o

p_h l i s … s o

  - Similar to … Wieling et. al,'12]
- DTW requ… …symbols used.
  - 20 produ…
  - HMM is b…
    - HMM =…
  - Acoustic … each HM…

Vowels and Consonants used in Acoustic Analysis

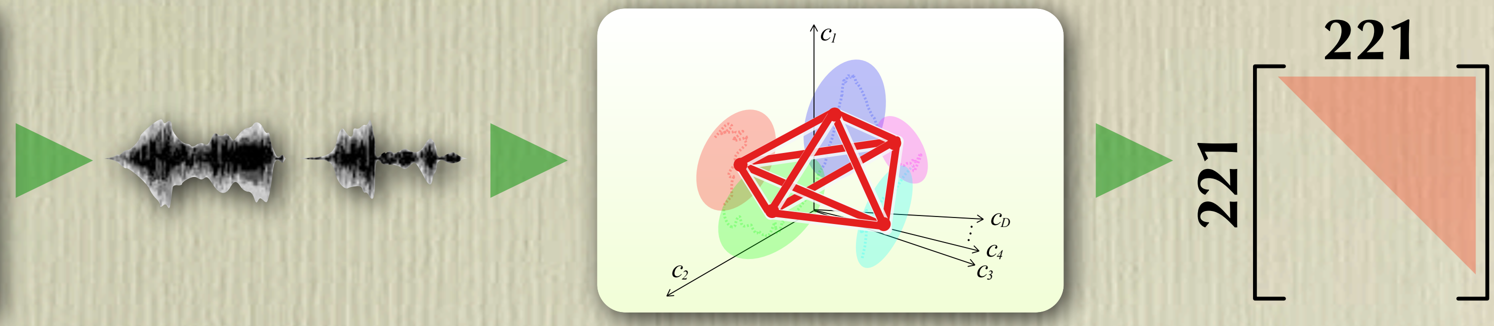
1. i	2. ĭ	3. iː	4. i̥	5. ï	6. ĩ
7. y	8. ɪ̥	9. ɪ	10. ɪː	11. ɪ̯	12. ɪ̃
13. e	14. ë	15. ẽ	16. ɛ	17. ɛ̈	18. ɛ̃
19. æ̝	20. æ	21. æː	22. æ̃	23. a	24. ã
25. ɨ	26. ɨ̥	27. ɨ̃	28. ʉ	29. ʉ̥	30. ɚ
31. ɜ	32. ɝ	33. ɐ	34. ɐ̃	35. ʉ̃	36. ɵ
37. ɵ̃	38. ə	39. ə̆	40. ə̥	41. ə̃	42. ə̯
43. ɯ	44. ɯ̈	45. ɯ̃	46. u	47. ŭ	48. uː
49. ü	50. ũ	51. ũː	52. ʊ	53. ɤ	54. o
55. ö	56. õ	57. ʌ	58. ʌ̃	59. ɔ	60. ɔː
61. ɔ̈	62. ɔ̃	63. ɑ	64. ɑː	65. ɑ̈	66. ɑ̃
67. p	68. pʰ	69. p̚	70. b	71. b̚	72. b̥
73. ɸ	74. β	75. β̦	76. β̃	77. f	78.v
79. y̥	80. ʋ	81.m	82. m̥	83. m̩	84.n
85. n̥	86. n̪	87. n̩	88. ɲ	89. ŋ	90. ɴ
91.t	92. tʰ	93. t̪	94. ţ	95. t’	96. t̚
97. d	98. d̪	99. d̚	100. d̥	101. s	102. s̪
103. sʲ	104. z	105. z̥	106. ɹ	107. ɹ̥	108. ɹ̝
109. r	110. ɾ	111. ɽ	112. l	113. l̩	114. lˠ
115. θ	116. ð	117. ɕ	118. ʑ	119. ʑ̥	120. ʃ
121. ʒ	122. ç	123. j	124. j̩	125. k	126. kʰ
127. k̩	128. k’	129. k̩ʰ	130. k̚	131. ɡ	132. g
133. ɡ̚	134. ɡ̊	135. x	136. ɣ	137. ɣ̥	138. ɰ
139. ʔ	140. h	141. ɦ	142. w	143. ɥ	144. pɸ
145. tθ	146. dð	147. ts	148. dz	149. tɕ	150. dʑ
151. tʃ	152. dʒ	153. kx			

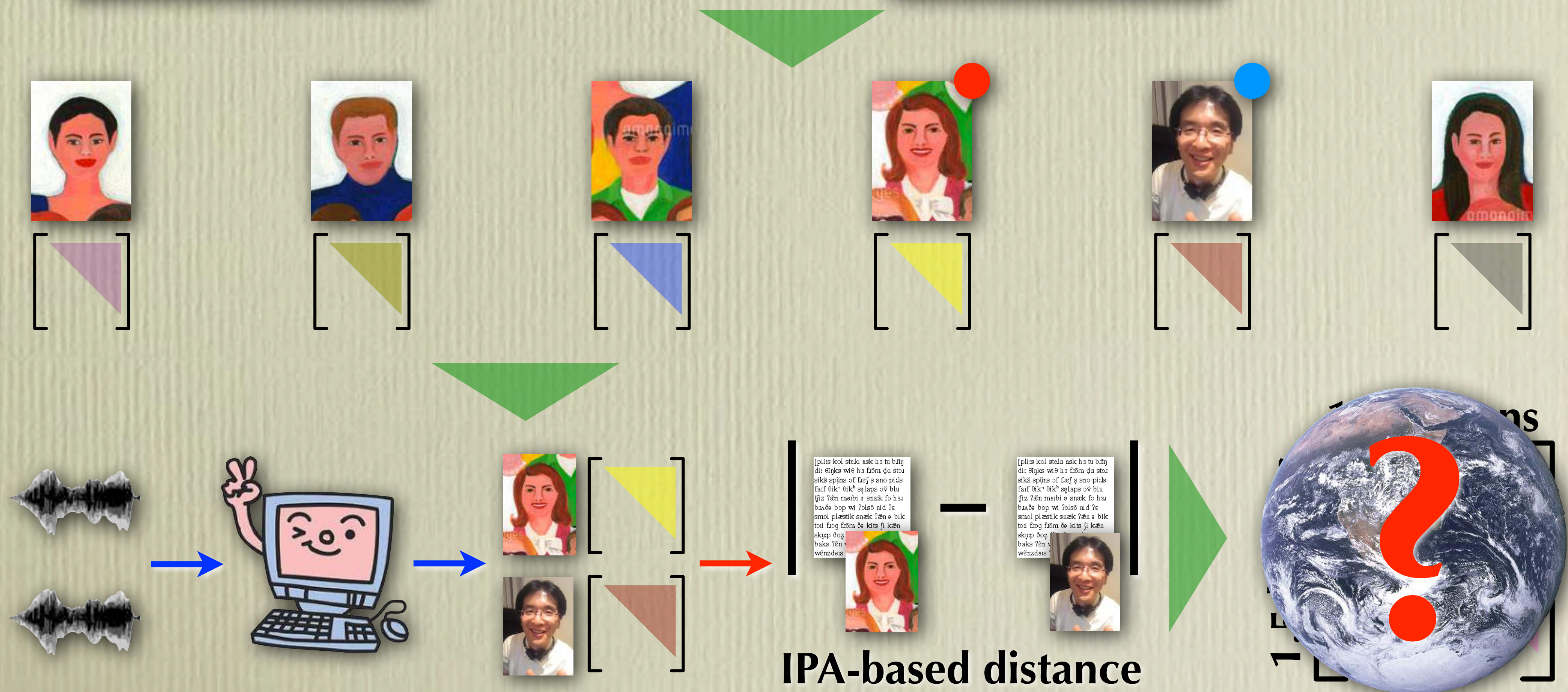


Pron. distance calculation using structure

A common paragraph to pron. structure

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack ....

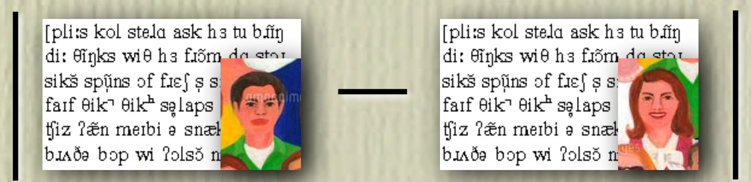




Pron. clustering using real data of WE

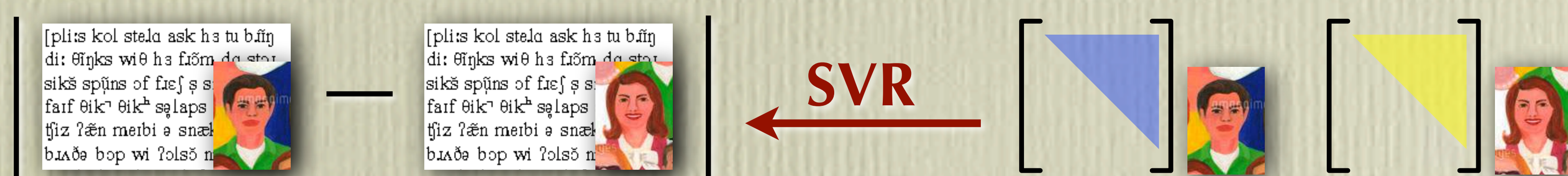
Use of IPA transcripts to prepare reference distances

- DTW-based calculation of the reference distance bet. transcripts



Prediction of the ref. distances using pron. structures

- SVR-based supervised prediction using structures as input features



Use of phonemic transcripts to calculate distances

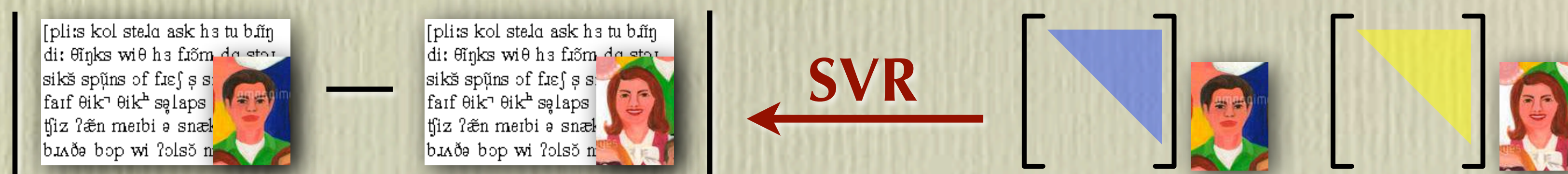
- Corresponds to calculate pron. distances somewhat coarsely.

/ pə̆liːz̥ kɔlˠ stɪlə æ̆sk hɜɹ tʷŏ bɹɪ̃ŋ /   #symbols = 153

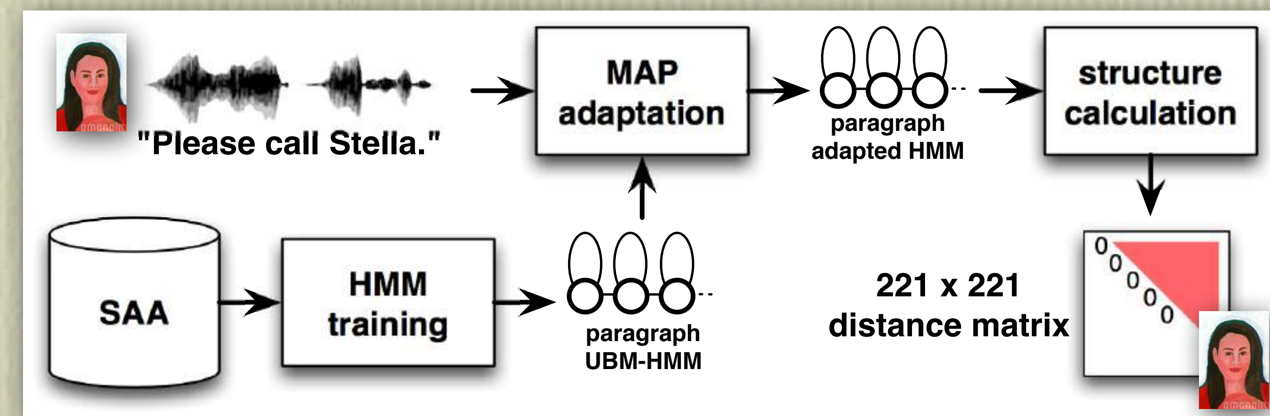
[p ah l iy z k ao l s t ih l ah ae s k hh ah r t ow b r ih ng]   #symbols = 39

Pron. clustering using real data of WE

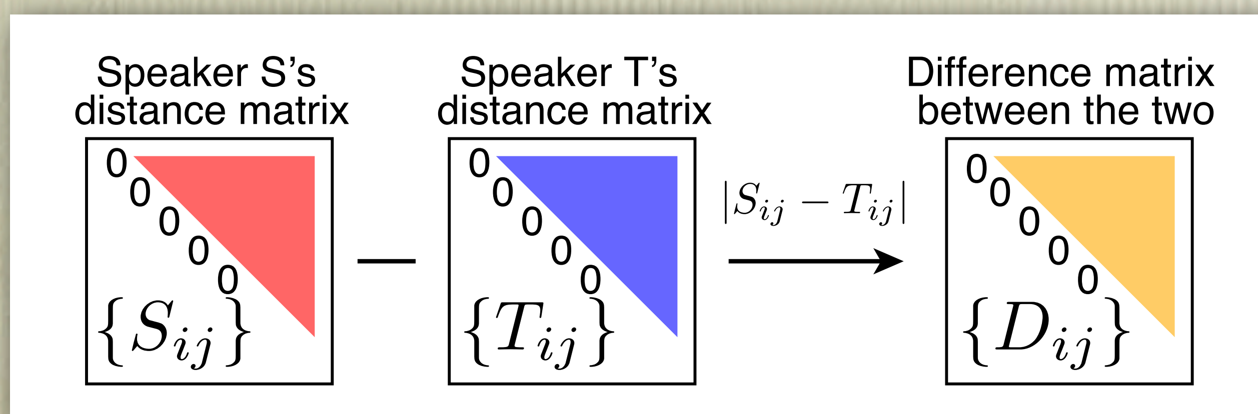
SVR-based prediction of IPA distances [Kasahara'14]



- Pronunciation structure extraction from an SAA sample



- Differential features from two pronunciation structures



Pron. clustering using real data of SAA

Three modes of preparing training data and testing data

- Speaker-open mode
  - SAA → two speaker groups of training and testing
- Speaker-**pair**-open mode
  - SAA → speaker pairs → two speaker pair groups of training and testing
- Speaker-open and speaker-pair-open mode

speaker-open

training	testing
A - B	D - H
B - C	Y - D
B - F	G - X
Z - A	M - J
:	:

Speakers are *not* shared.
Speaker pairs are *not* shared.

speaker-pair-open

training	testing
A - B	A - C
B - C	B - D
B - F	C - F
Z - A	Z - B
:	:

Speakers are shared.
Speaker pairs are *not* shared.

training $\{ T_i \}$	testing $\{ X_i \}$
T_1 - T_2	X_1 - T_1
T_1 - T_3	X_1 - T_2
T_4 - T_7	X_1 - T_3
T_5 - T_9	X_2 - T_8
:	:

← speaker-open (the training $\{ T_i \}$ / testing $\{ X_i \}$ split)

← speaker-pair-open (the pair lists)

Speakers are shared only partially.
Speakers pairs are *not* shared.

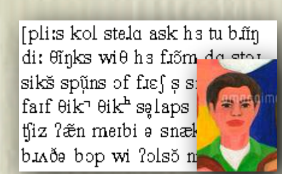
Pron. clustering using real data of SAA

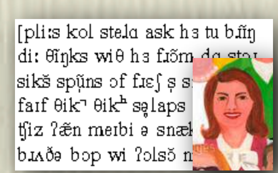
Corr. bet. IPA distances and predicted distances [Sato+'15]

mode	spk-open	spk-pair-open	both
corr.	**0.5**	**0.87**	**0.77**

Comparison with other possible methods

- Transcript-to-transcript distance based on phonemes
  - Phone : minimum unit of sounds perceived by phoneticians
  - Phoneme : minimum unit of sounds perceived by general listeners
- Rule-based conversion from IPA trans. to AE phonemic trans.
  - Trans.-to-trans. distances were obtained with phoneme HMMs + DTW.
  - Corr. = **0.75**
- Automatic AE phoneme recognition for SAA utterances
  - Phoneme recognition accuracy = 73.5%
  - Corr. = **0.46**





Pron. clustering using real data of SAA

Three modes of preparing training data and testing data

- Speaker-open mode
  - SAA → two speaker groups of training and testing
- Speaker-**pair**-open mode
  - SAA → speaker pairs → two speaker pair groups of training and testing
- Speaker-open and speaker-pair-open mode

speaker-open

training	testing
A - B	D - H
B - C	Y - D
B - F	G - X
Z - A	M - J
:	:

Speakers are *not* shared.
Speaker pairs are *not* shared.

speaker-pair-open

training	testing
A - B	A - C
B - C	B - D
B - F	C - F
Z - A	Z - B
:	:

Speakers are shared.
Speaker pairs are *not* sha

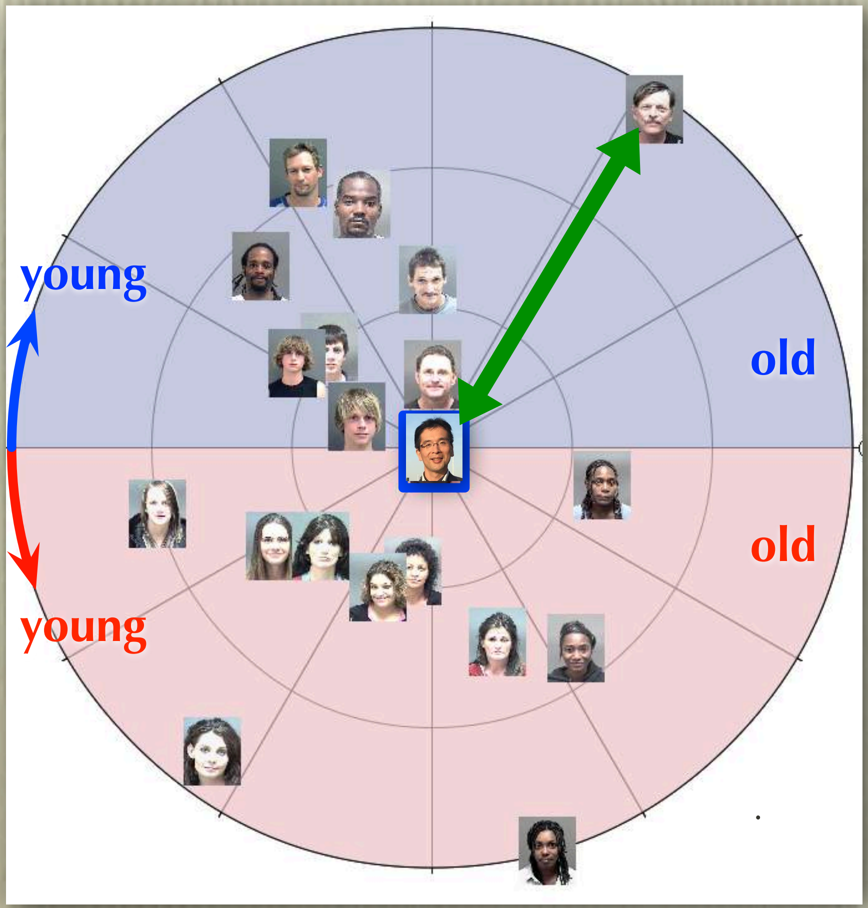
training $\{ T_i \}$	testing $\{ X_i \}$	
		← speaker-open
T_1 - T_2	X_1 - T_1	
T_1 - T_3	X_1 - T_2	← speaker-pair-open
T_4 - T_7	X_1 - T_3	
T_5 - T_9	X_2 - T_8	
:	:	

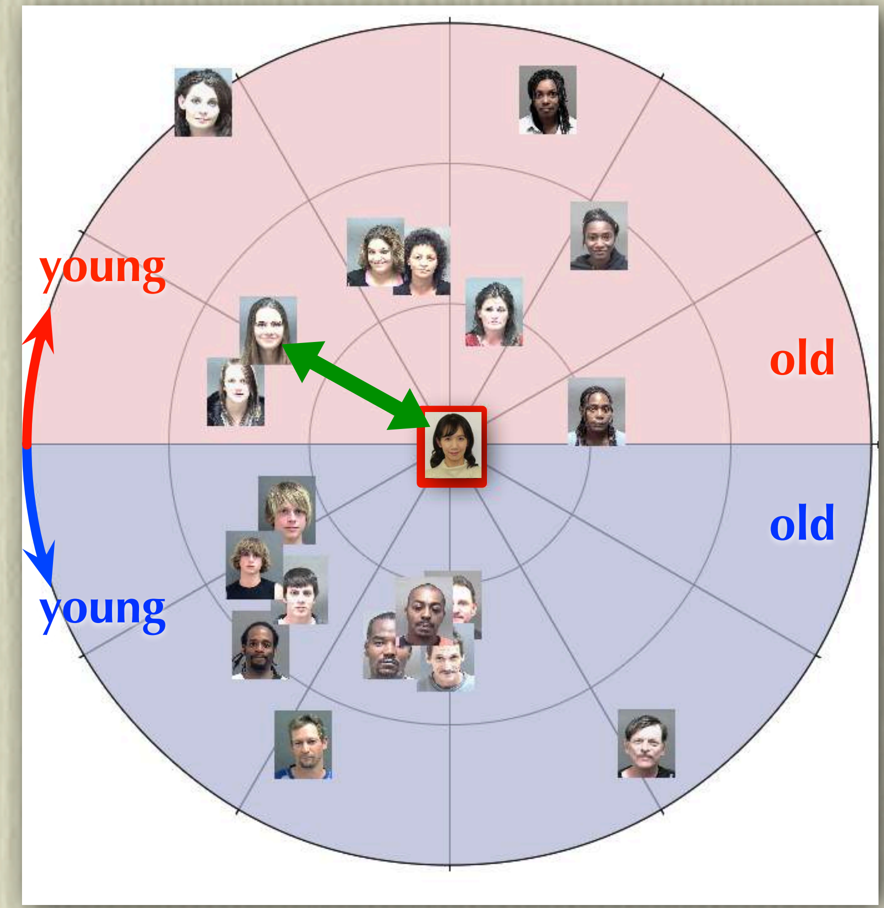
Speakers are shared only partially.
Speakers pairs are *not* shared.

A possible application[Kawase+'14]

Accent-based browser of WE from "your" viewpoint

- Your pronunciation is placed at the origin.
- Accent distance is represented as geometric distance from you.
- Gender and age is also shown in the visualization.





Menu of the last four lectures

- **Robust processing of easily changeable stimuli**
  - Robust processing of general sensory stimuli
  - Any difference in the processing between humans and animals?
- **Human development of spoken language**
  - Infants' vocal imitation of their parents' utterances
  - What acoustic aspect of the parents' voices do they imitate?
- **Speaker-invariant holistic pattern in an utterance**
  - Completely transform-invariant features -- *f*-divergence --
  - Implementation of word Gestalt as relative timbre perception
  - Application of speech structure to robust speech processing
- **Radical but interesting discussion**
  - A hypothesis on the origin and emergence of language
  - What is the definition of "human-like" robots?