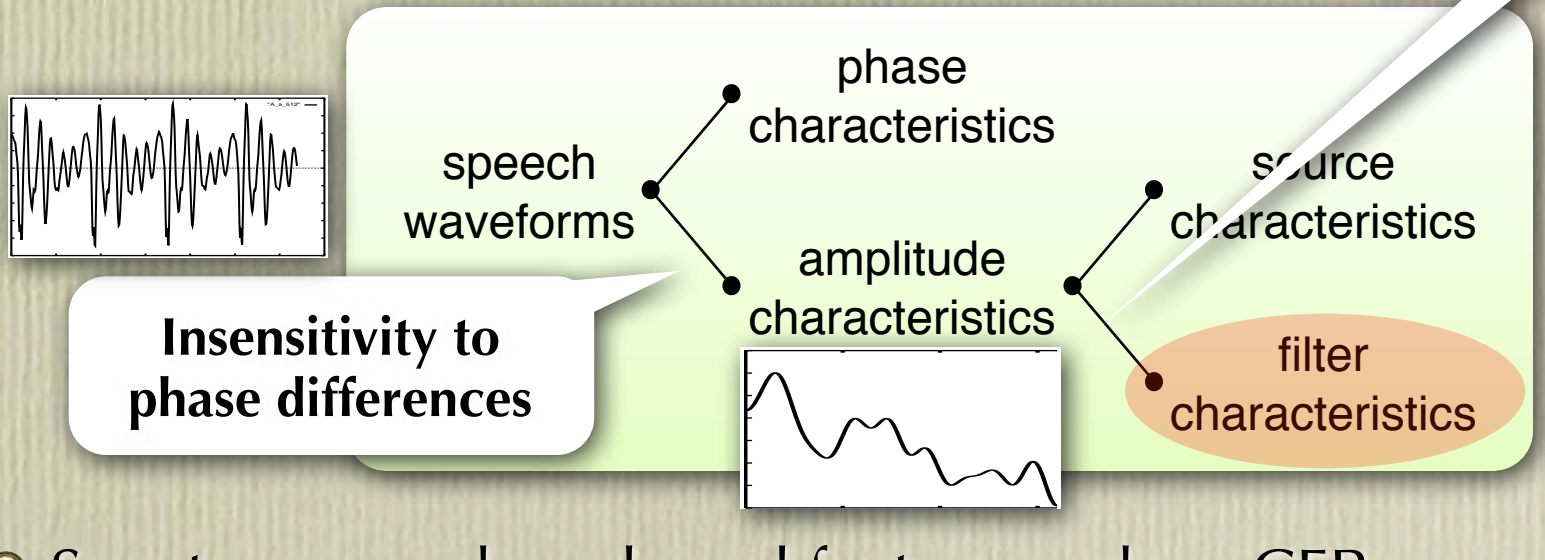
Cognitive Media Processing #11

Nobuaki Minematsu

Feature separation to find specific info.

De facto standard acoustic analysis of s



$$o \rightarrow o_w, o_s$$

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

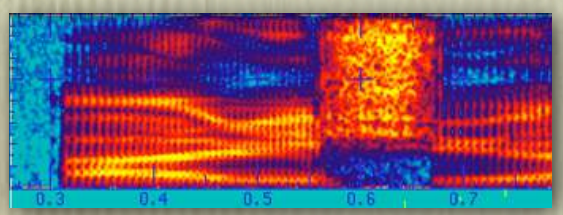
A claim found in classical linguistics

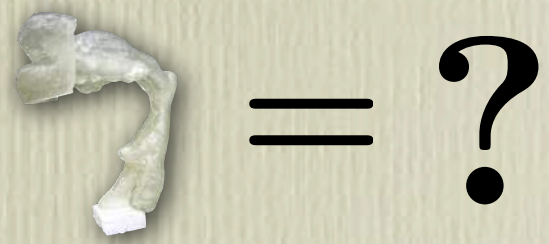
Theory of relational invariance [Jakobson+'79]

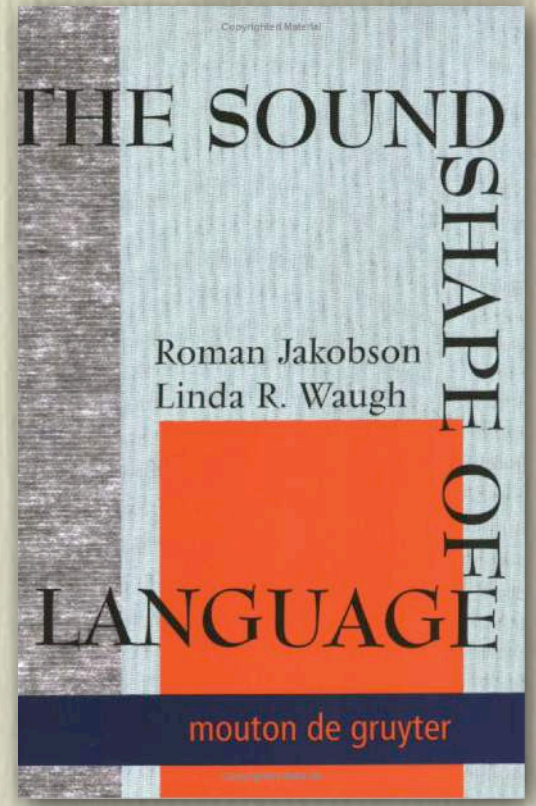
- Also known as theory of distinctive features
- Proposed by R. Jakobson

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

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



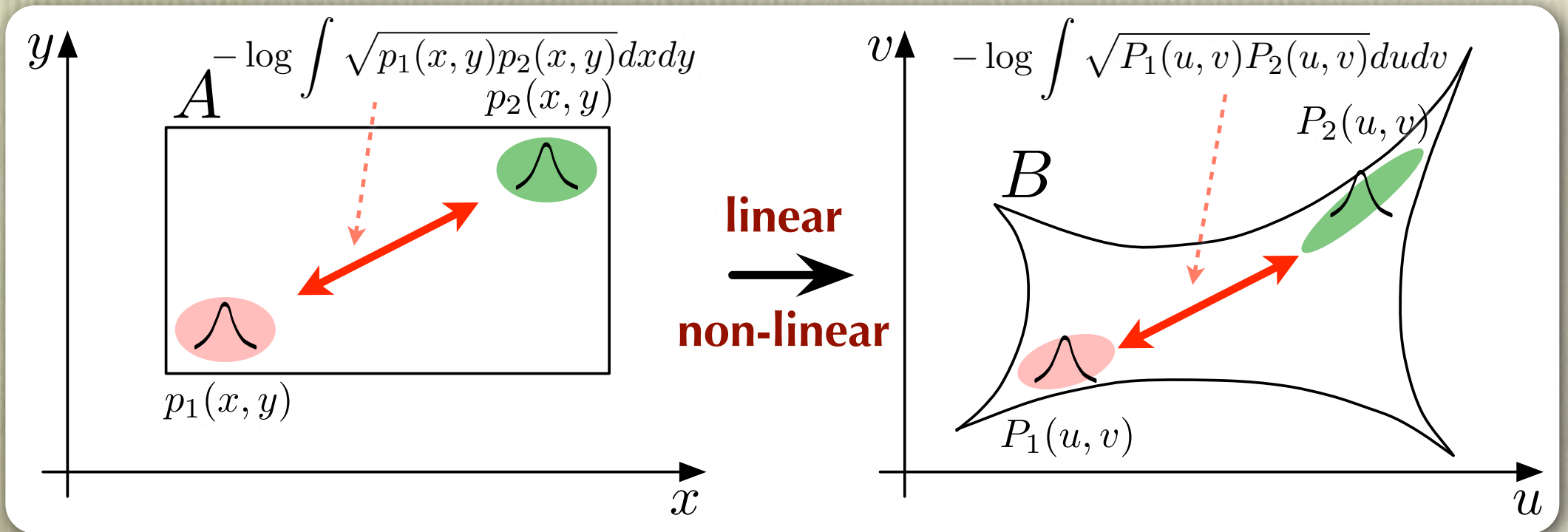




Complete transform-invariance

Complete invariance between two spaces

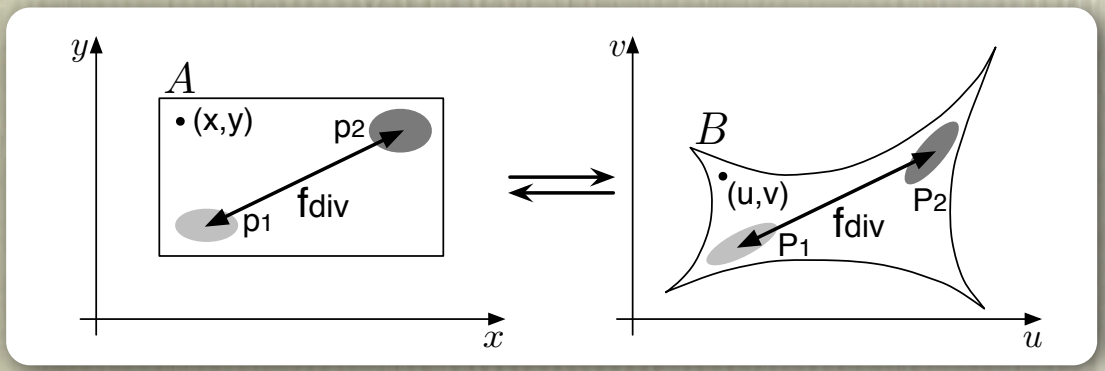
- An assumption
  - The transform is convertible and differentiable anywhere.
- An event in a space should be represented as distribution.
  - Event p in space A is transformed into event P in space B
  - p and P are physically different (/a/ of speaker A and /a/ of speaker B)



Complete transform-invariance

Any general expression for invariance? [Qiao'10]

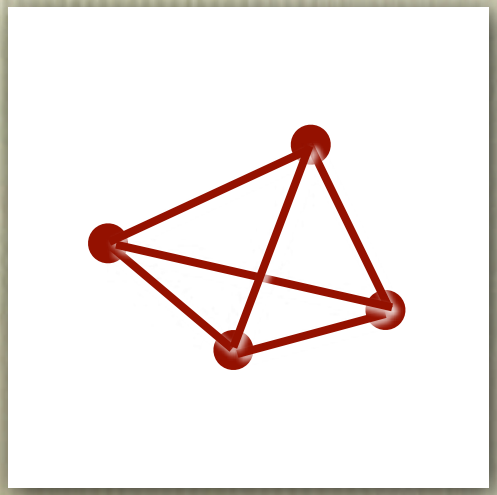
- BD is just one example of invariant contrasts.
- f-divergence is invariant with any kind of transformation.
  - $f_{div}(p_1, p_2) = \int p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}$
  - $g(t) = t\log(t) \rightarrow f_{div} = \mathrm{KL-div.} \qquad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

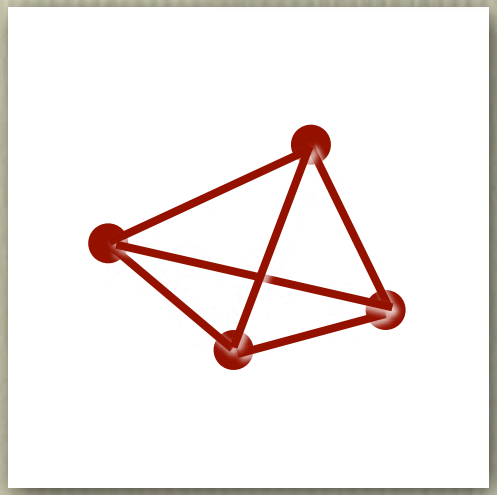


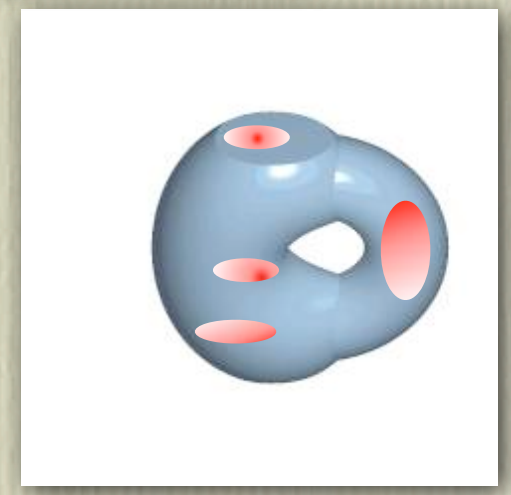
Invariance in variability

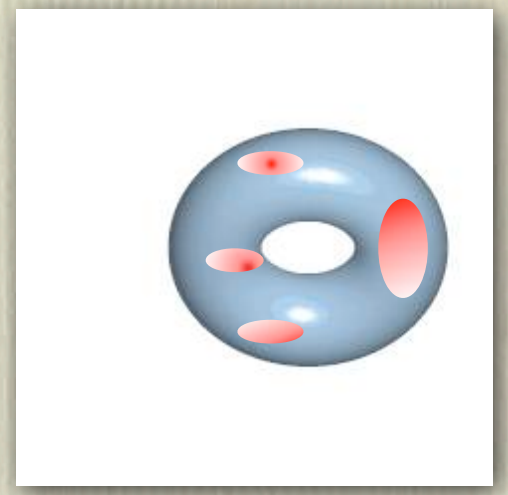
Topological invariance [Minematsu'09]

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



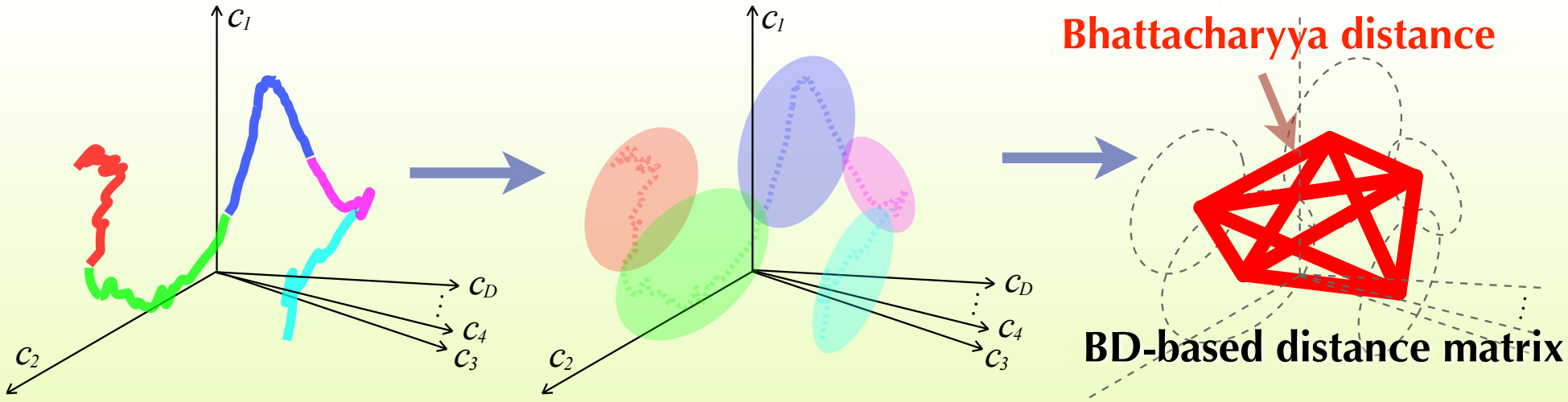


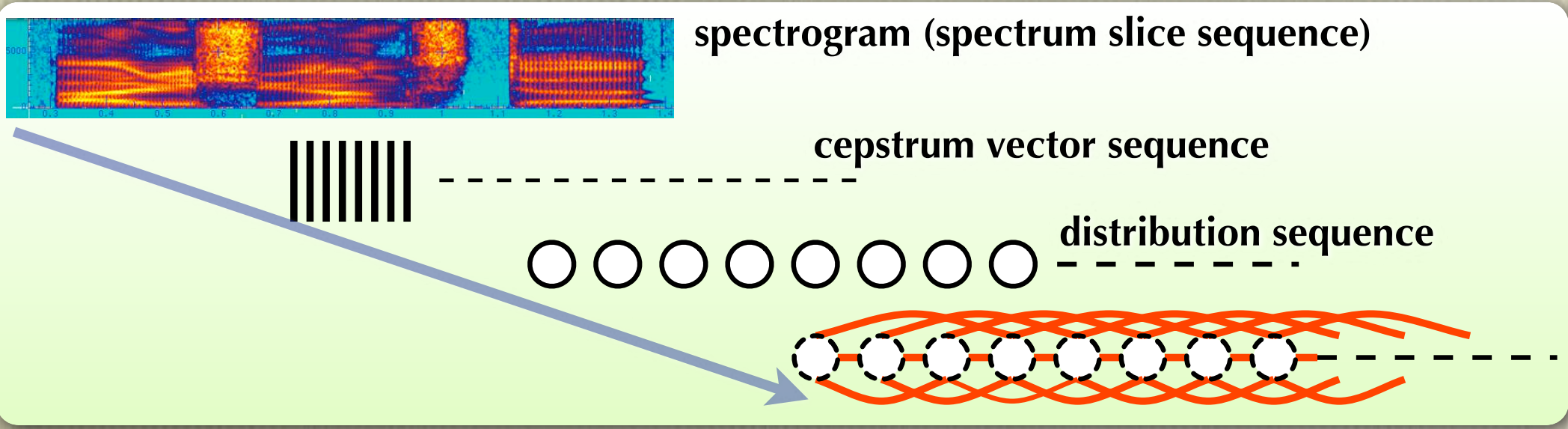




Invariant speech structure

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





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

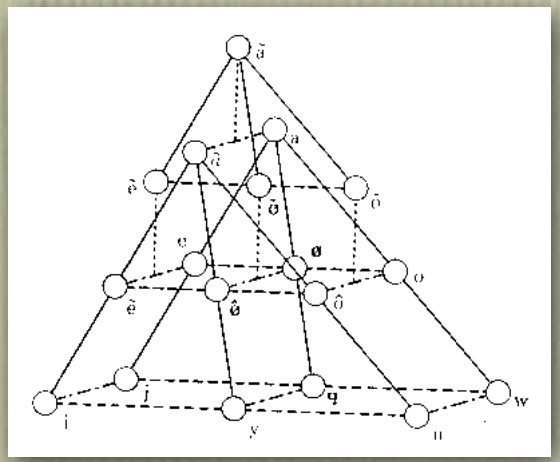
A claim found in classical linguistics

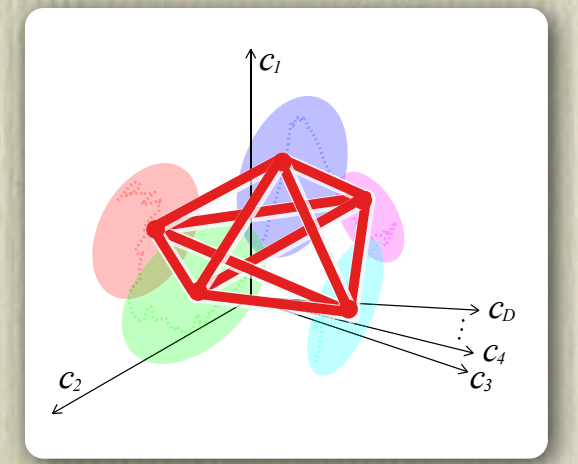
Theory of relational invariance [Jakobson+'79]

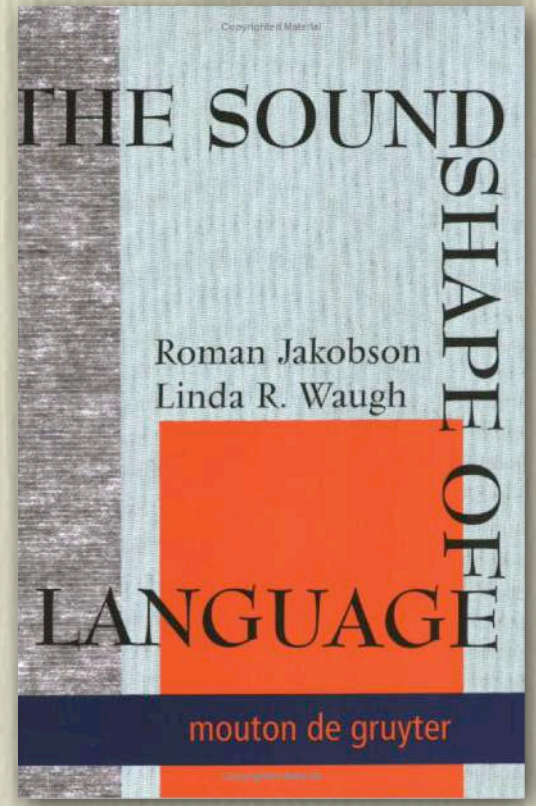
- Also known as theory of distinctive feature
- Proposed by R. Jakobson

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.



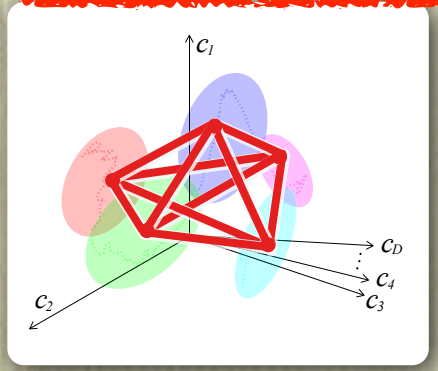




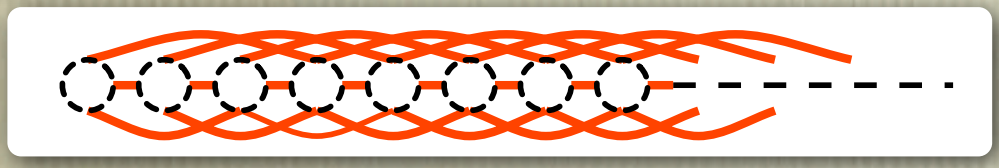
More classical claims in linguistics

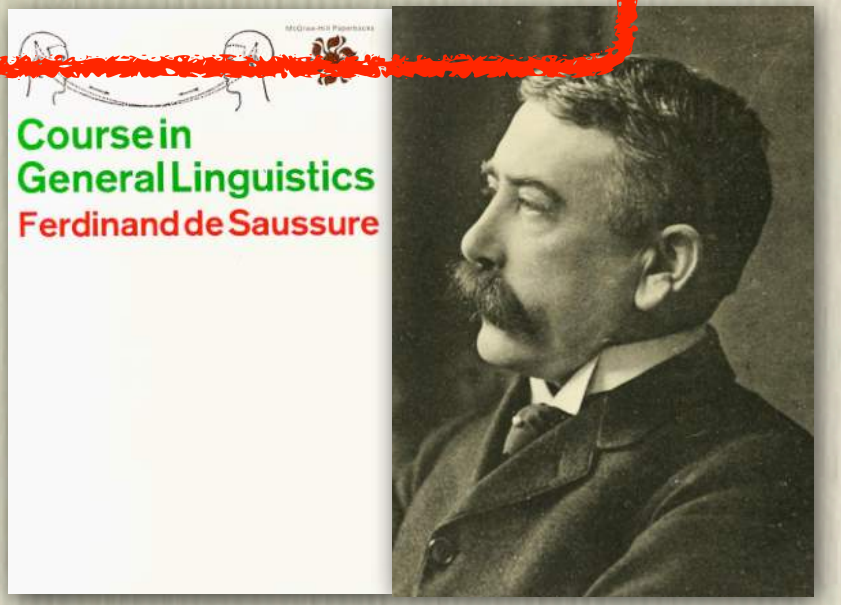
Ferdinand de Saussure (1857-1913)

- Father of modern linguistics
- "Course in General Linguistics" (1916)
- What defines a linguistic element, conceptual or phonic, is the relation in which it stands to the other elements in the linguistic system.
- The important thing in the word is not the sound alone but the phonic differences that make it possible to distinguish this word from the others.
- Language is a system of only conceptual differences and phonic differences.



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

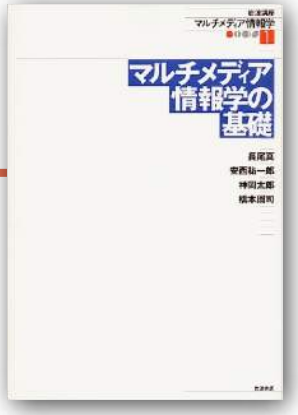




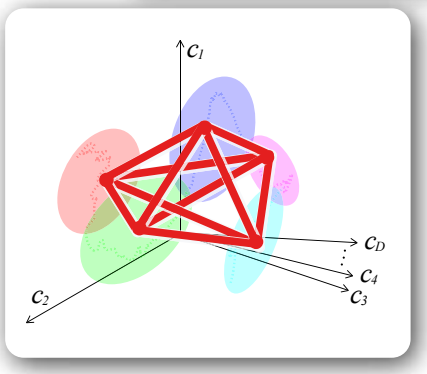
A new framework for “human-like” speech machines #3

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

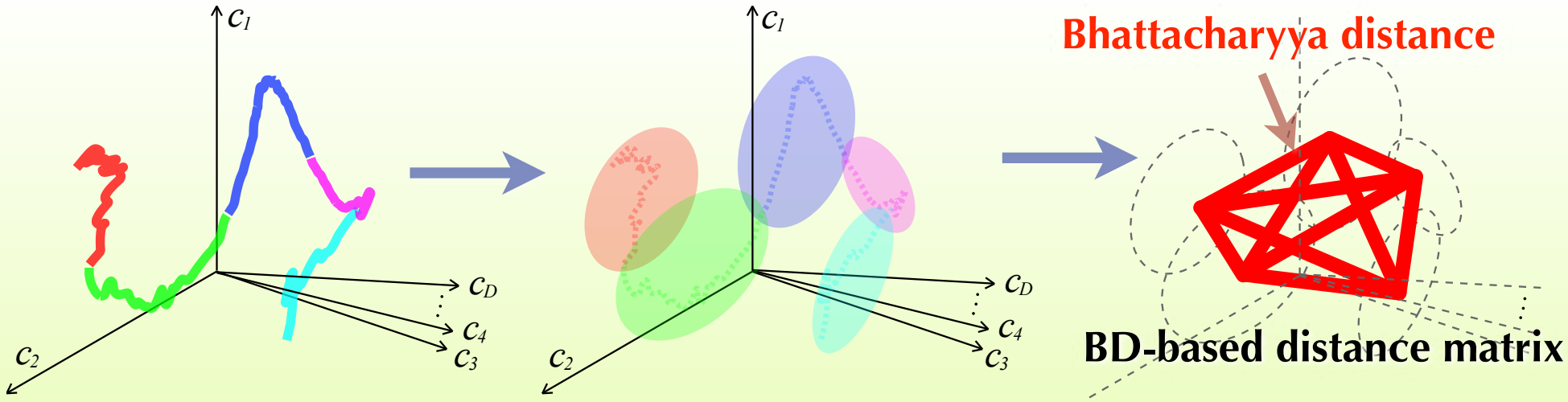
- Completely transform-invariant features -- *f*-divergence --
- Implementation of word Gestalt as relative timbre perception
- Application of speech structure to robust speech processing

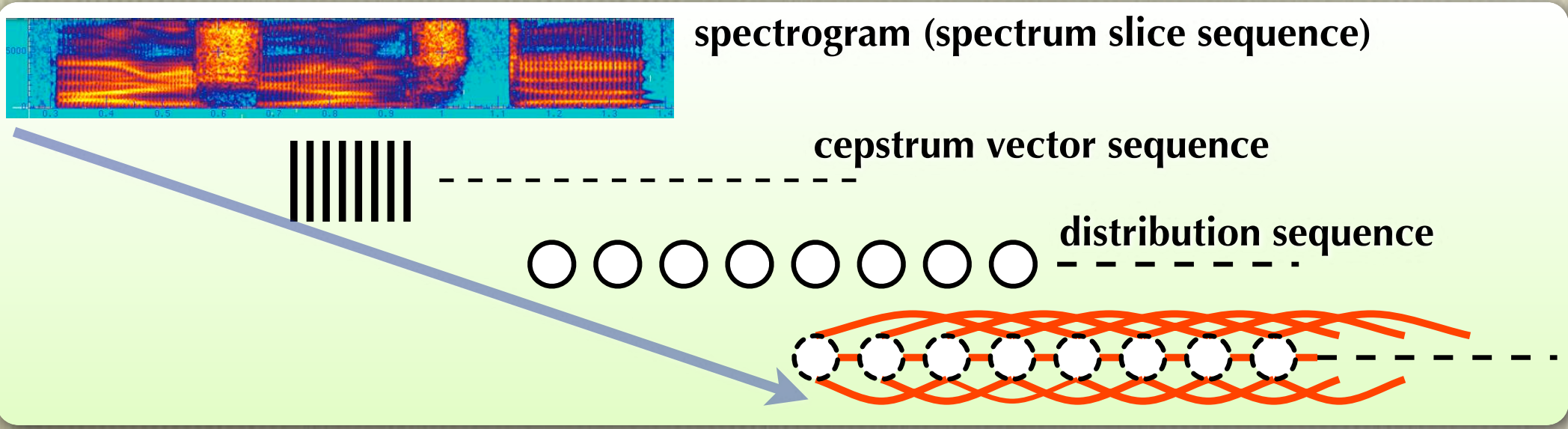
Radical but interesting discussion

- A hypothesis on the origin and emergence of language
- What is your definition of "human-like" machines?

Invariant speech structure

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

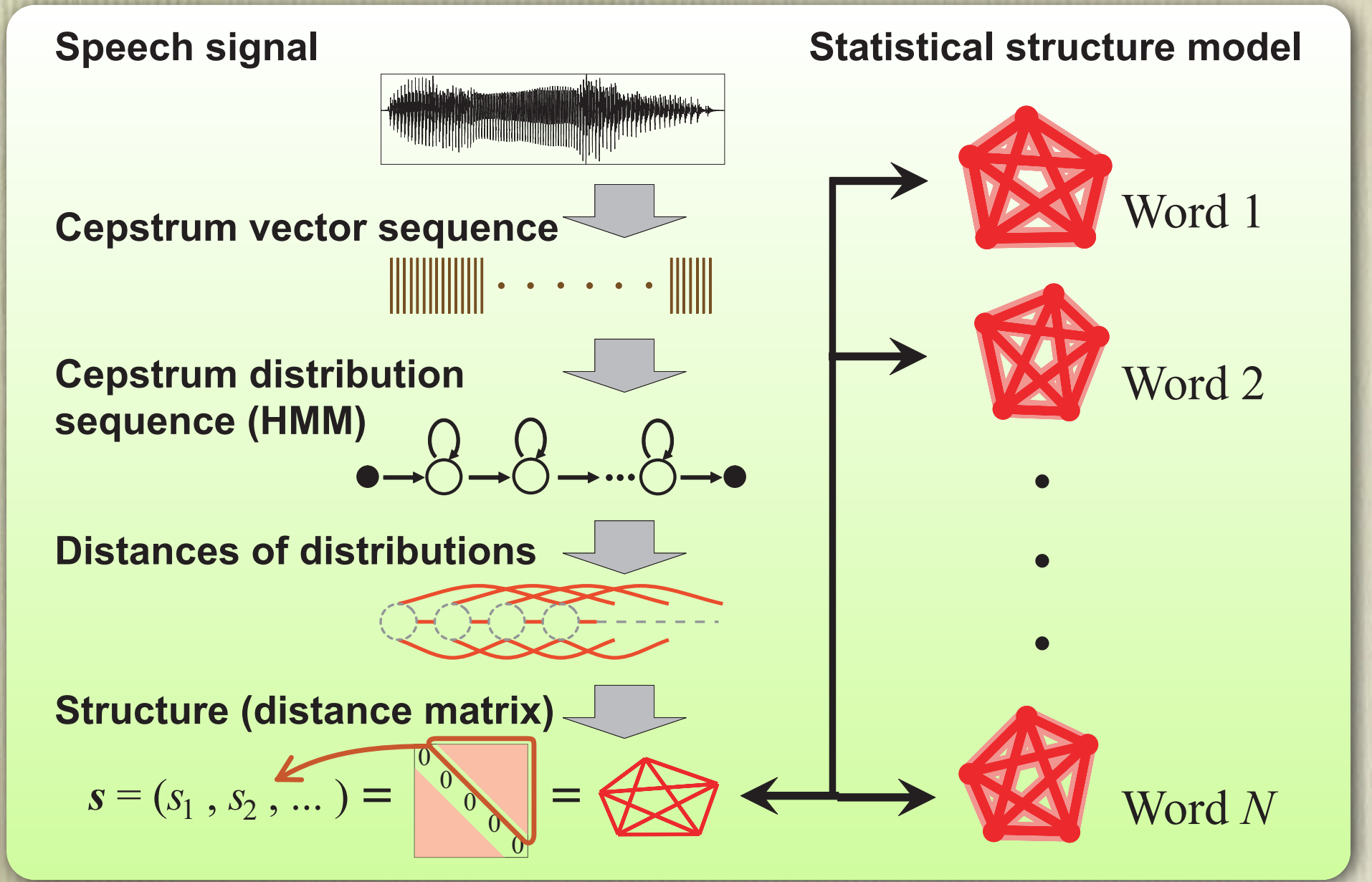




- An event (distribution) has to be much smaller than a phoneme.

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 ($\boldsymbol{c'} = \boldsymbol{Ac}$)

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

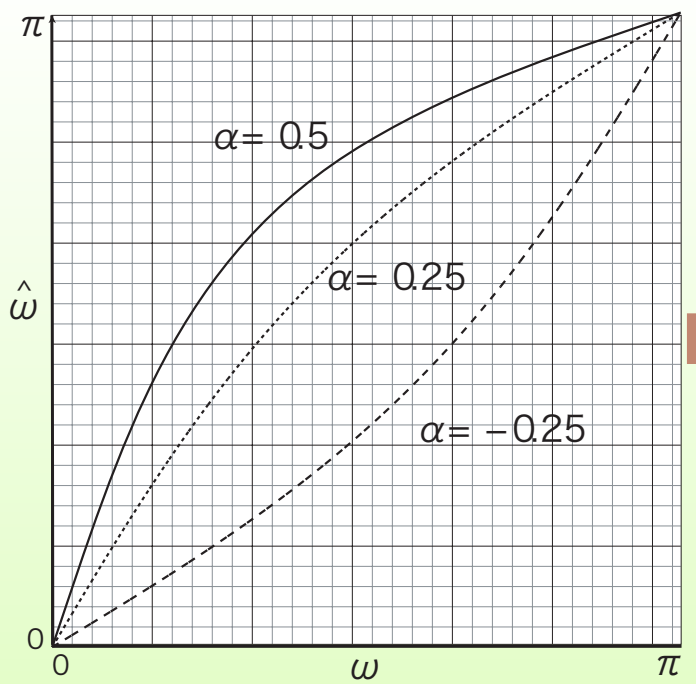
$$\begin{pmatrix} \boldsymbol{c'_{1,n}} \\ \boldsymbol{c'_{n+1,N}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{c_{1,n}} \\ \boldsymbol{c_{n+1,N}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_{1,n}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

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.



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

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

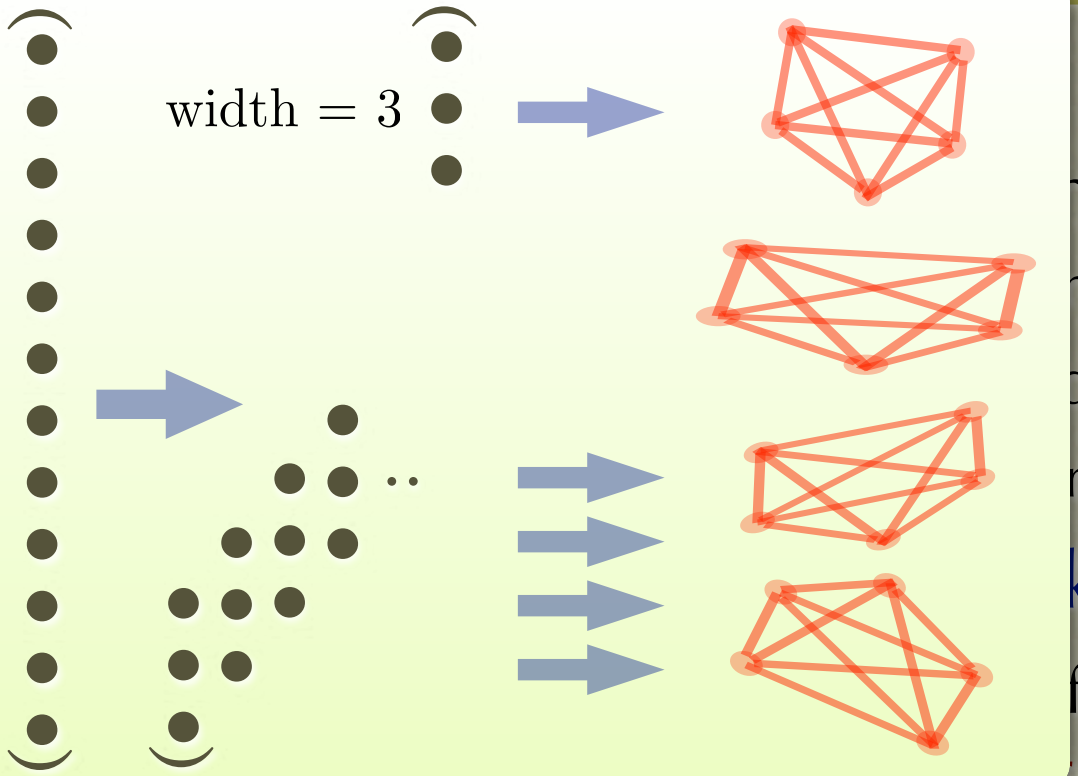
$$\boldsymbol{c} = (c_1\ c_2\ c_3\ c_4 \cdots)^t.$$

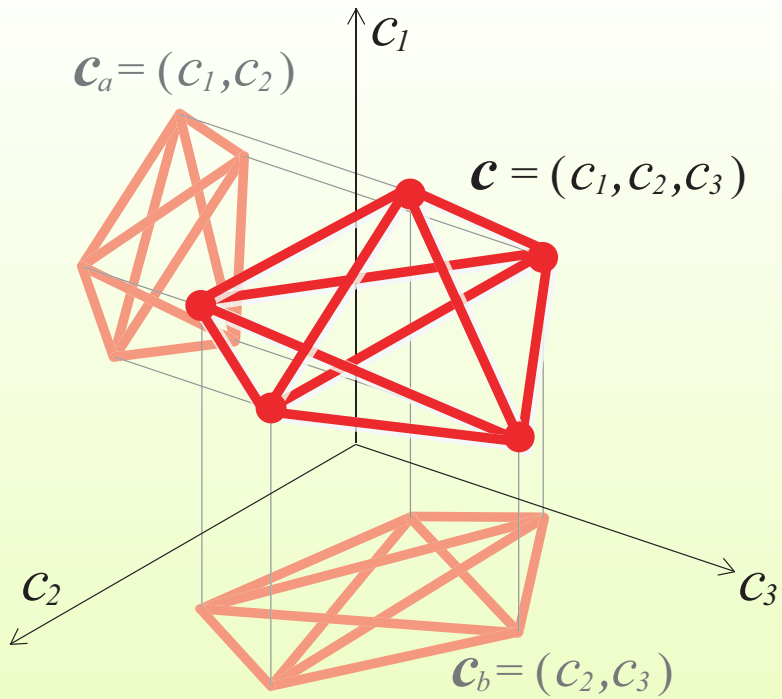
$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=\max(0,j-i)}^{j} \binom{j}{m} \times \frac{(m+i-1)!}{(m+i-j)!} (-1)^m \alpha^{(2m+i-j)}$$

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

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

Application of structures to ASR





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

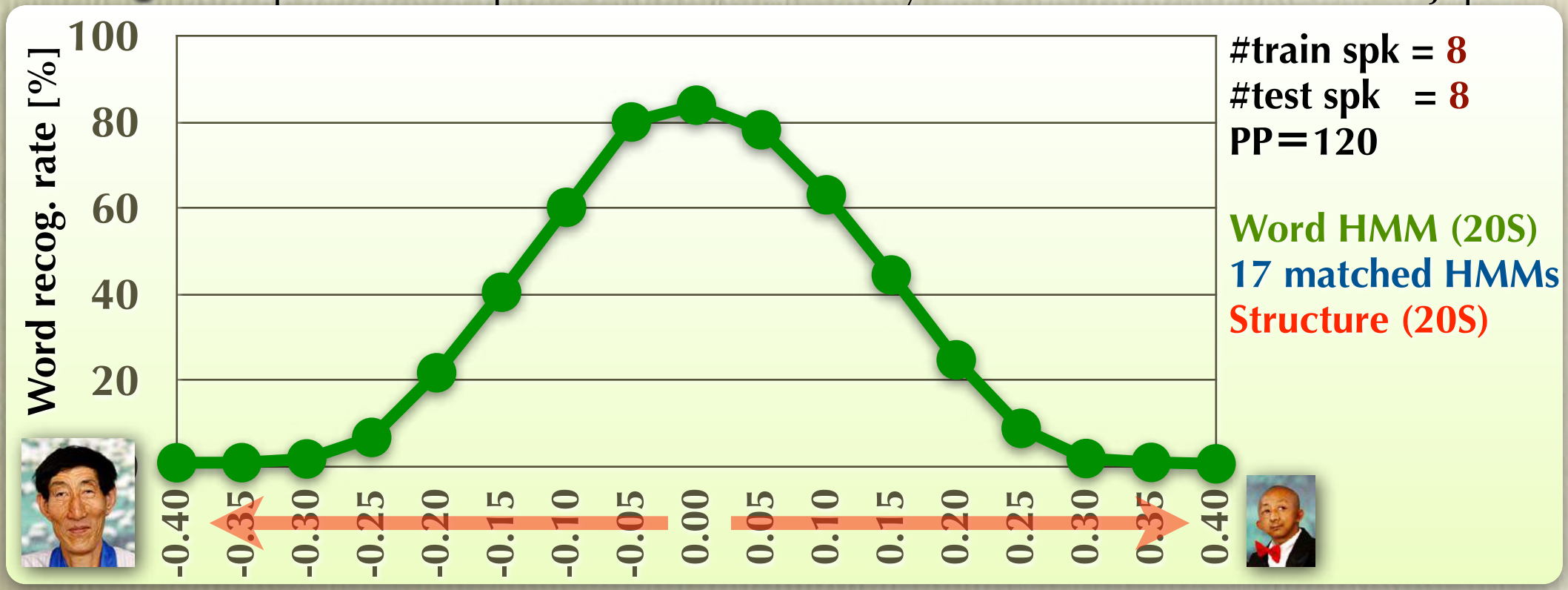
$$A = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots &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

$$\begin{pmatrix} c'_{1,n} \\ c'_{n+1,N} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} c_{1,n} \\ c_{n+1,N} \end{pmatrix} + \begin{pmatrix} b_{1,n} \\ b_{n+1,N} \end{pmatrix}$$

Application of structures to ASR

Isolated word recognition using warped utterances

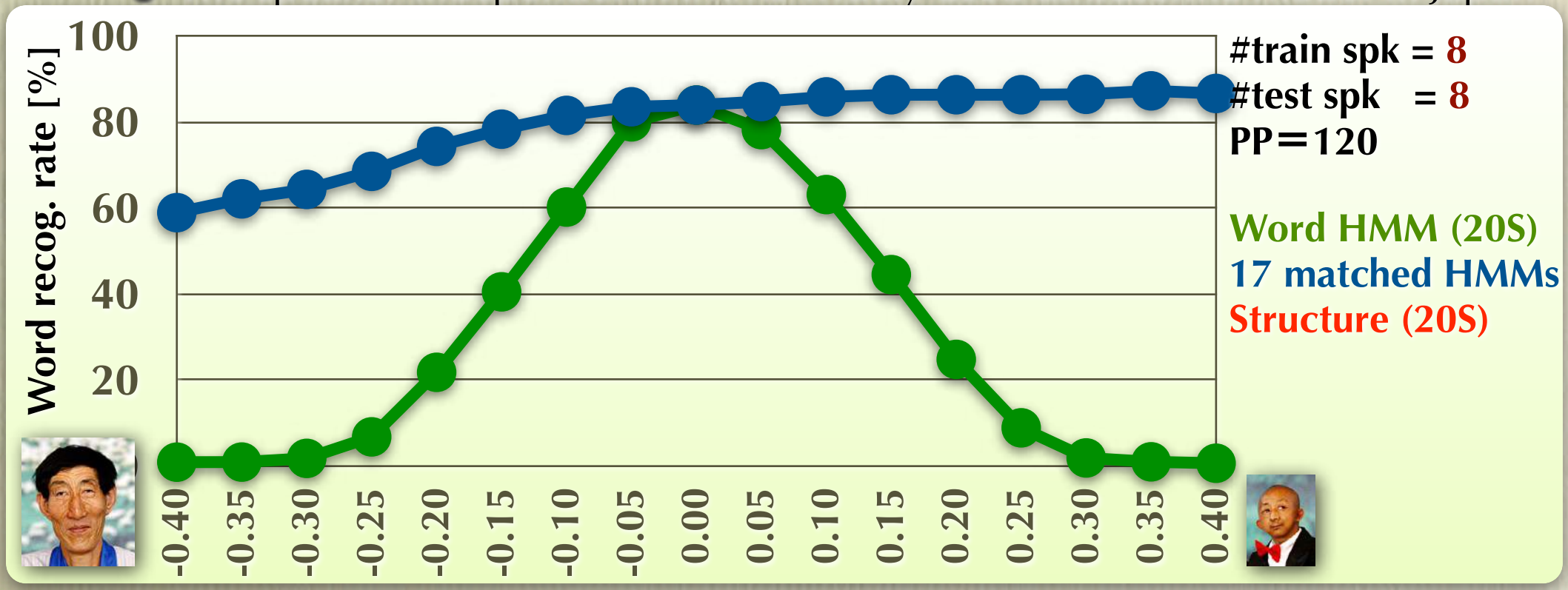
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used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



Application of structures to ASR

Isolated word recognition using warped utterances

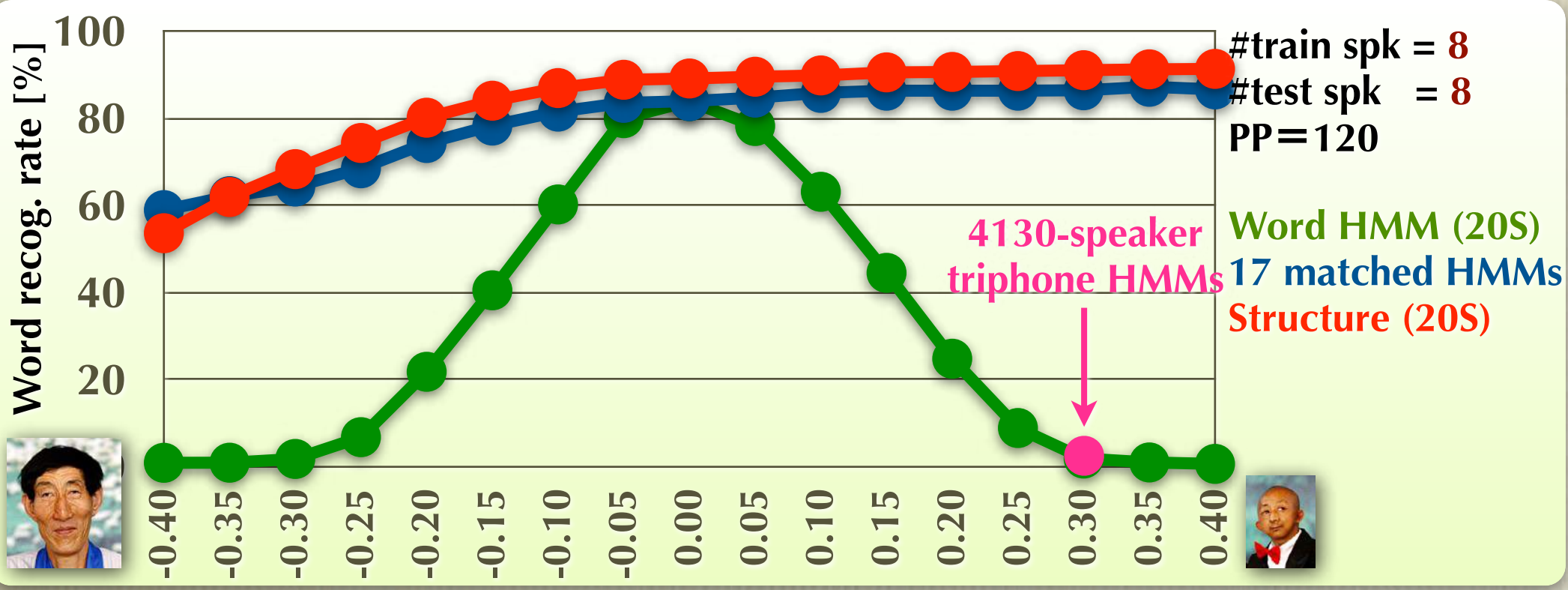
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used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



Application of structures to ASR

Isolated word recognition using warped utterances

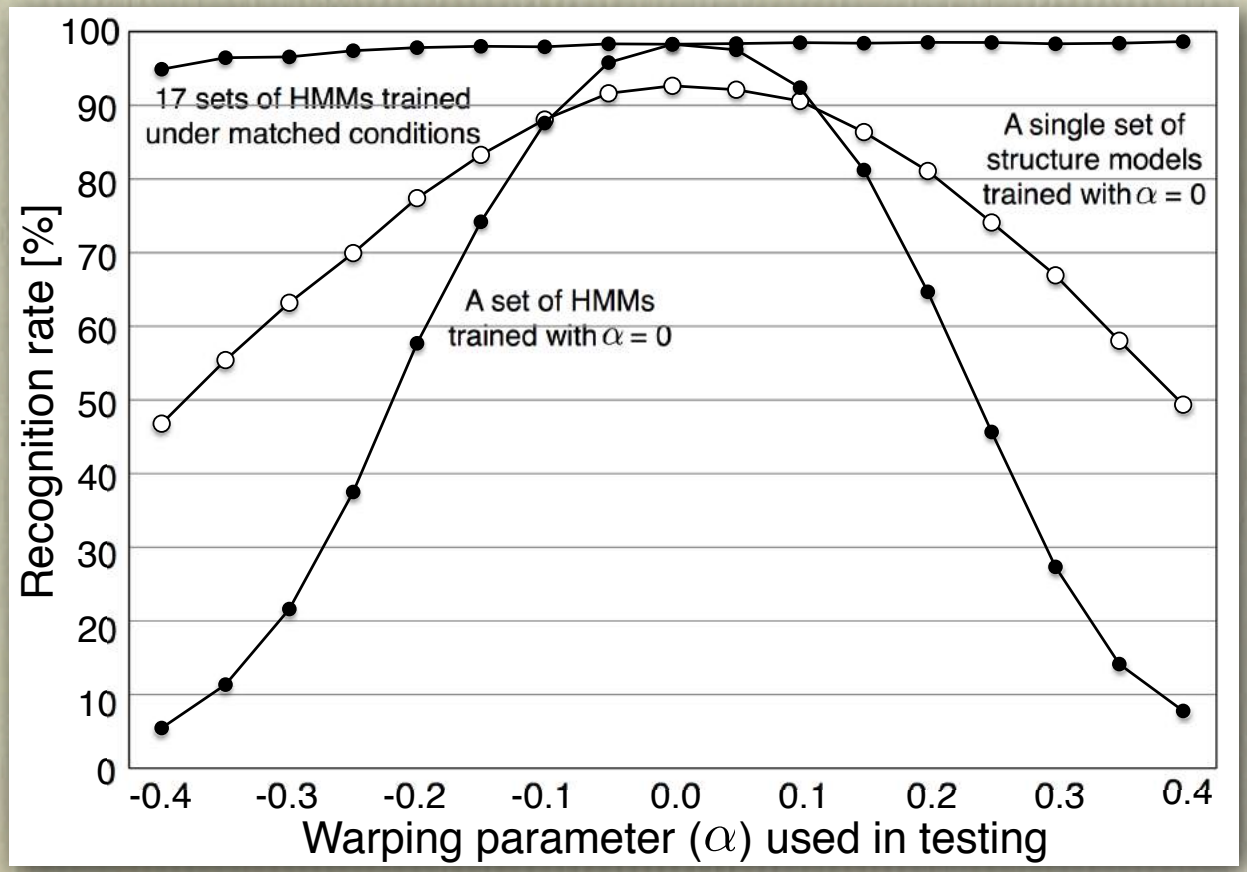
- Word = $V_1V_2V_3V_4V_5$ such as /eoaui/, PP = 120 (CL=0.8%)
- Word-based HMMs (20 states) vs. word-based structures (20 events)
  - Training = 4M+4F adults, testing = other 4M+4F with various VTLs
- 4,130-speaker triphone HMMs are also tested with 0.30.
  - The speaker-independent HMMs widely used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



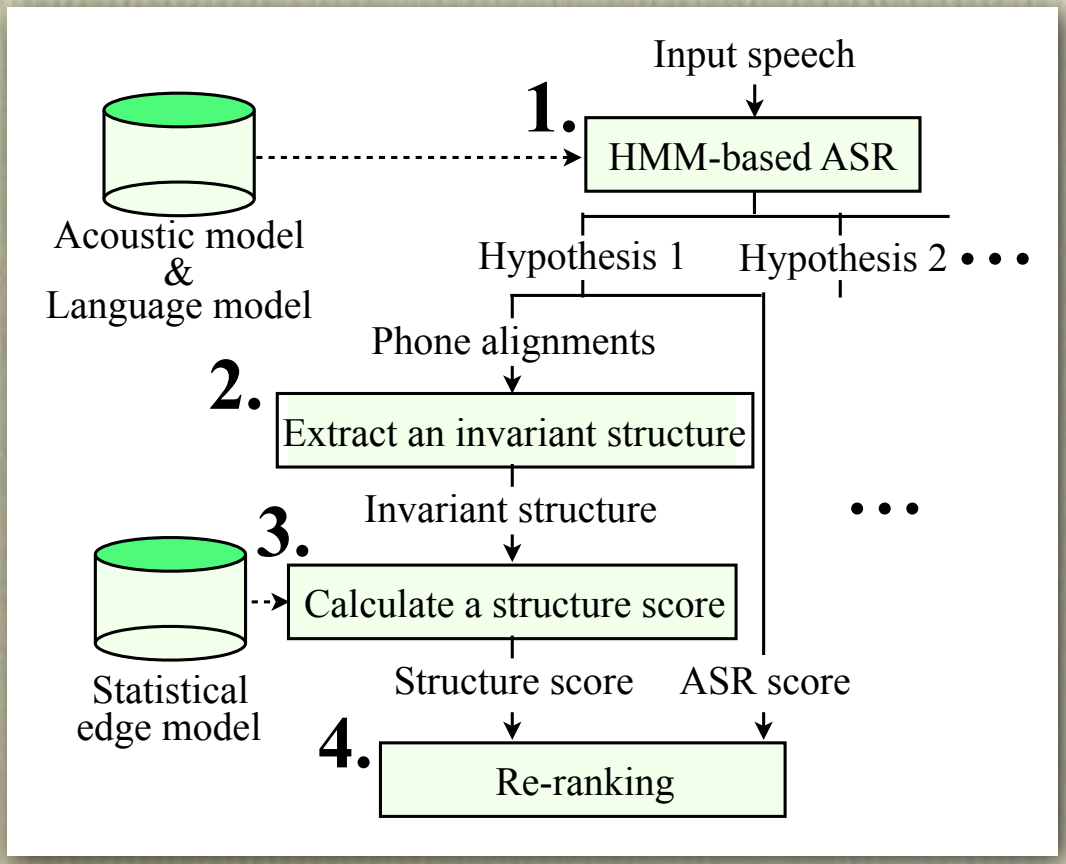
Application of structures to LVCSR

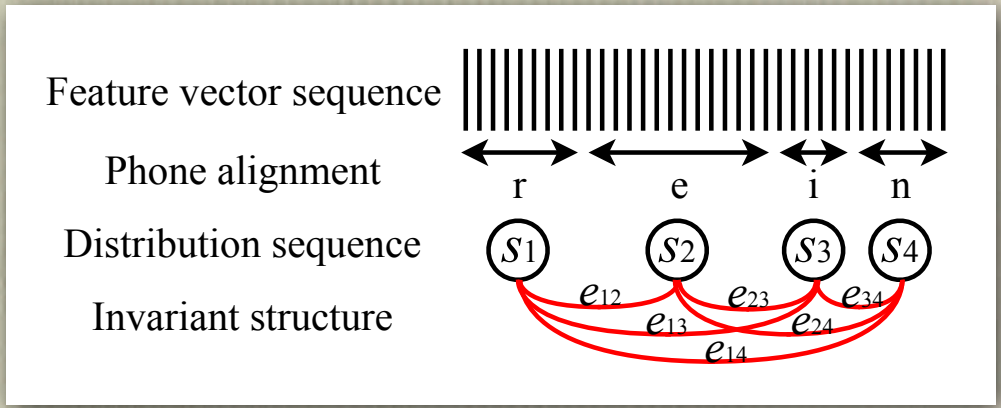
Application to more realistic ASR tasks [Suzuki+'15]

- Digits recognition and LVCSR (dictation)

Use of structural features in discriminative reranking

- Str. scores and ASR scores are combined with average perceptron.

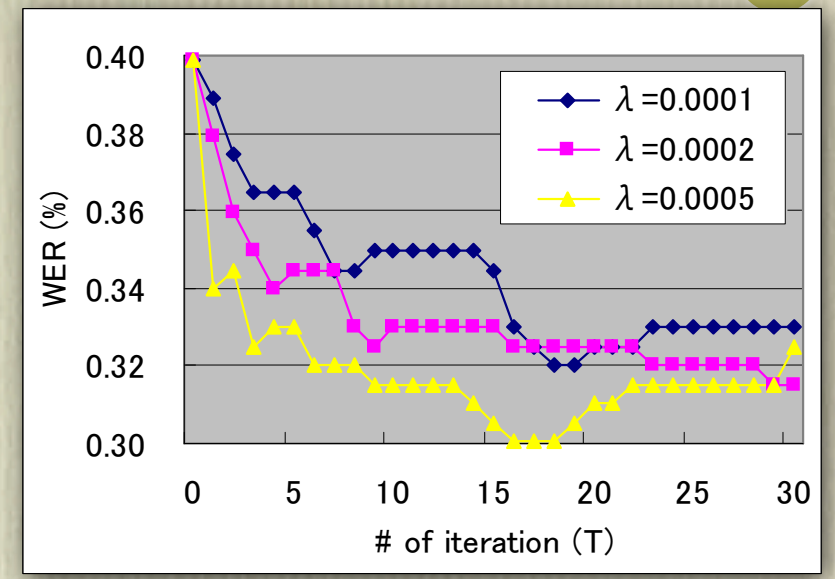




Application of structures to LVCSR

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%

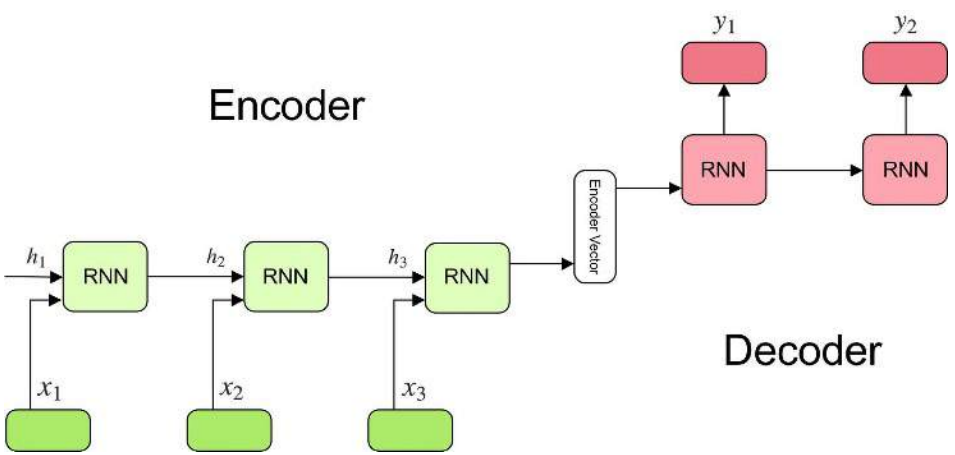
Many errors are due to a large number of homonyms in Japanese.

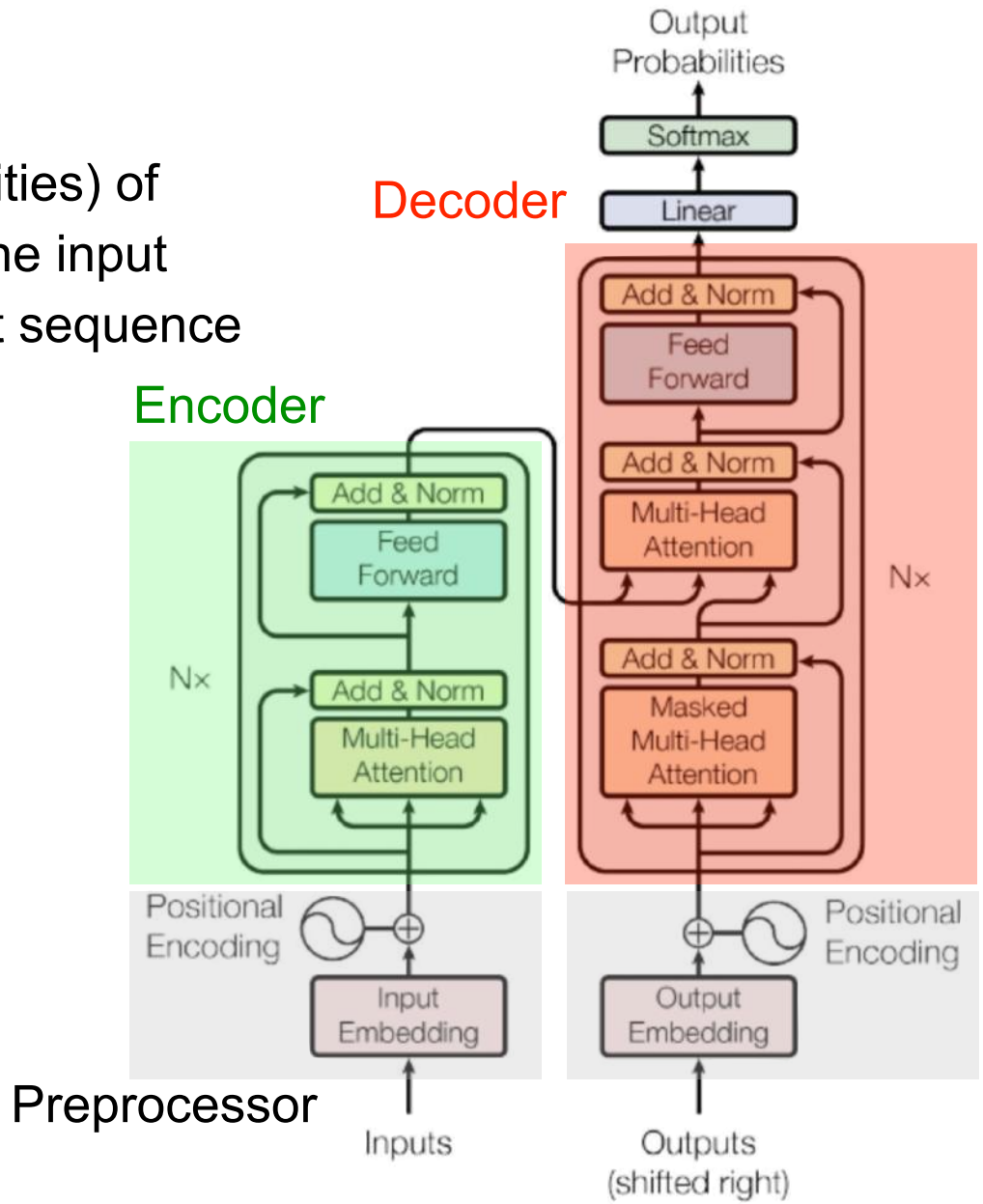
Table 6: CERs of the LVCSR experiment.

Baseline	Proposed	Relative improvement
2.67%	2.53%	5.24%

Transformer model

- **Attention is all you need !!**
  - https://arxiv.org/abs/1706.03762
  - Explicit modeling of the relations (similarities) of the current input token to other ones in the input sequence and to the tokens in the output sequence generated so far.
    - Self-attention mechanism

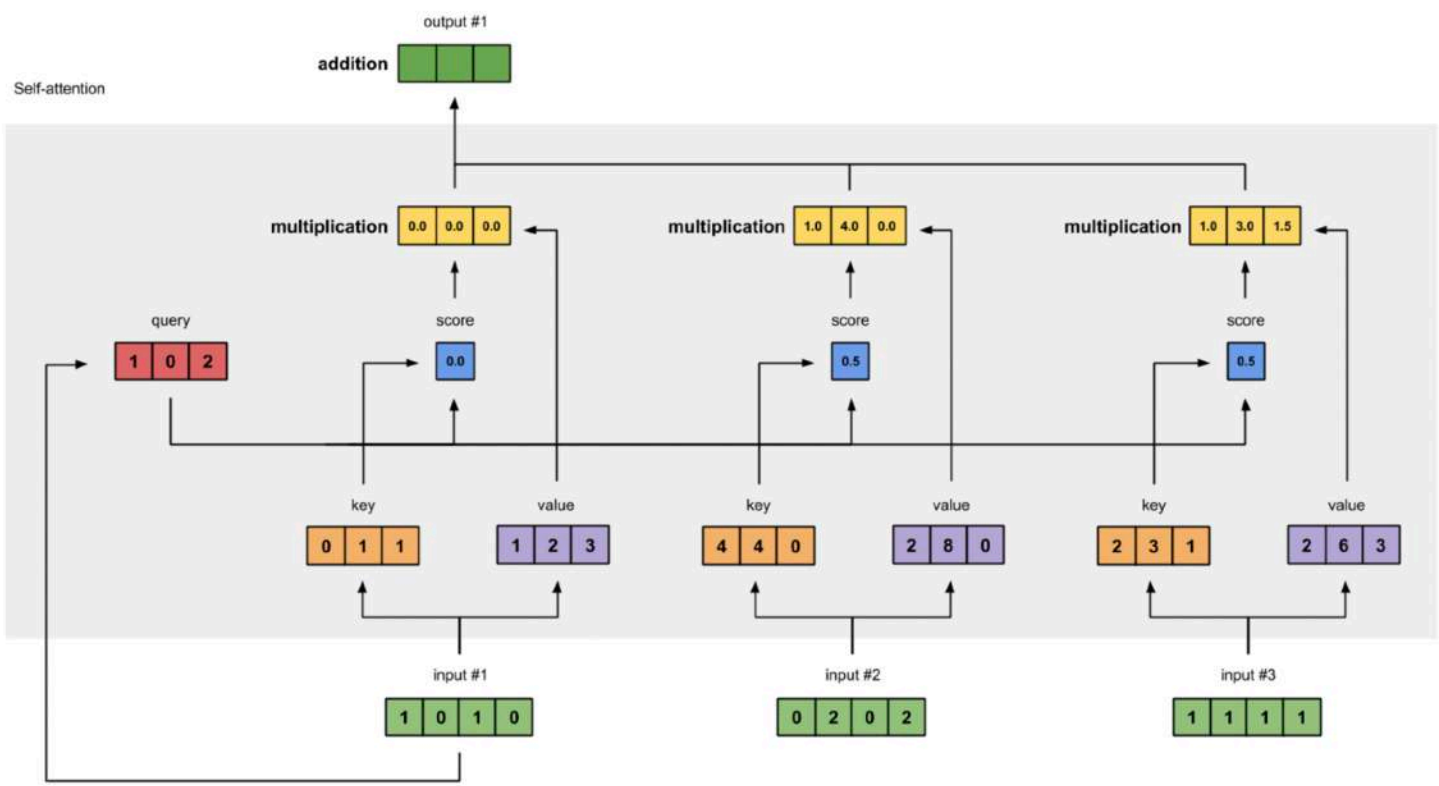




Self-attention mechanism

- Relatedness (similarity) of the current input token
  - to the other tokens in the input sequence and
  - to the tokens in the output sequence generated so far.
- A token is converted to its three components.
  - Value vector, key vector, and query vector.

$$\boldsymbol{E} \underset{(\vec{e_1}, \vec{e_2}, ..., \vec{e_V})}{} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \vec{e_2}$$

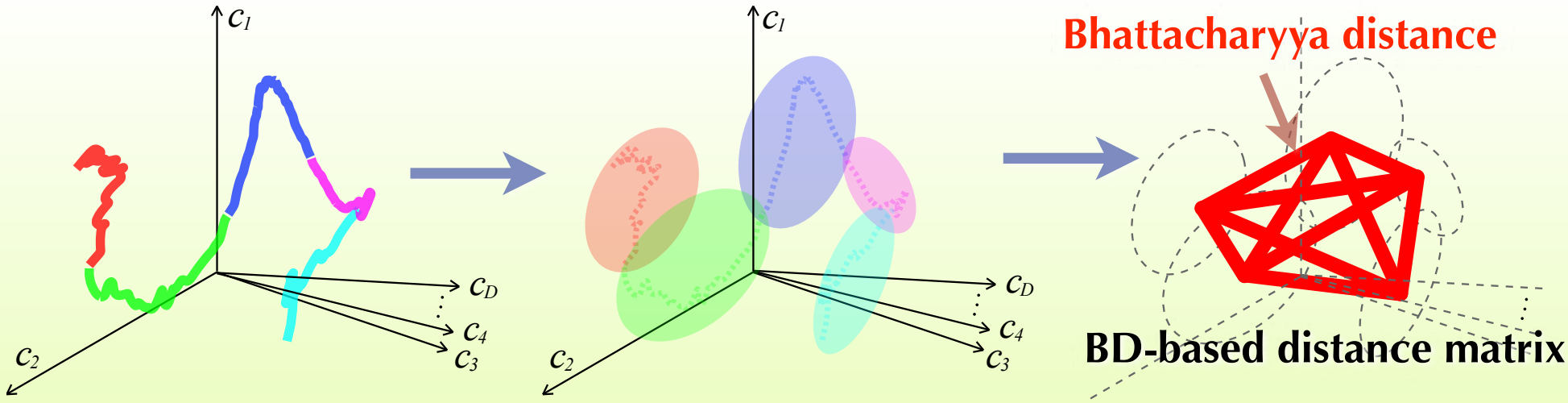


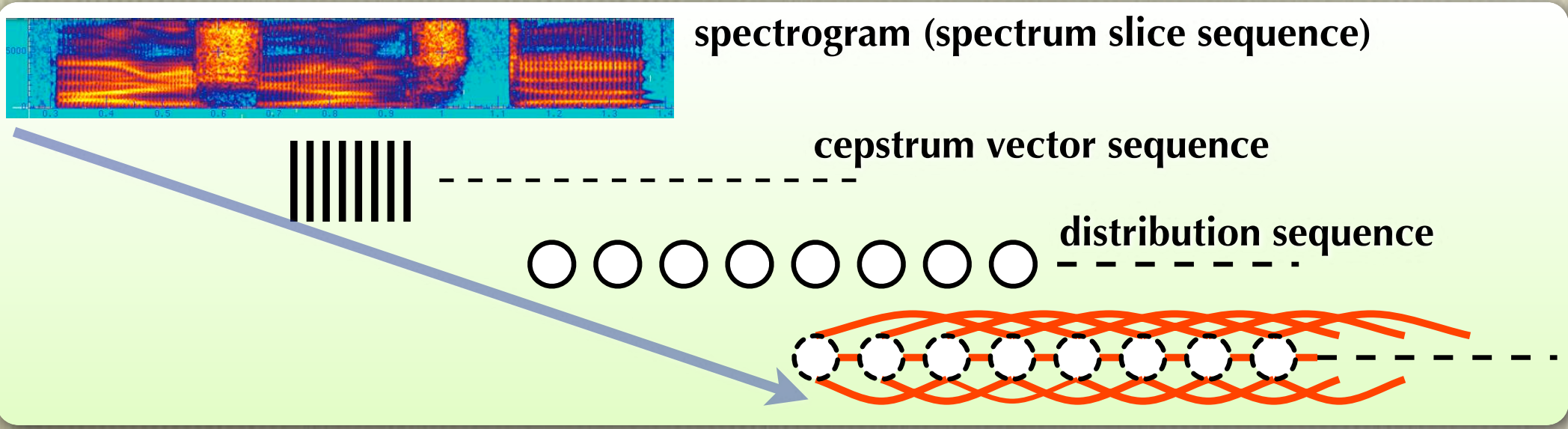
https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

In … re

Similarity ∝ 1/Distance
Distance ∝ 1/Similarity

Utterance to … div. [Minematsu'06]



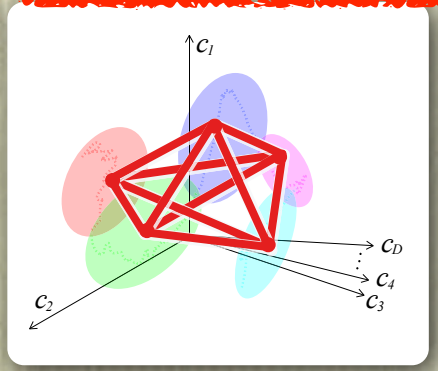


- An event (distribution) has to be much smaller than a phoneme.

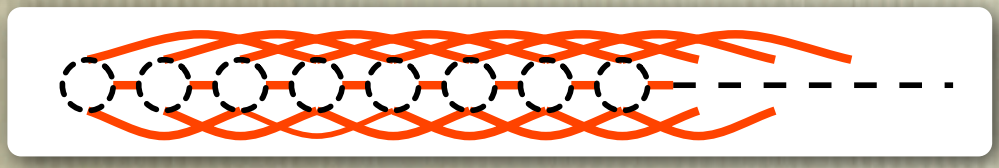
More classical claims in linguistics

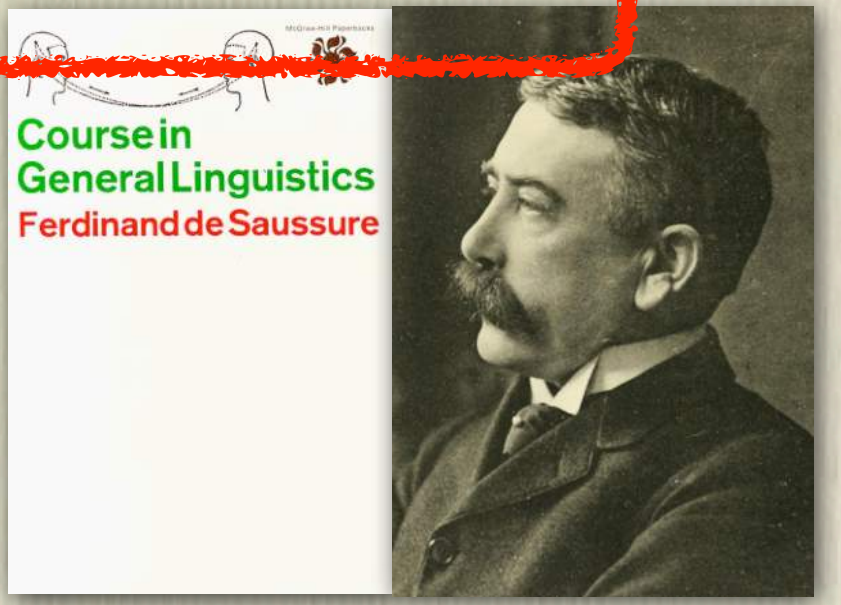
Ferdinand de Saussure (1857-1913)

- Father of modern linguistics
- "Course in General Linguistics" (1916)
- What defines a linguistic element, conceptual or phonic, is the relation in which it stands to the other elements in the linguistic system.
- The important thing in the word is not the sound alone but the phonic differences that make it possible to distinguish this word from the others.
- Language is a system of only conceptual differences and phonic differences.



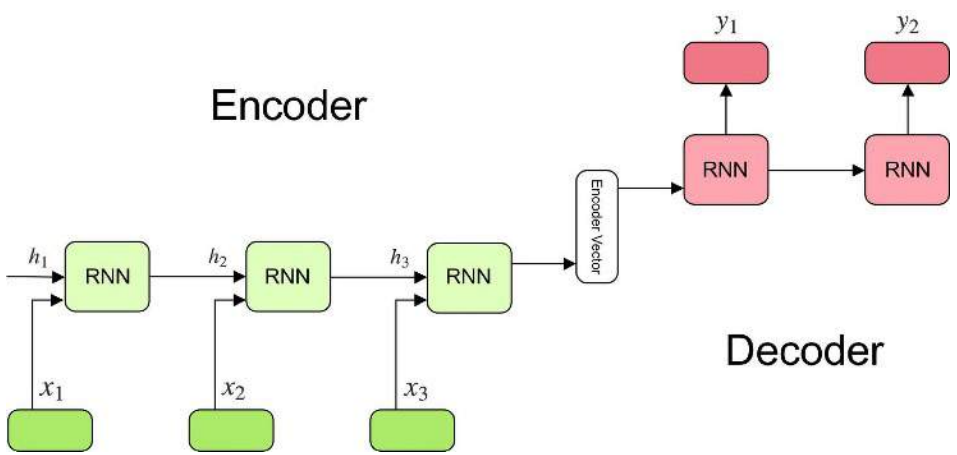
$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ d_{31} & & & \\ \vdots & & & \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix}$$

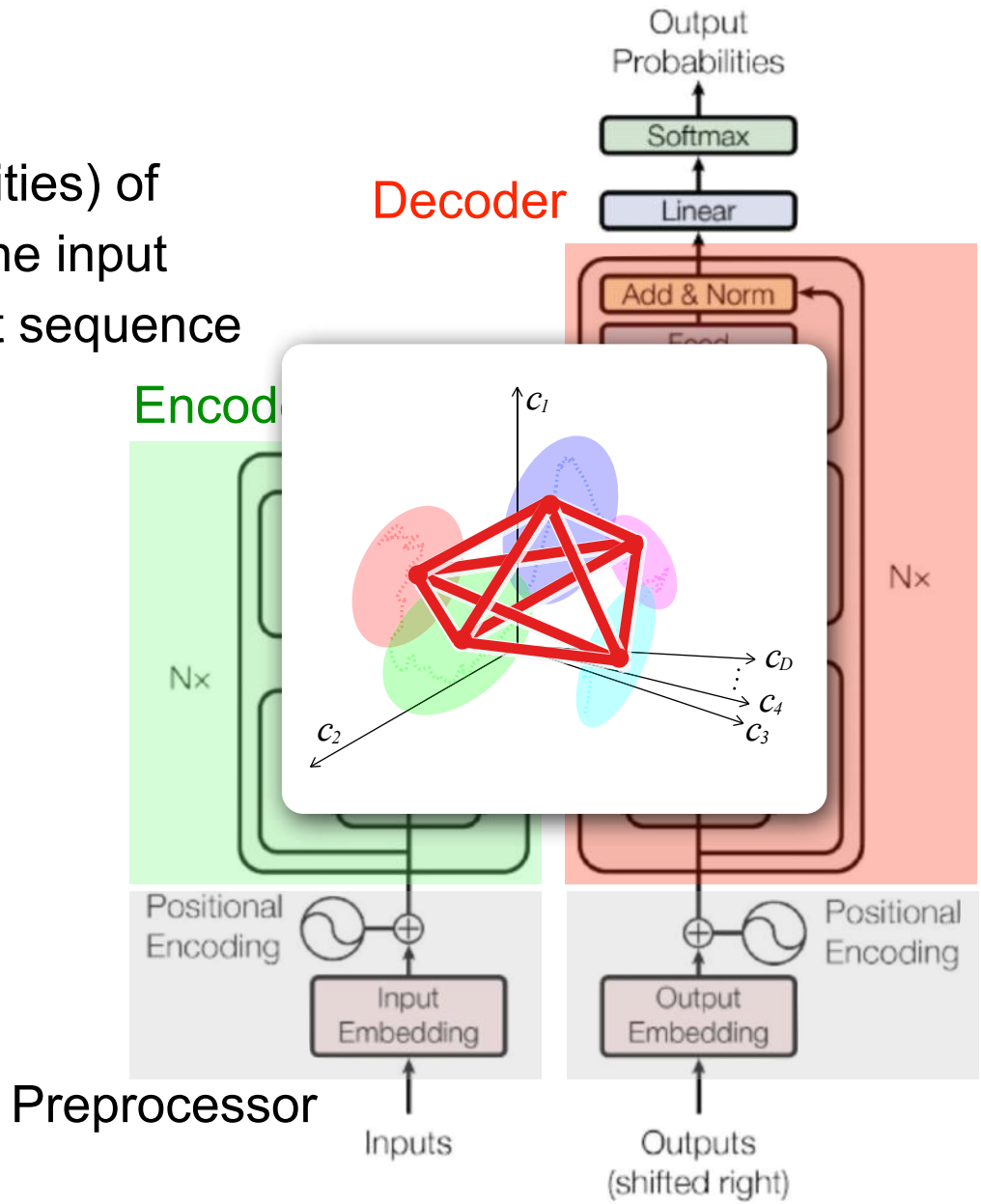




Transformer model

- **Attention is all you need !!**
  - https://arxiv.org/abs/1706.03762
  - Explicit modeling of the relations (similarities) of the current input token to other ones in the input sequence and to the tokens in the output sequence generated so far.
    - Self-attention mechanism

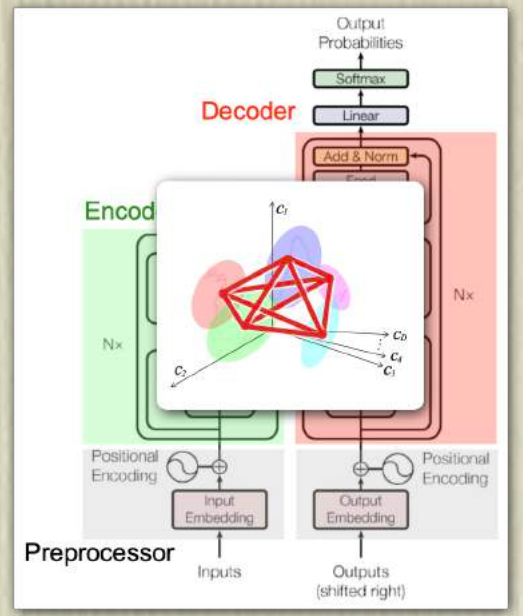


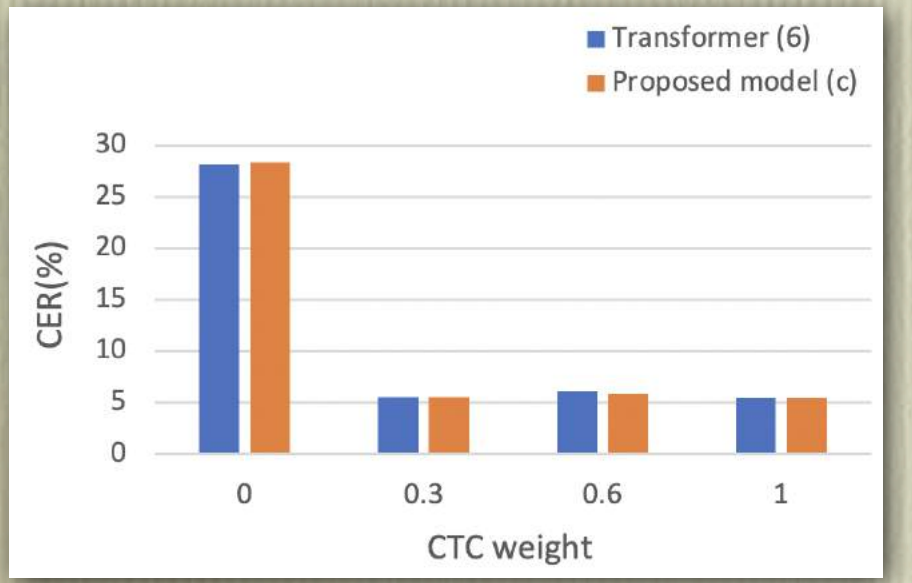


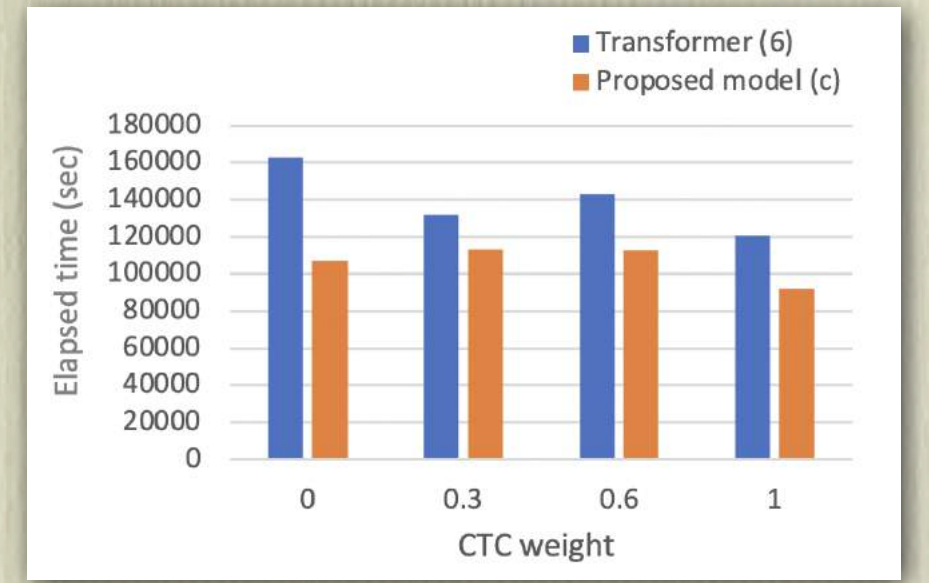
Structure-based transformer

Comparison between the two transformers [Wang+'22]

CTC weight	Transformer (6)			Proposed model (c)		
	CER-dev	CER-test	training time	CER-dev	CER-test	training time
0	28.2	29.5	163134	28.4	29.9	107156
0.3	5.5	6.2	132001	5.5	6.2	113071
0.6	6.1	6.6	143398	5.9	6.6	112970
1.0	5.4	5.9	120568	5.4	5.9	91890

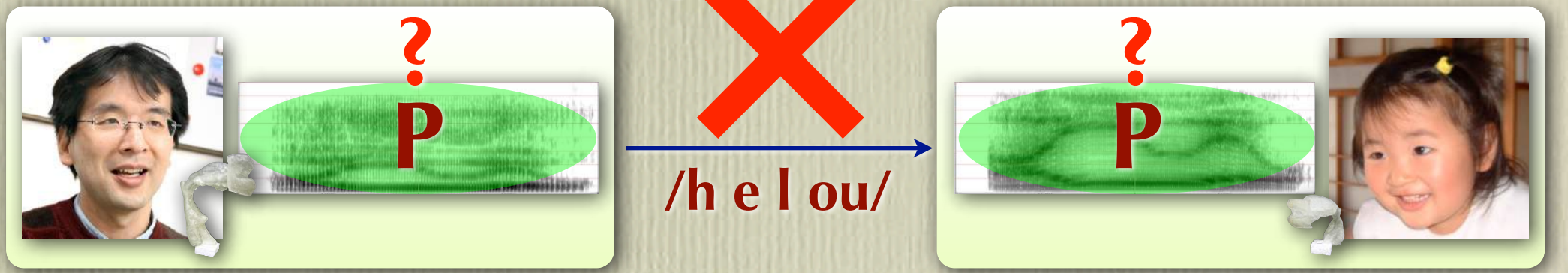






Language acquisition through vocal imitation

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]

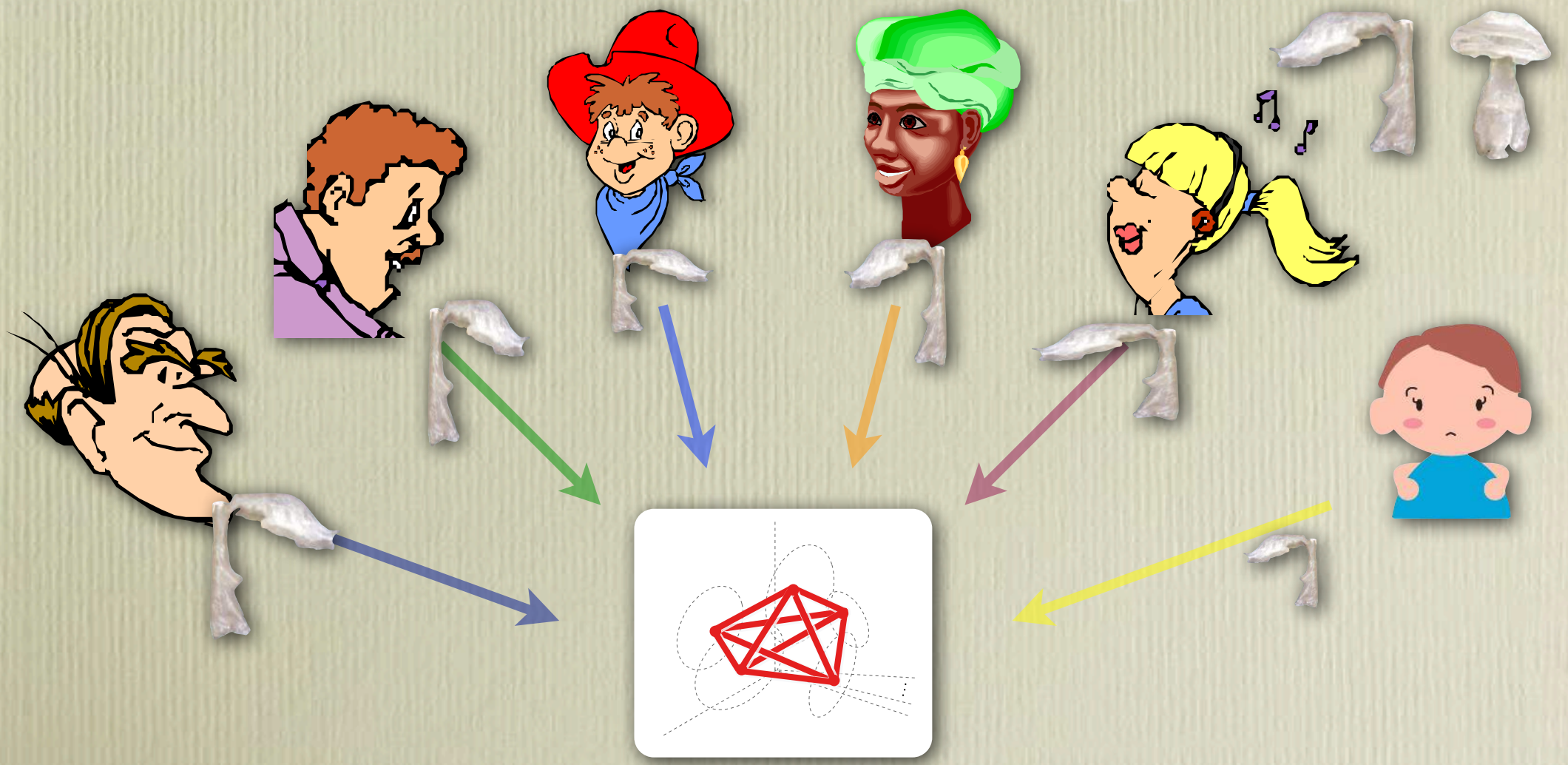
→ **No mathematical formulation**

- 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?

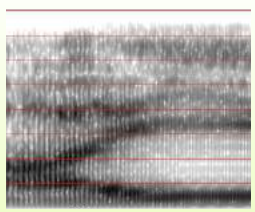
Structure-to-speech conversion

Speech representation with extra-ling. features removed

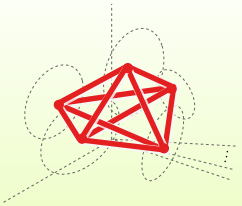
- Speaker-specific vocal tract features are removed.
- With them, we can identify speakers by hearing voices.

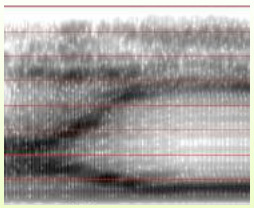


Structure-to-speech conversion



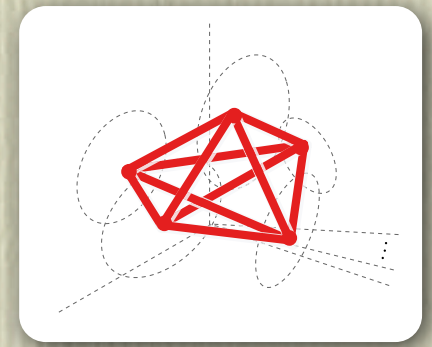
= Gestalt



=


/aiueo/

/aiueo/



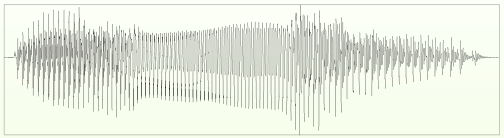
/aiueo/

Speech generation based on
infant-like vocal imitation

How to implement the vocal imitation?

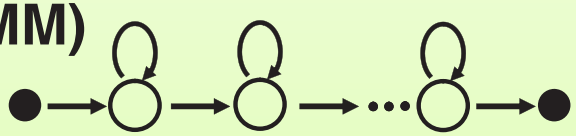
Extraction of a structure through training of an HMM

1. Speech waveforms



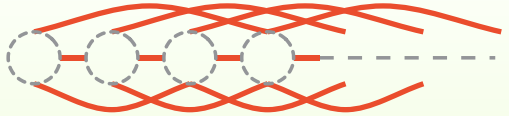
2. Cepstrum vector sequence

3. Cepstrum distribution sequence (HMM)



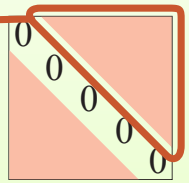
MAP estimation

4. Bhattacharyya distances

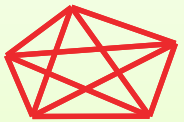


5. Structure (distance matrix)

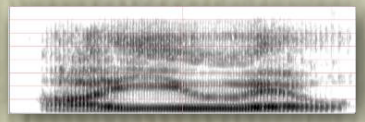
$s = (s_1, s_2, \ldots) =$

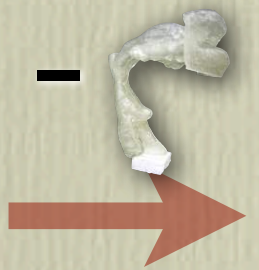


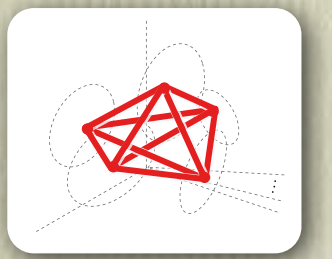
$=$



structure vector



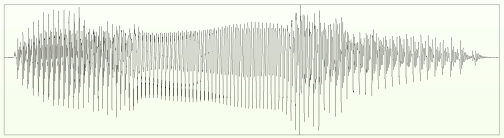




How to implement the vocal imitation?

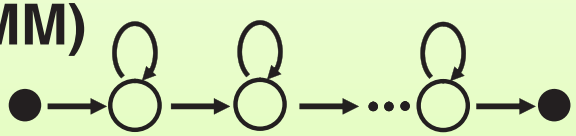
Extraction of a structure through training of an HMM

1. Speech waveforms



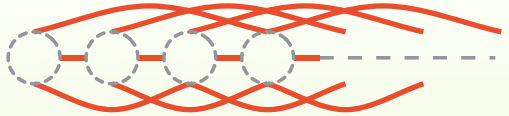
2. Cepstrum vector sequence

3. Cepstrum distribution sequence (HMM)



MAP estimation

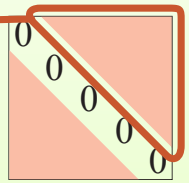
4. Bhattacharyya distances



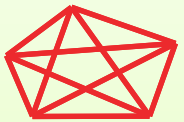
5. Structure (distance matrix)

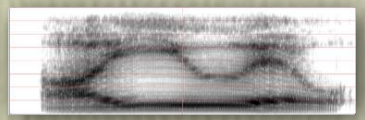
$s = (s_1, s_2, \ldots) =$

structure vector

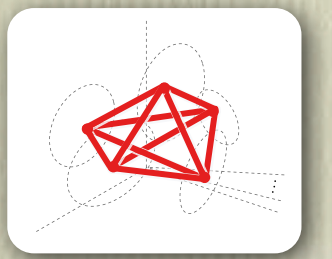


=





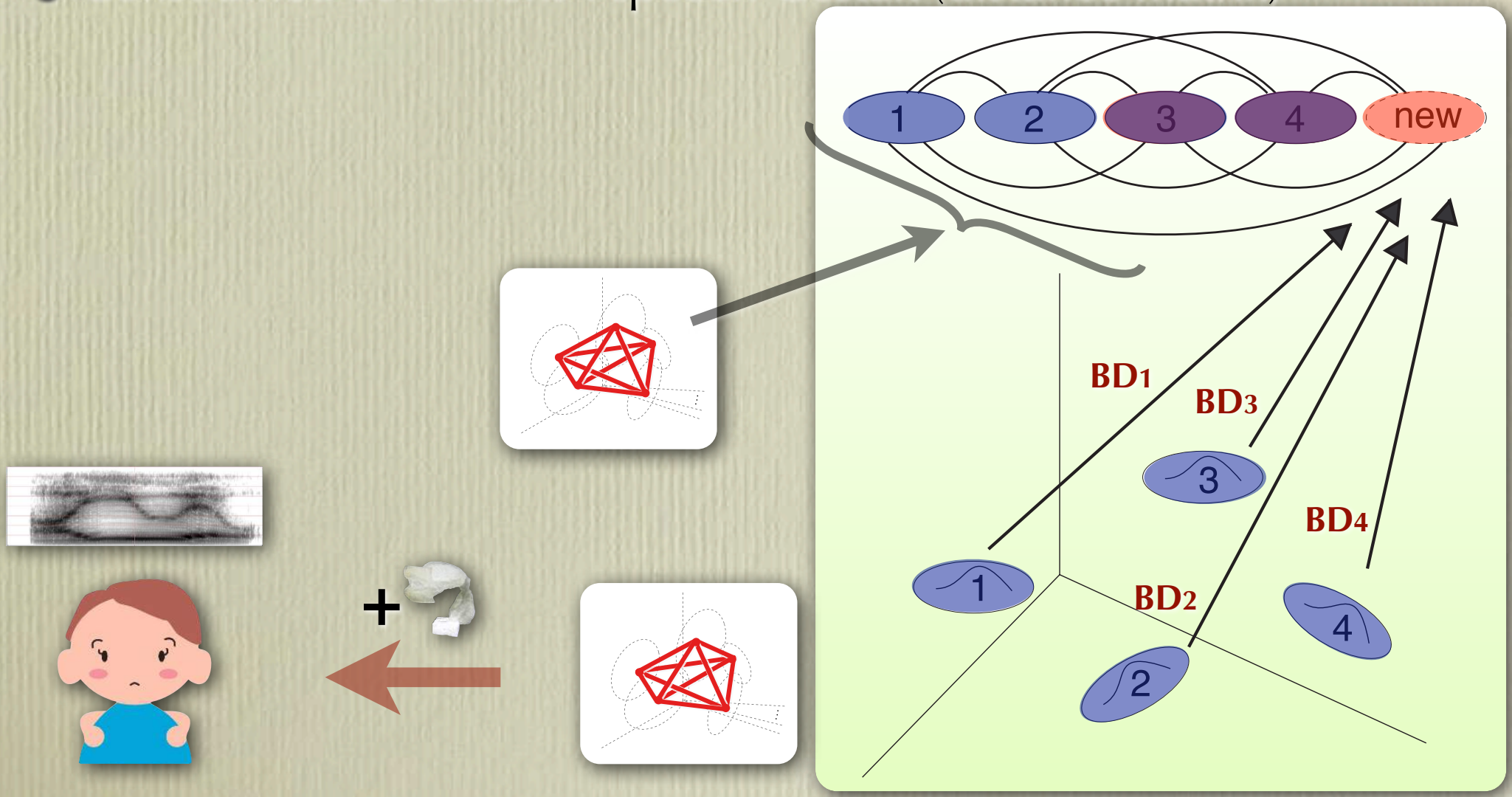
+



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)

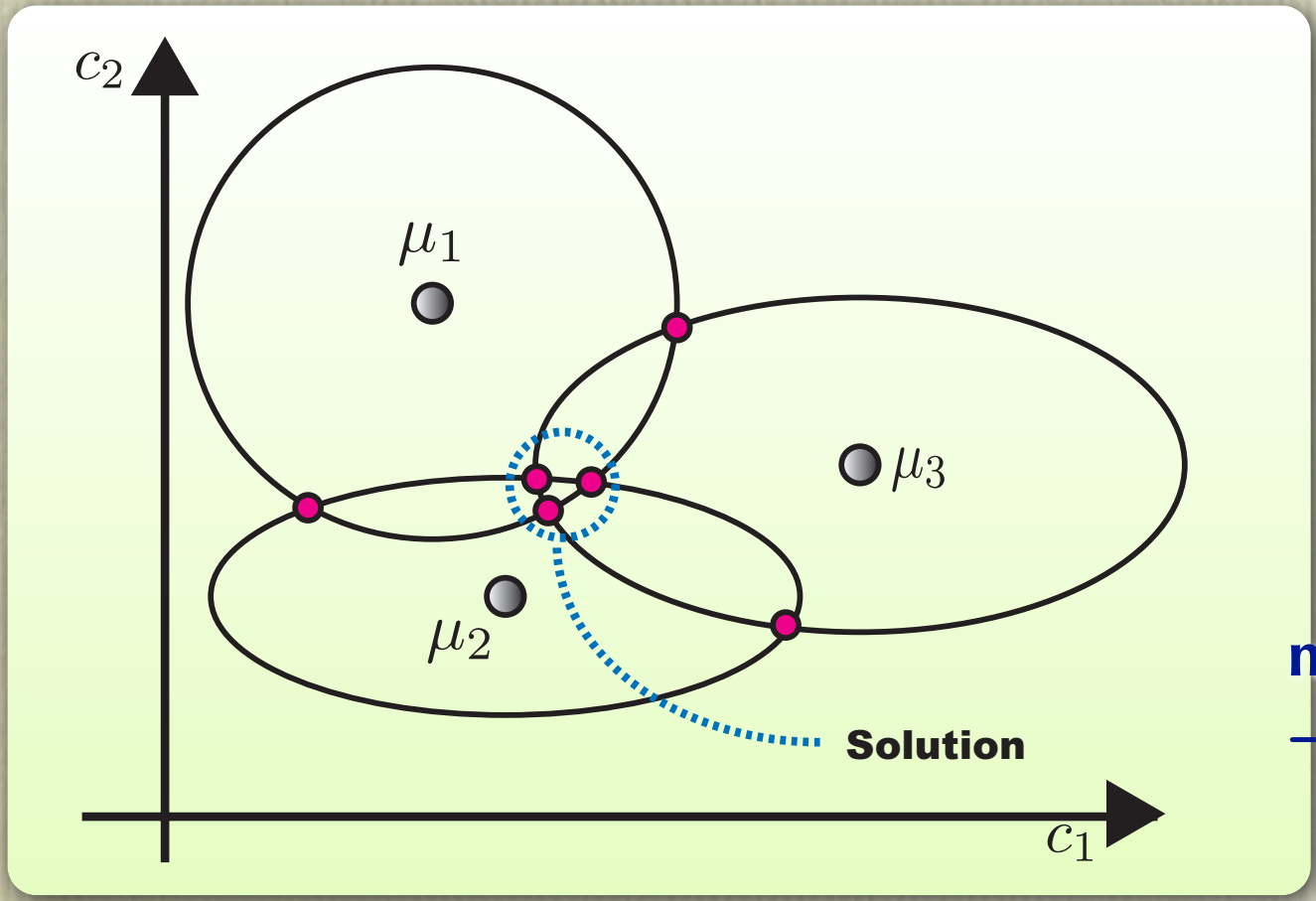


How to implement the vocal imitation?

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.

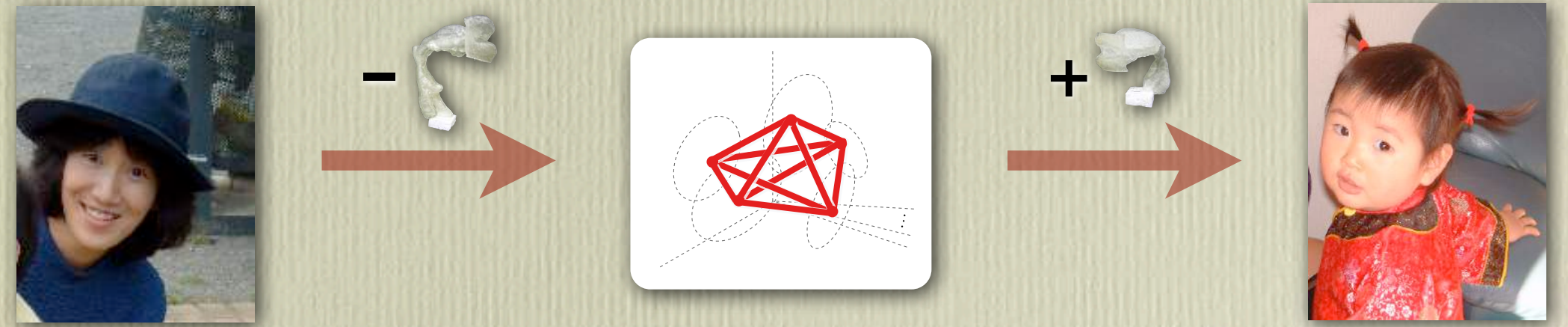


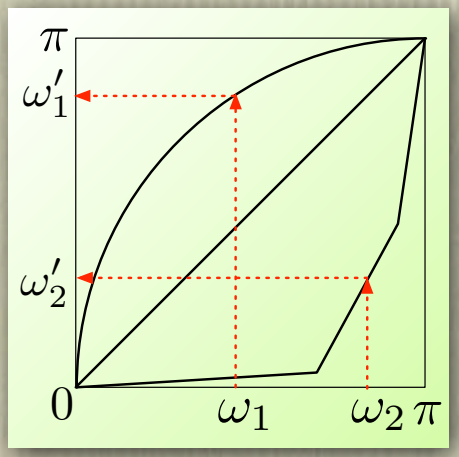
multiple solutions
⟶ averaging

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.

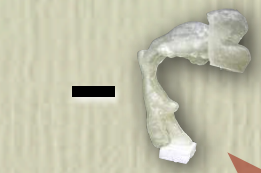


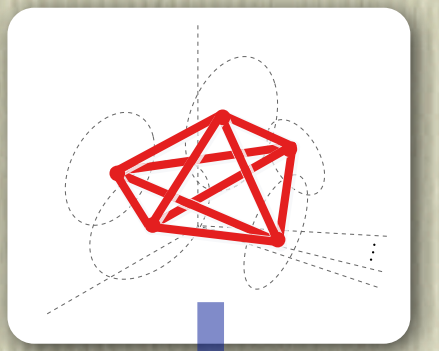


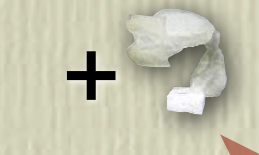
An experiment with real vocal imitation

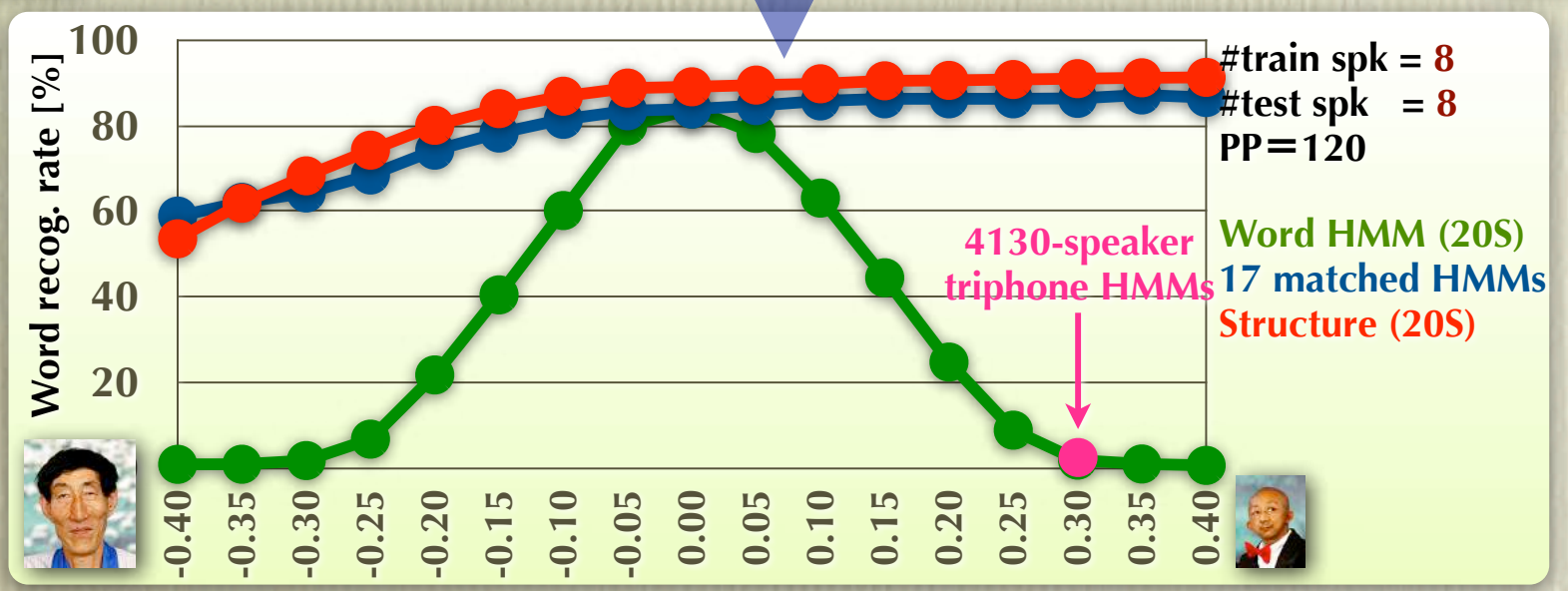
Demonstration with my wife and daughter

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







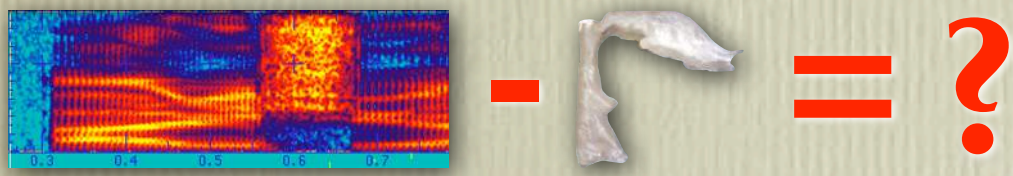


A big problem in CALL development

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 ?

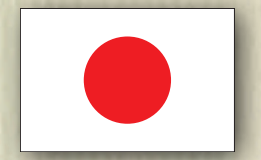
speaker	USA/F12	✗	Minematsu	○	Minematsu
gender	female	✗	male	○	male
age	?	✗	37	○	37
mic	Sennheiser	✗	cheap mic	○	cheap mic
room	recording room	✗	living room	○	living room
AD	SONY DAT	✗	PowerBook	○	PowerBook
proficiency	perfect	△	good	✗	Japanized

(Minematsu@ICSLP 2004)

A big solution for CALL development

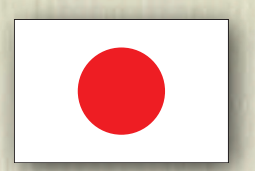
Proficiency estimation based on P(o | M)

USA/F12



Minematsu
(Japanized)

USA/M08



Minematsu
(Japanized)

(Minematsu@ICSLP 2004)

A big solution for CALL development

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

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

USA

matsu
nized)

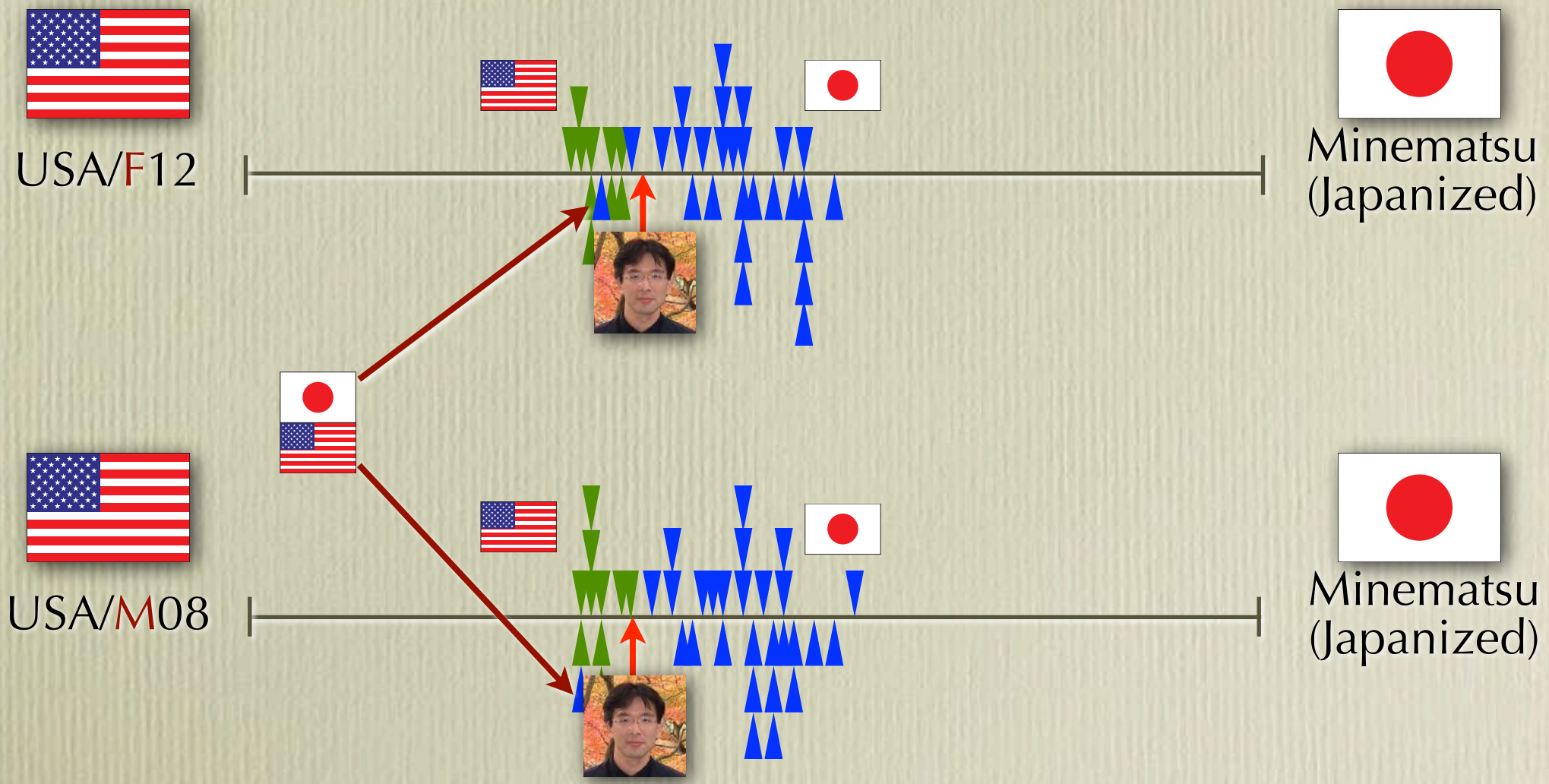
USA

matsu
nized)

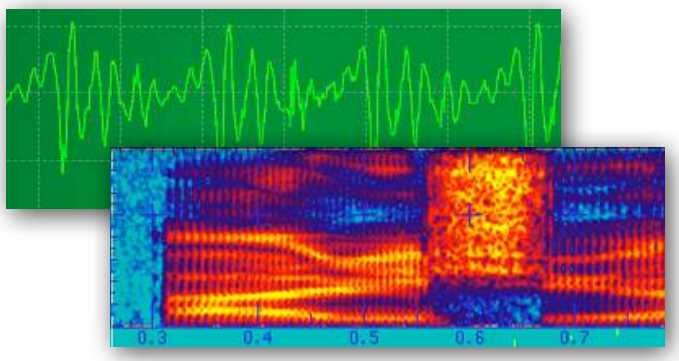
(Minematsu@ICSLP 2004)

A big solution for CALL development

Proficiency estimation based on structural distance



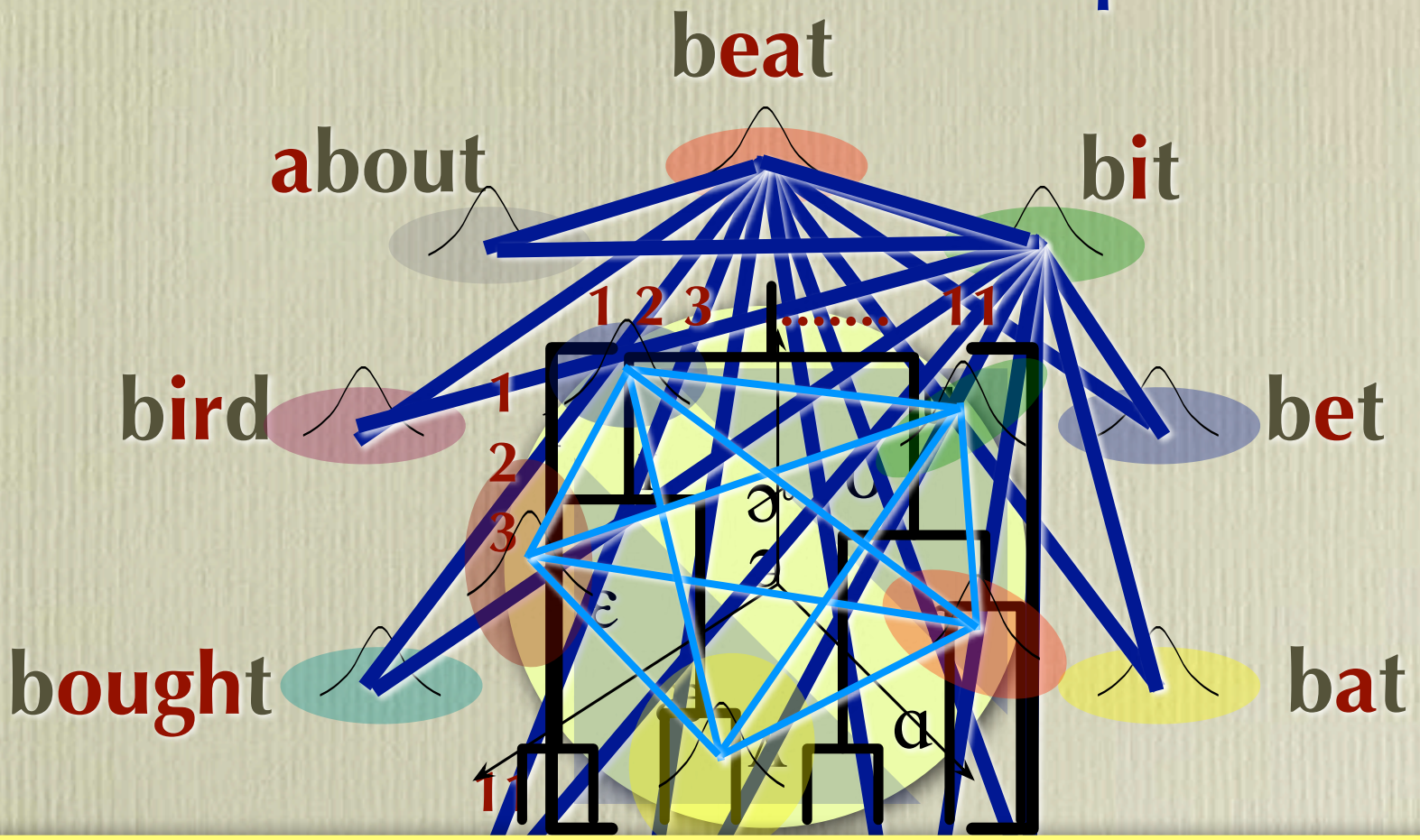
(Minematsu@ICSLP 2004)



proficiency	perfect	good	Japanized

Application of structures to CALL

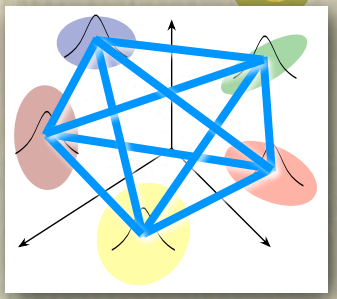
Vowel structure estimated from multiple utterances



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

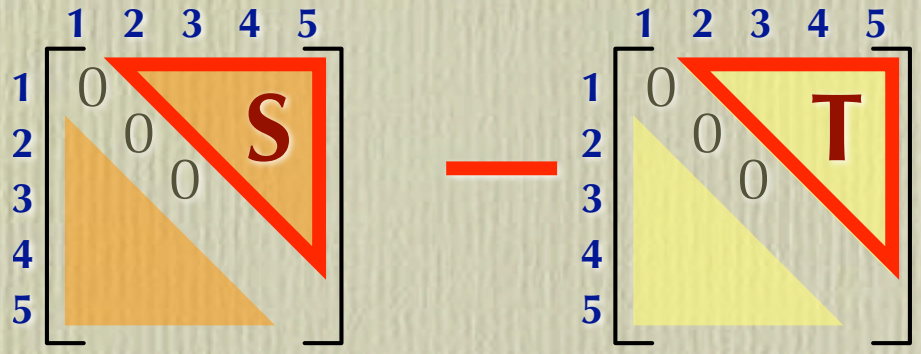
	ɑ	æ	ʌ	ə	ɚ	ɪ	i	ʊ	u	ɛ	ɔ
S1	J	J	J	J	J	J	J	J	J	J	J
S2	E	E	E	E	E	J	J	J	J	J	J
S3	J	J	J	J	J	E	E	E	E	E	E
S4	E	E	J	J	J	E	E	J	J	E	E
S5	J	J	E	E	E	J	J	E	E	J	J
S6	E	J	E	J	E	J	J	J	J	E	E
S7	J	E	J	E	J	E	E	E	E	J	J
S8	E	E	E	E	E	E	E	E	E	E	E

ɑ, æ, ʌ, ə, ɚ	a
ɪ, i	i
ʊ, u	u
ɛ	e
ɔ	o

Clustering of learners

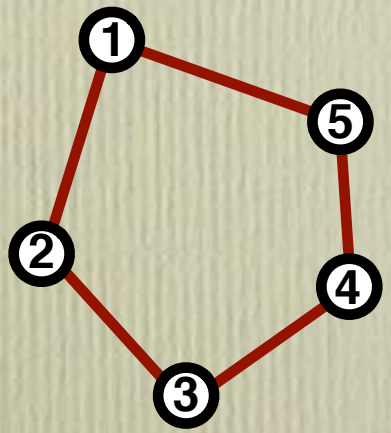
Structure-to-structure distance measure

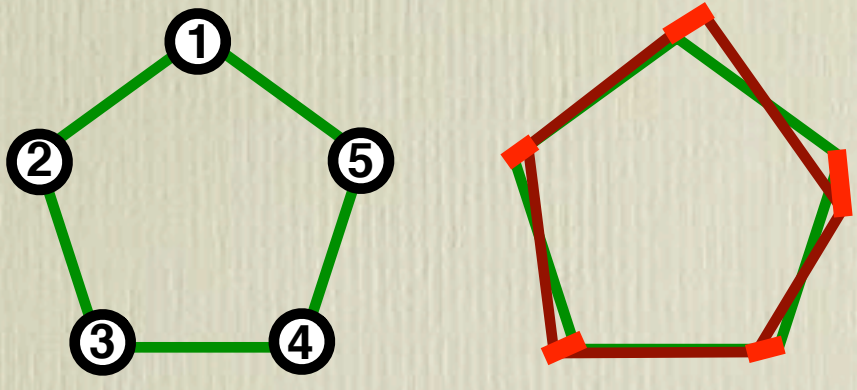
- Euclidian distance between two distance matrices



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

- Can approximate the structural distance after shift and rotation



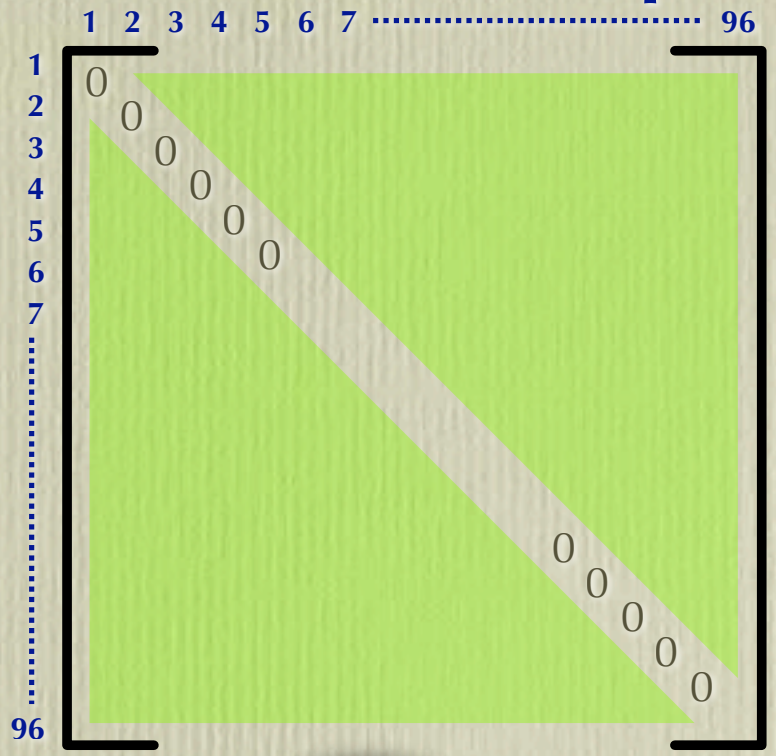


Minimum of the total distances between corresponding points

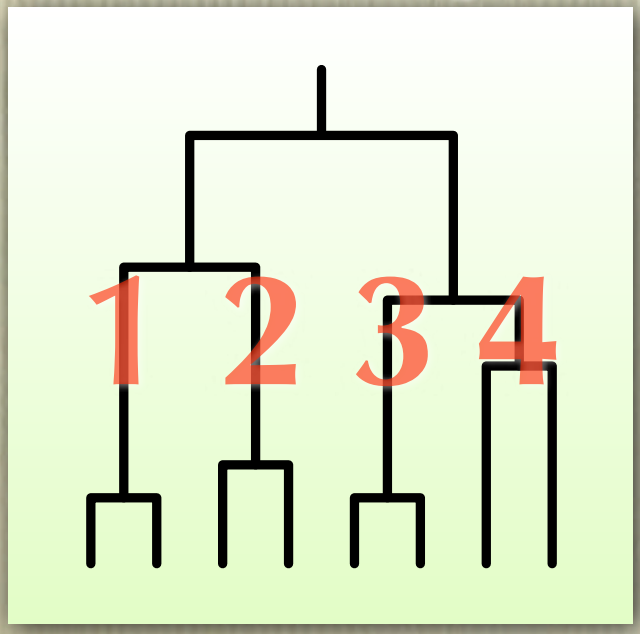
Clustering of learners

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

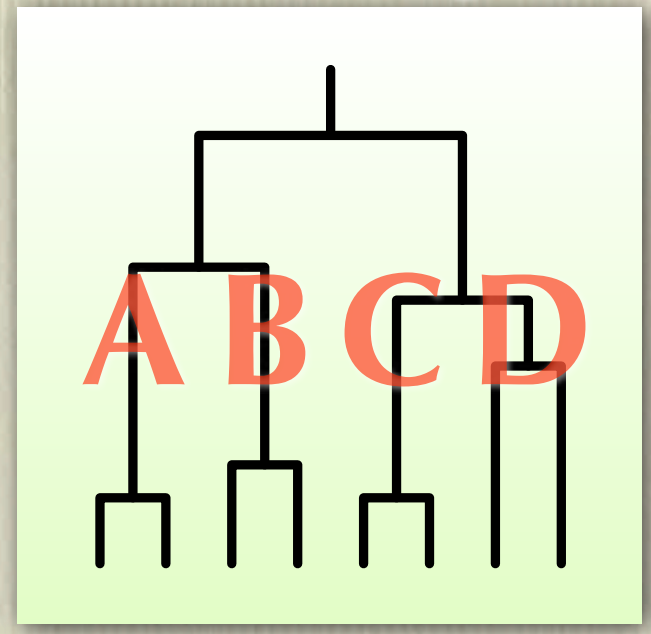
- Speakers: A to L
- Prons: 1 to 8



Pronunciation clustering



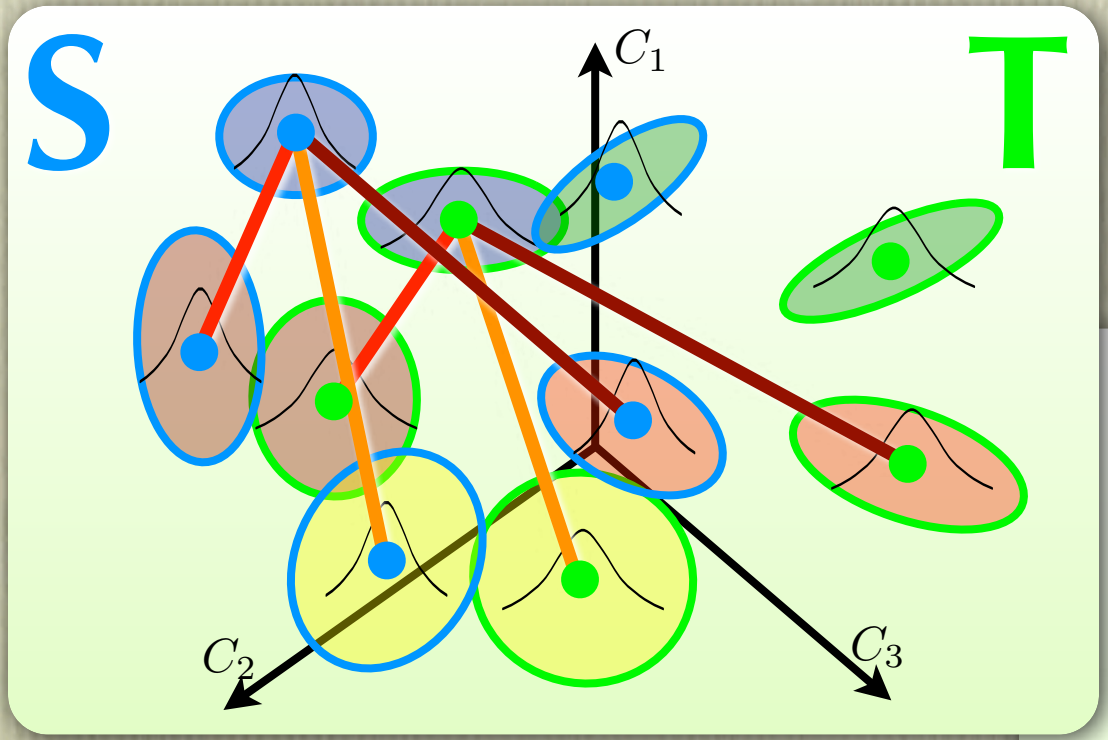
Speaker clustering



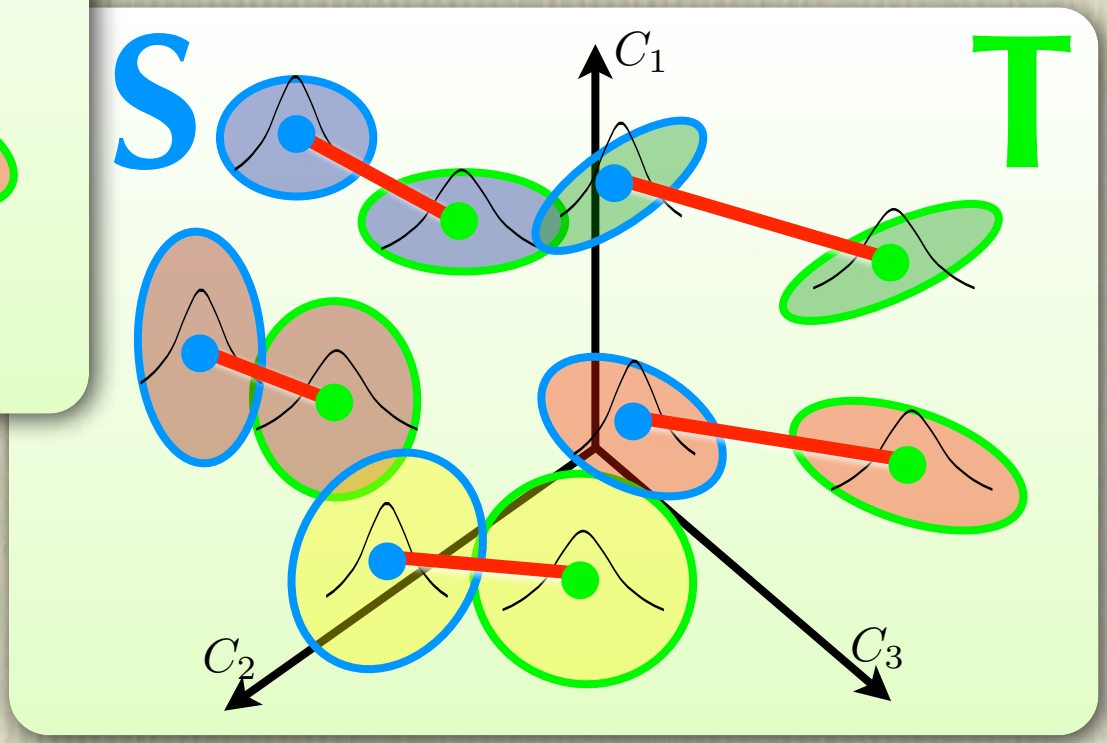
Clustering of learners

Another distance measure between two structures

- Contrast-based comparison
- Substance-based comparison



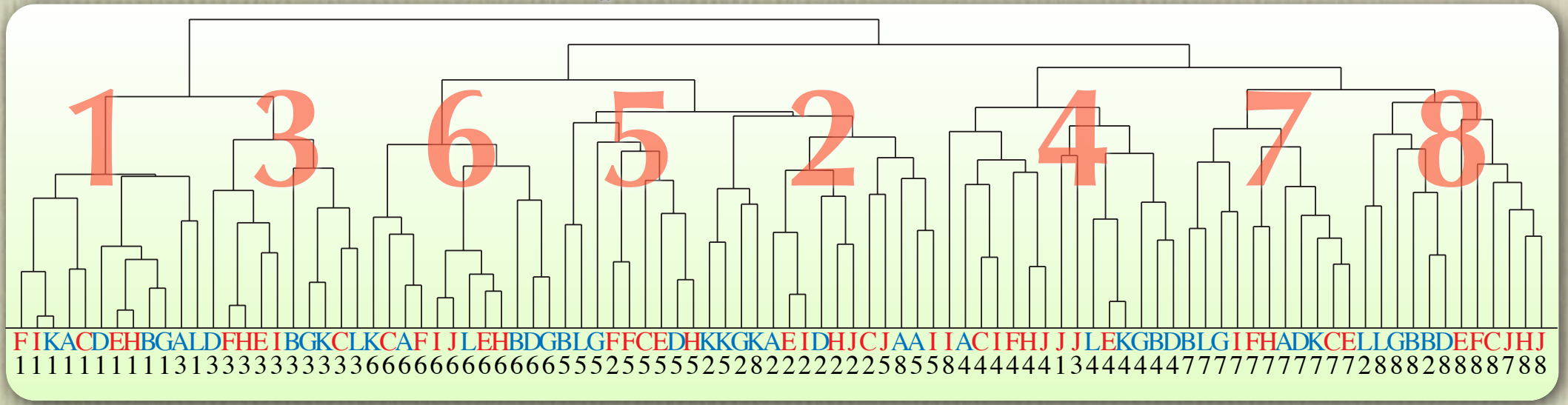
$$\sqrt{\frac{1}{M}\sum_{i<j}(S_{ij}-T_{ij})^2}$$



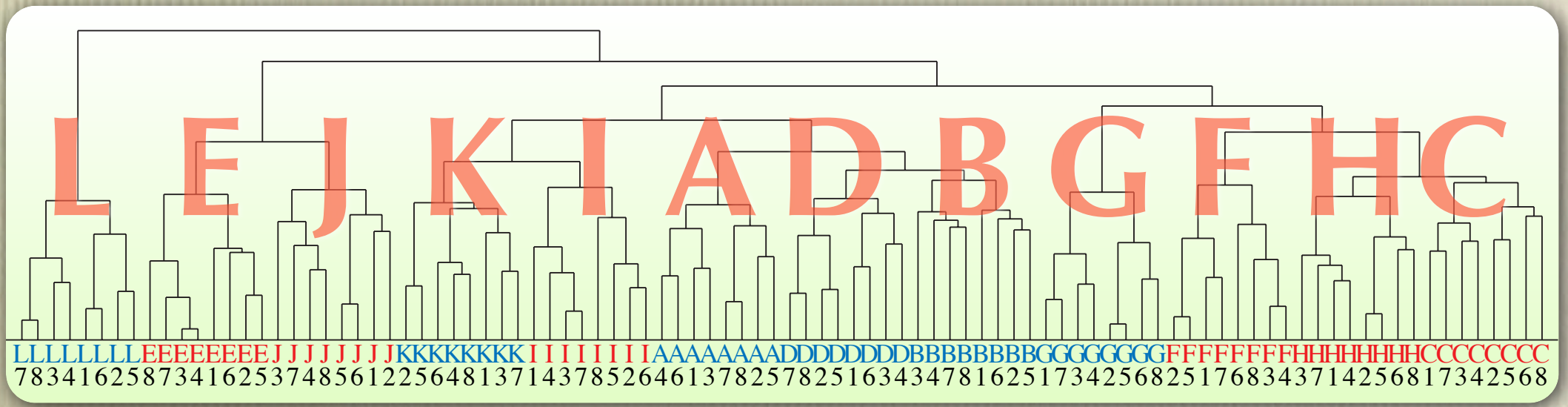
$$\sqrt{\frac{1}{M}\sum_{i} BD(v_i^S, v_i^T)}$$

Clustering of learners

Contrast-based comparison



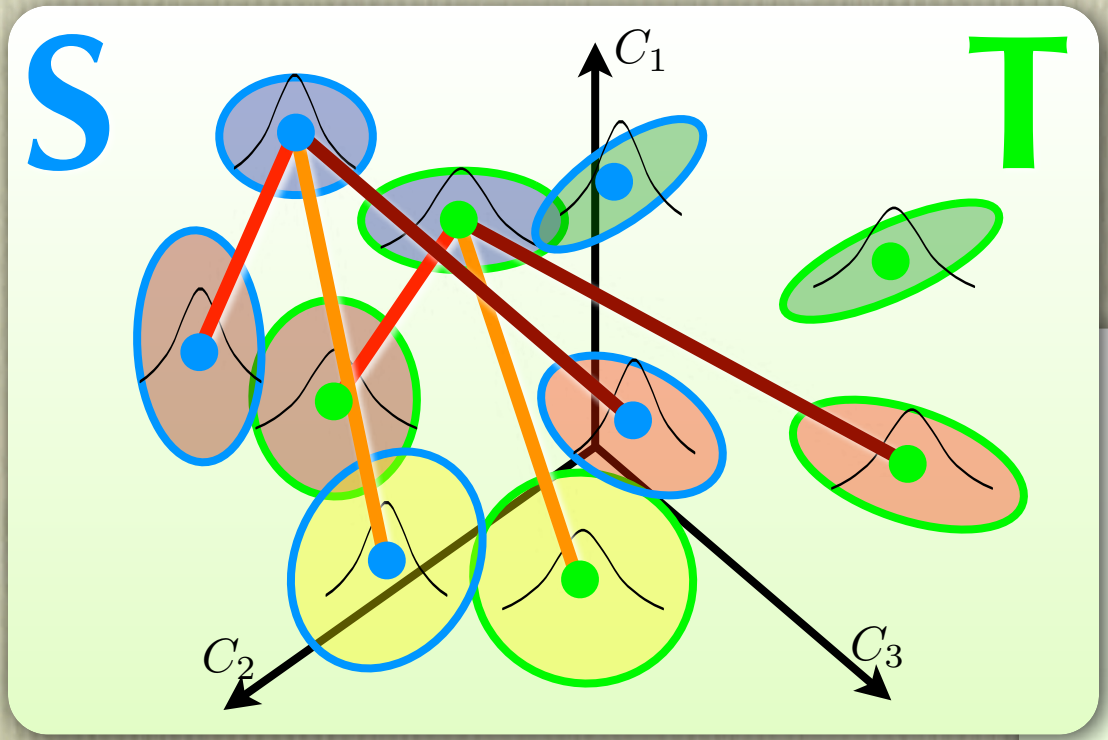
Substance-based comparison



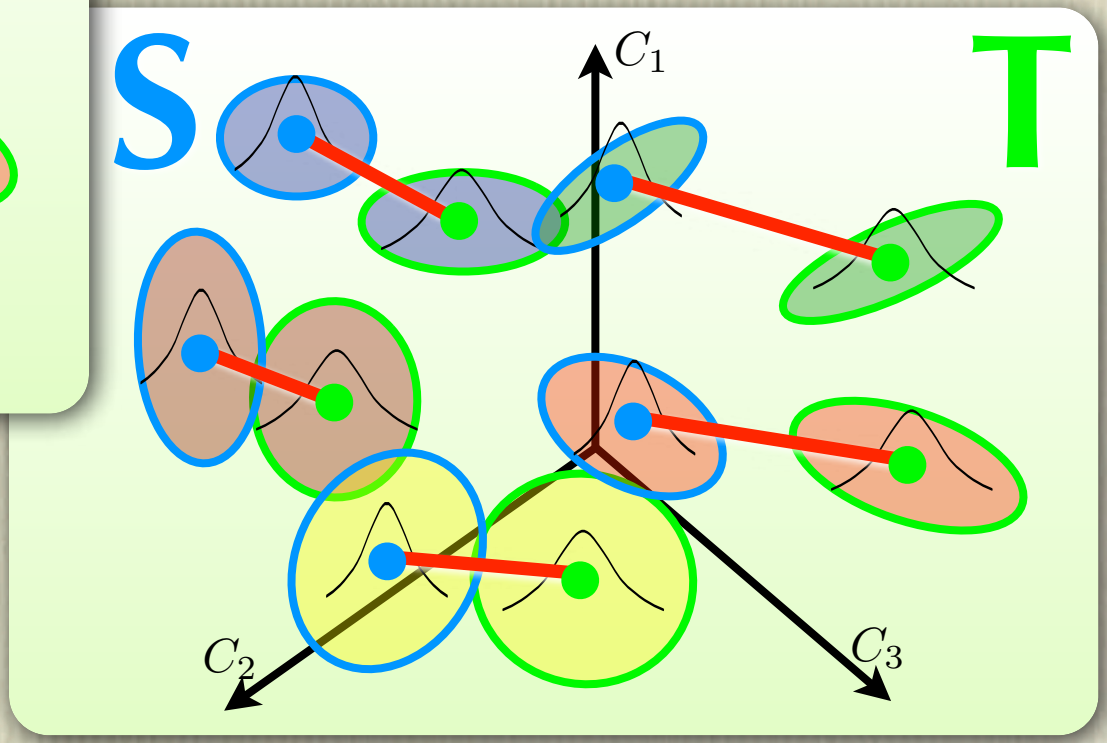
Clustering of learners

Another distance measure between two structures

- Contrast-based comparison
- Substance-based comparison



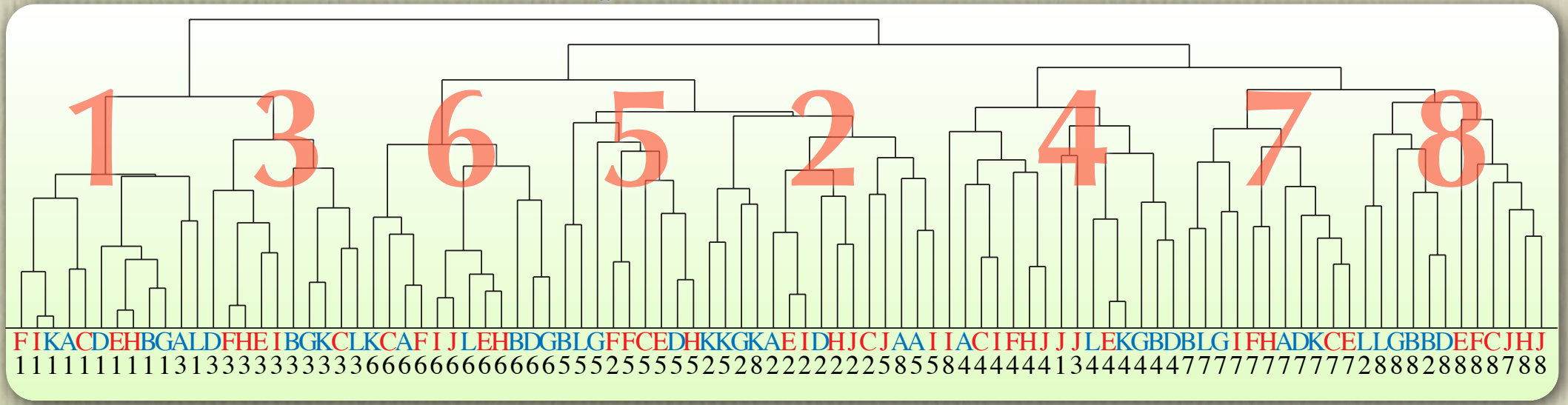
$$\sqrt{\frac{1}{M}\sum_{i<j}(S_{ij}-T_{ij})^2}$$



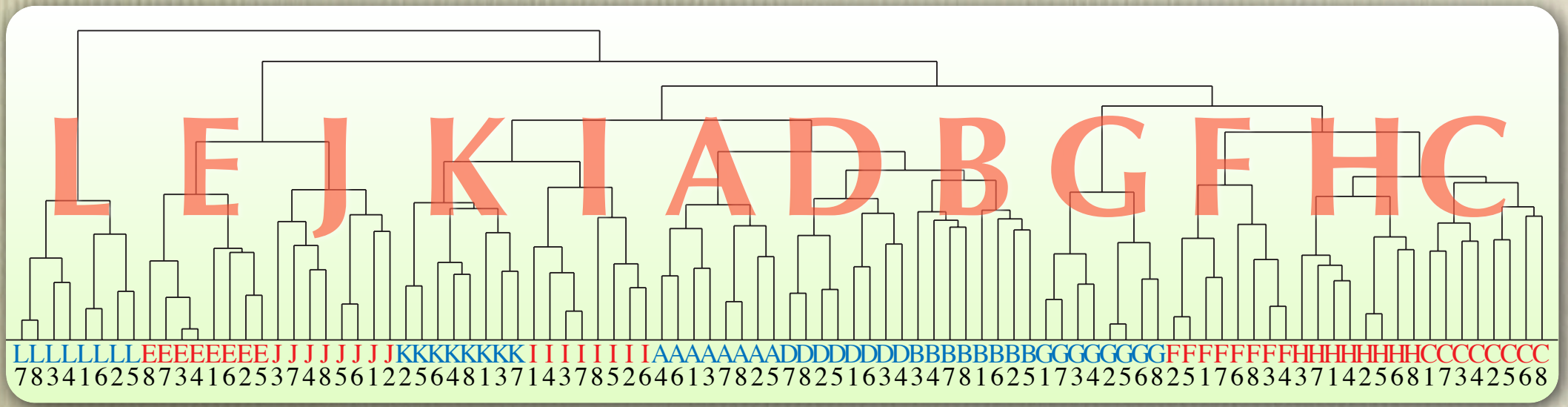
$$\sqrt{\frac{1}{M}\sum_{i} BD(v_i^S, v_i^T)}$$

Clustering of learners

Contrast-based comparison



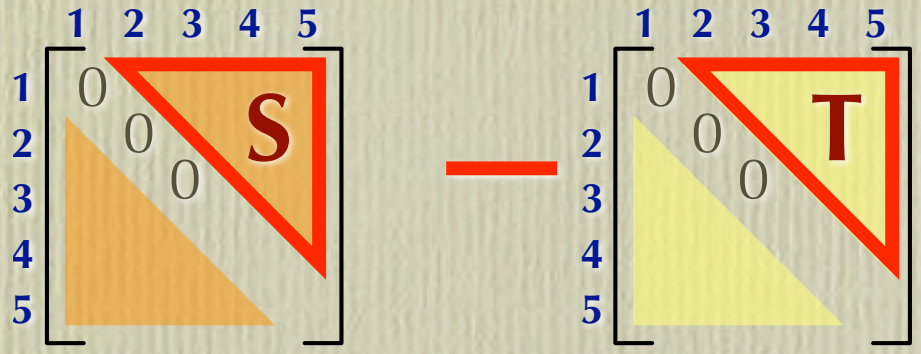
Substance-based comparison



Which vowels to correct at first?

Global difference between Student and Teacher

- Euclidian distance between two distance matrices



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

Can be decomposed into local differences

- Contribution of individual vowels to the global difference

$$d(v)=\sqrt{\frac{1}{M}\sum_{j=1}^{M}(S_{vj}-T_{vj})^2}$$

- Vowels of larger $d(v)$ are should be corrected at first!!!

Which vowels to correct at first?

Estimation of the order of vowel correction

- Only with given two matrices without the replacement table

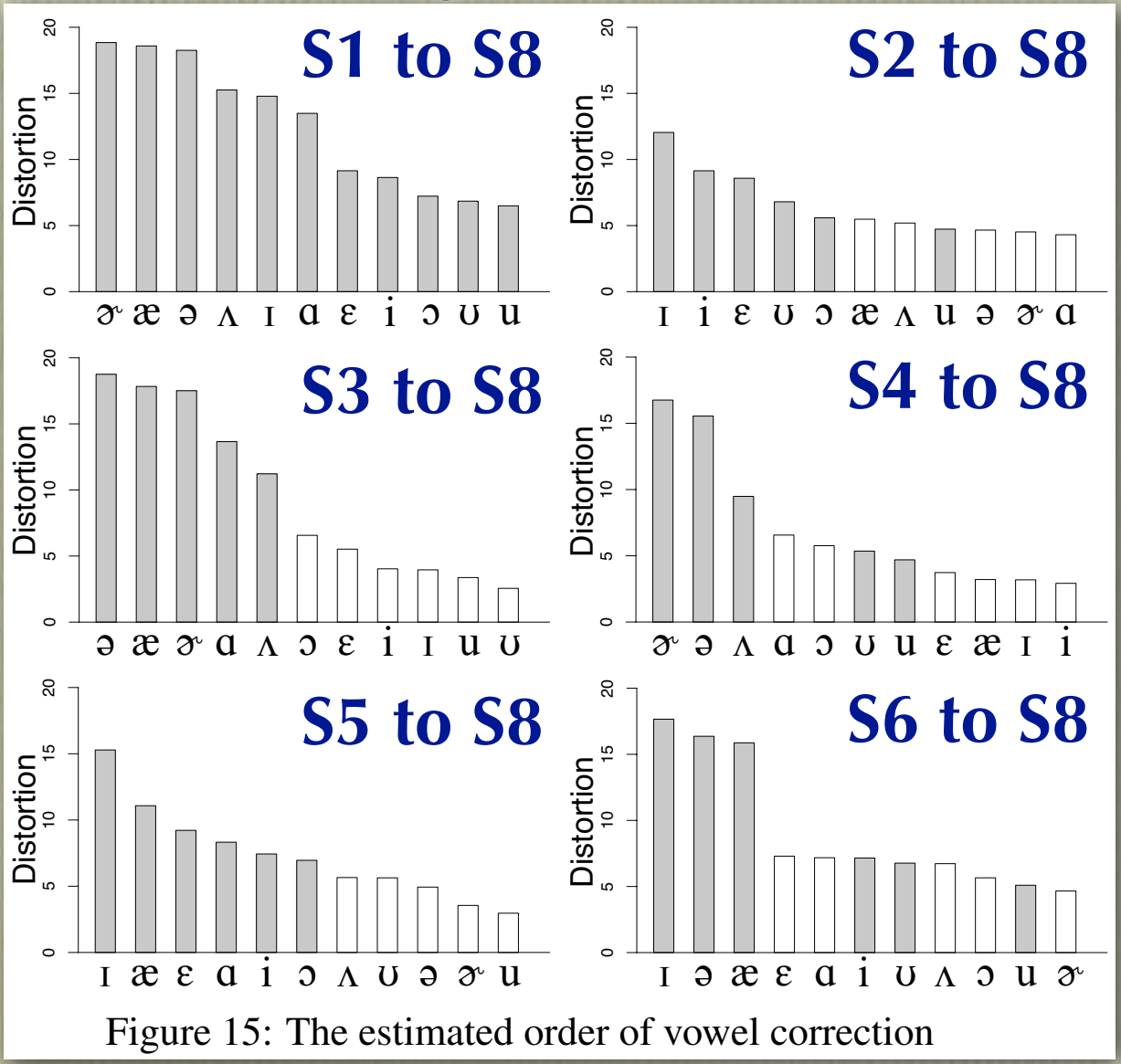


Figure 15: The estimated order of vowel correction

	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

S1 - S7 : Japanese English

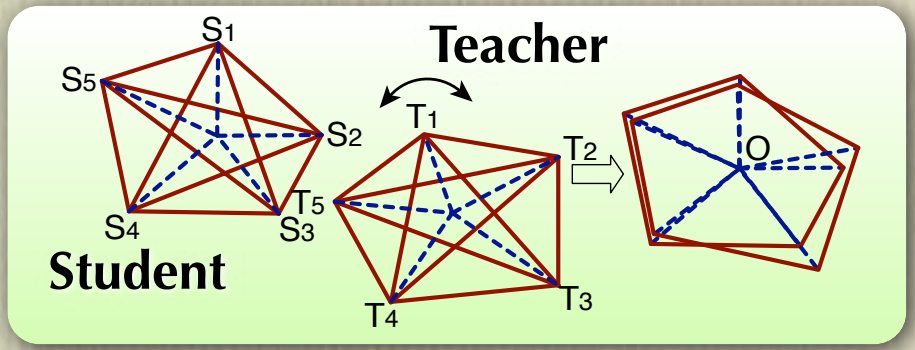
S8 : American English

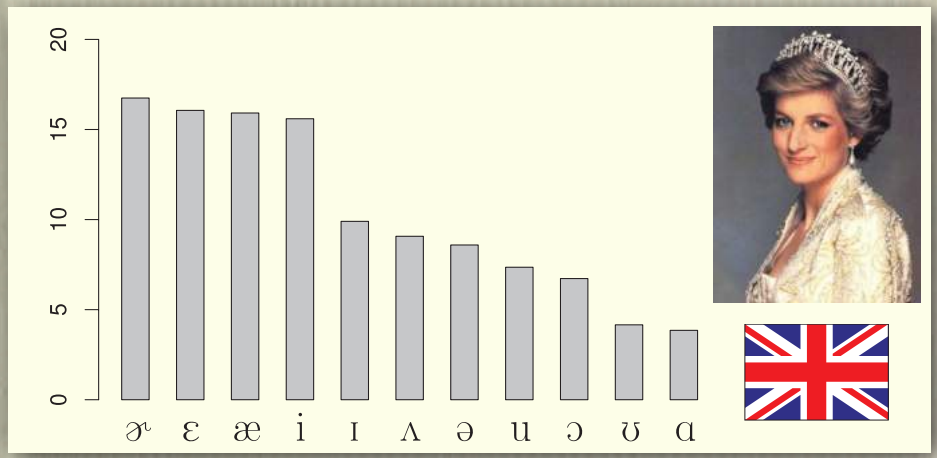
: replaced by J vowels

: no replacement

Very motivating interface for CALL

Select your favorite teacher!!





Shortest cut to your model!

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 your definition of "human-like" machines?

The final lecture of CMP

- **will be given on next Tuesday (Jan 16).**
- **The final assignment will also be given on that day.**
  - If you cannot attend it, you should view the video.