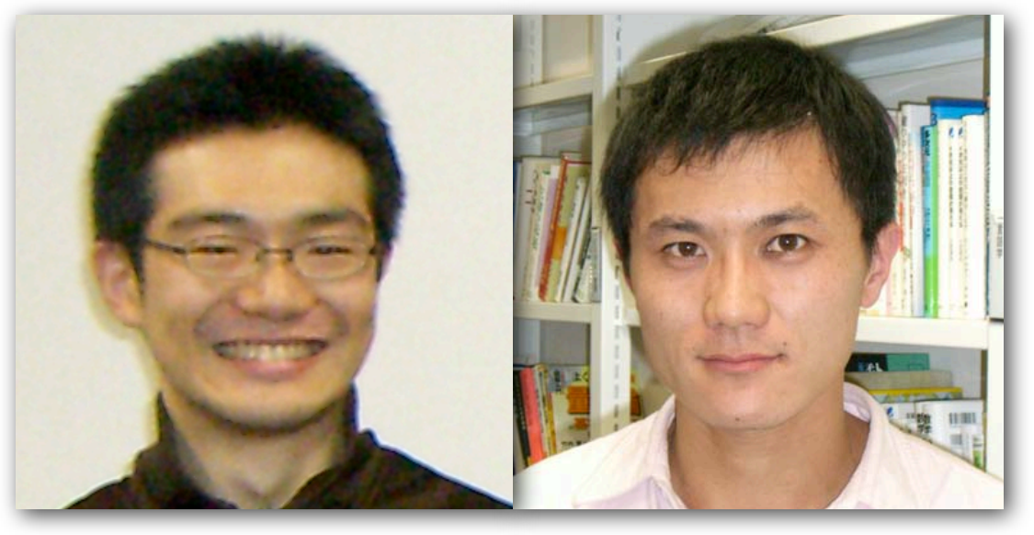
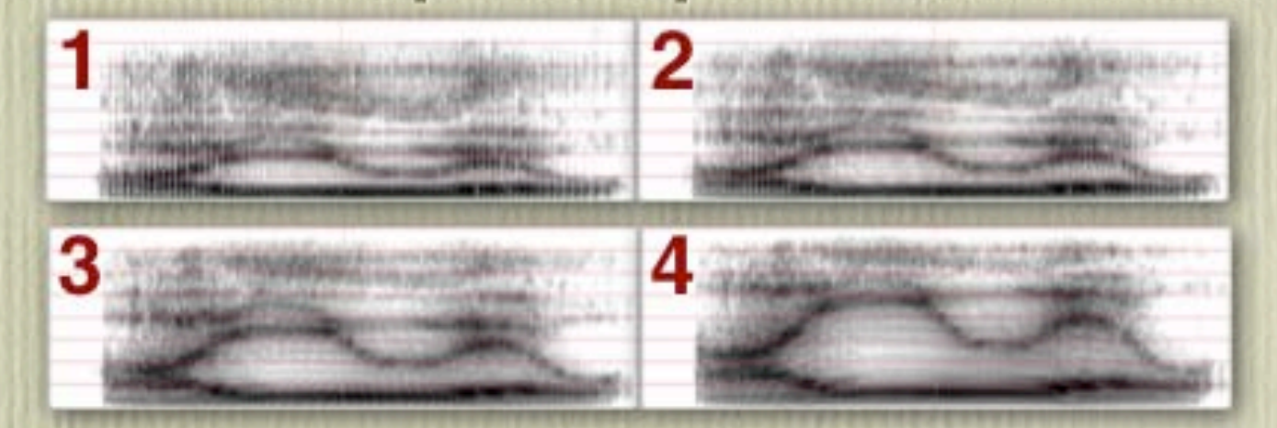
PRONUNCIATION CLINIC

Nobuaki Minematsu, Max Takazawa, Xuebin Ma @ The University of Tokyo

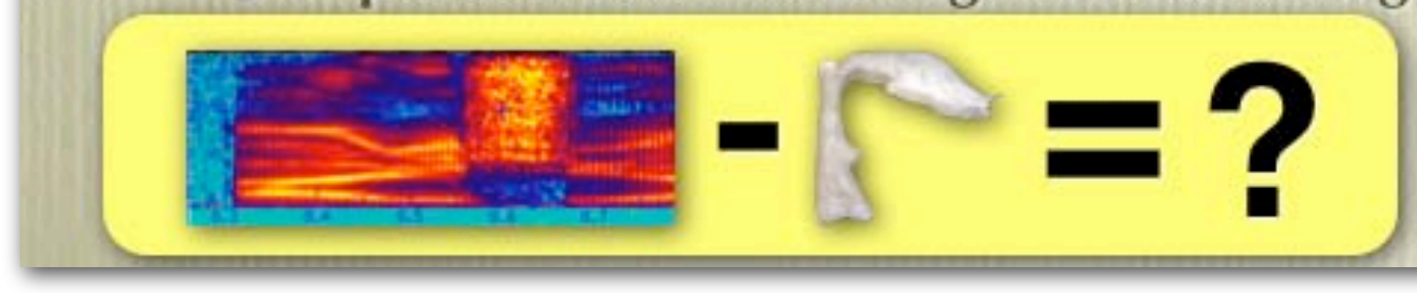


!! A big problem !!

- A very important and requisite function for CALL systems
  - The system has to be able to ignore speaker individualities.
    - Age and gender (the size and length of the vocal tube)
    - A desirable system must not be able to see differences in age and gender.
  - Some examples of speaker differences

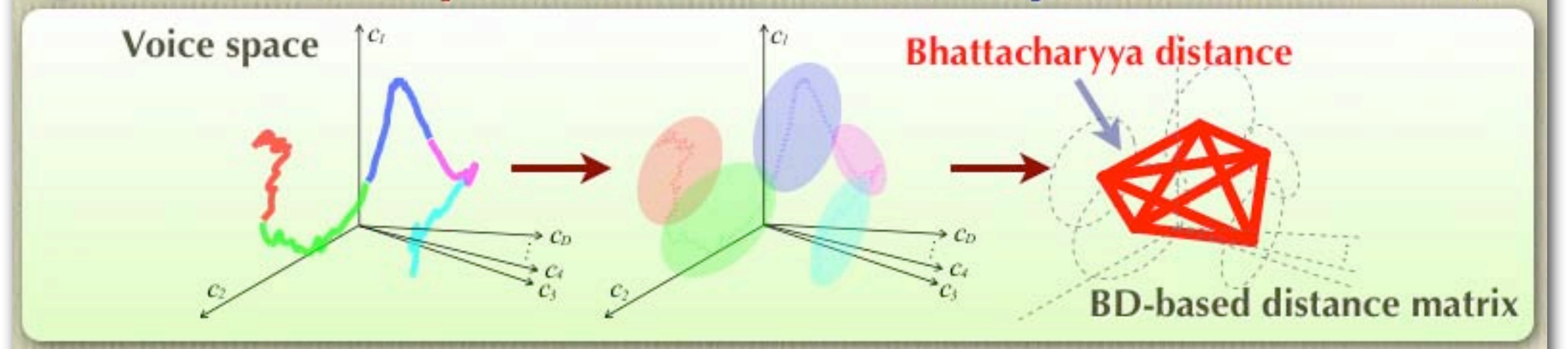


  - Source + filter model
    - Separation between source and filter
    - Separation between ling. and extra-ling.

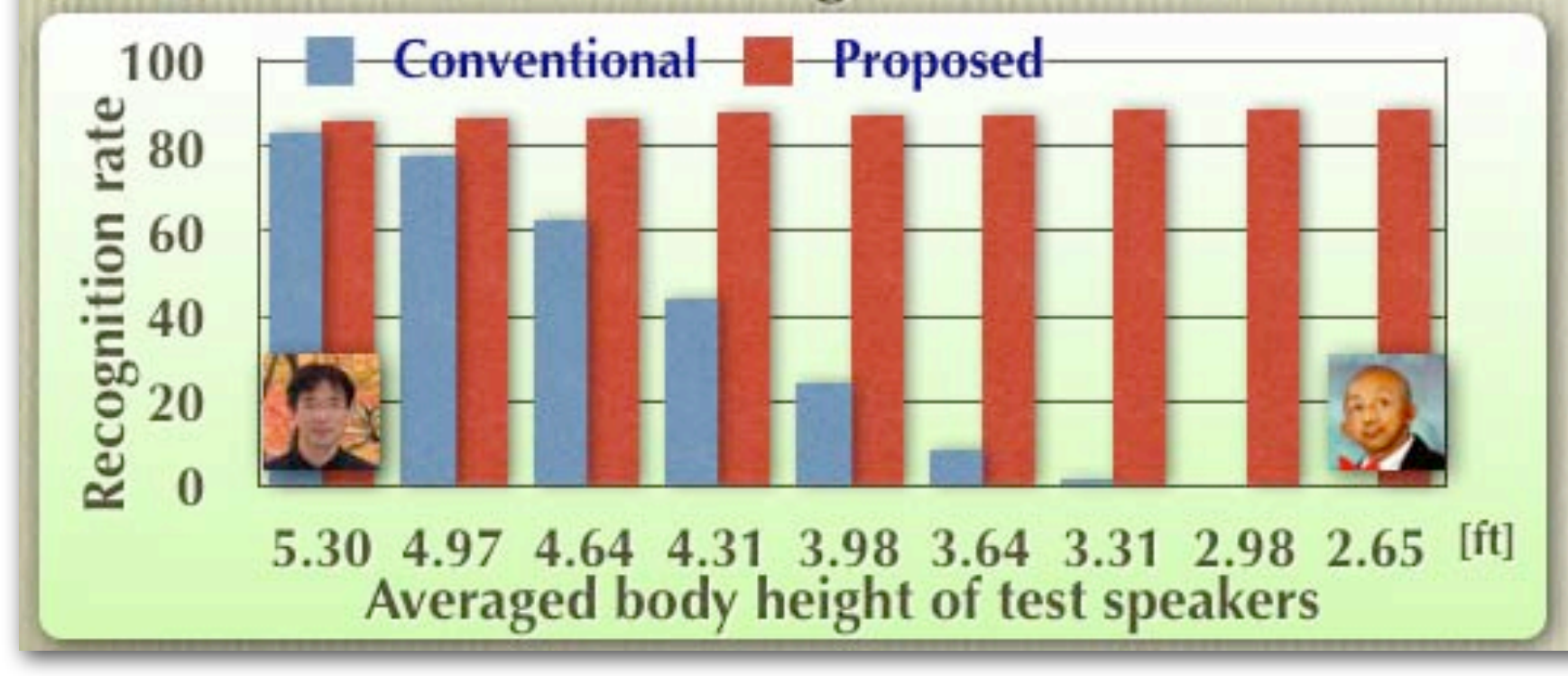


!! Our solution of that problem !!

- Holistic and speaker-invariant sound pattern (structure)

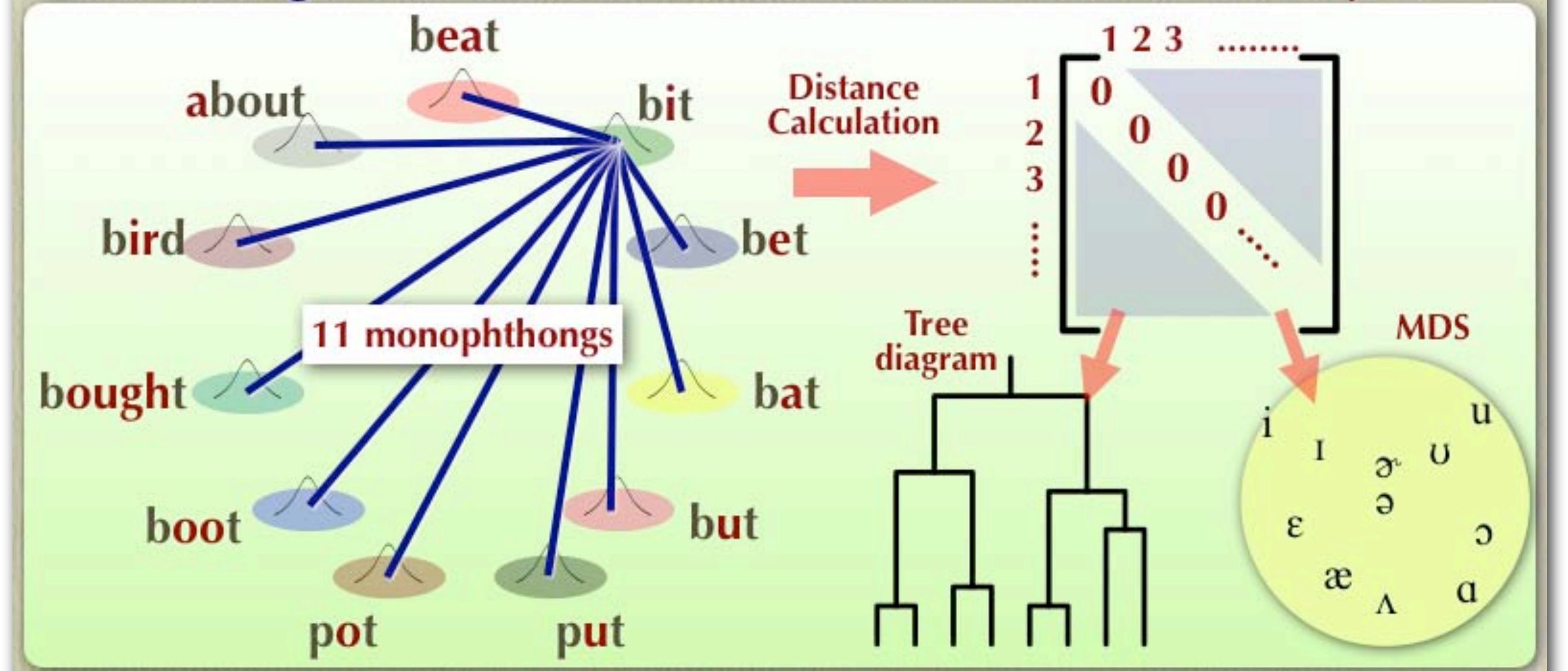


- Use of structures for automatic speech recognition
  - Isolated word recognition (word = 5 vowel sequence, e.g. /aeoui/)



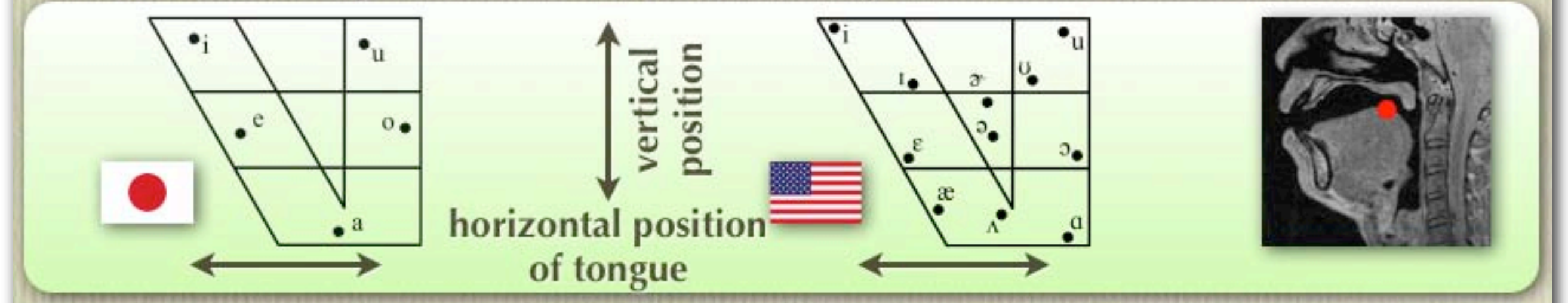
A vowel training system for everybody!!

- Learning not of individual vowels but of a vowel system



Structural representation of the vowel system

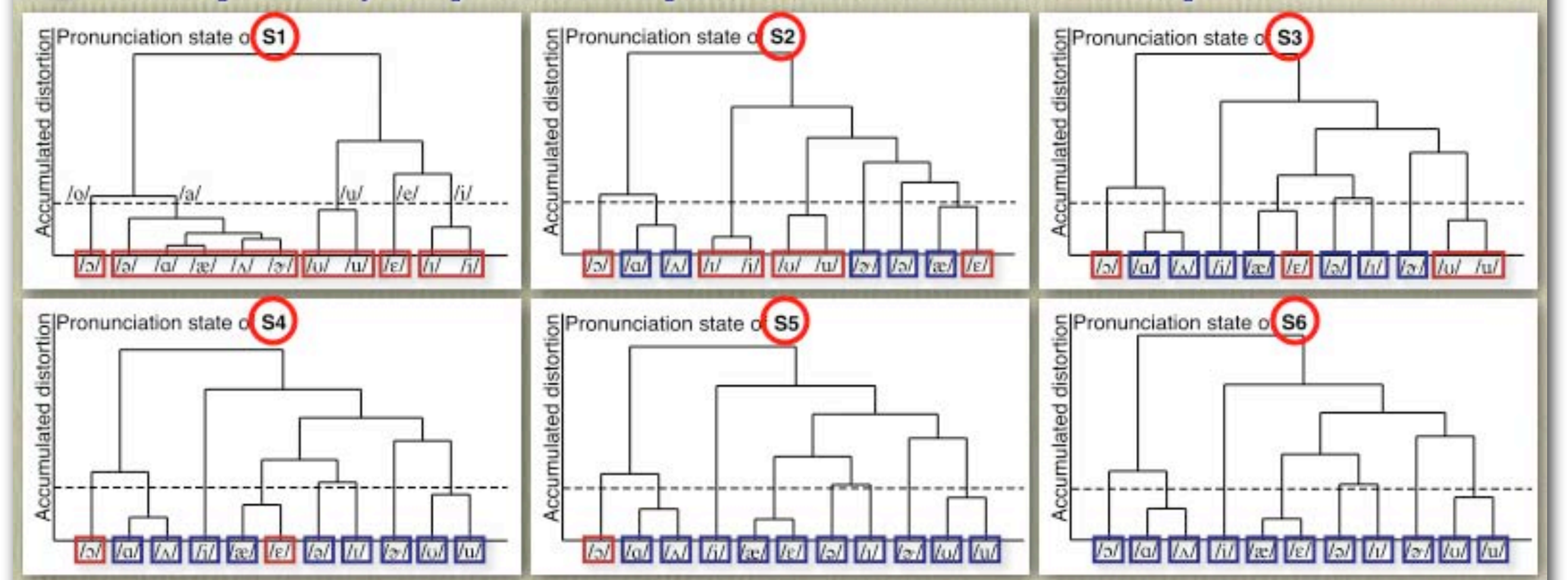
- Vowel system and vowel chart
  - Accented pronunciation = vowel system with some distortion

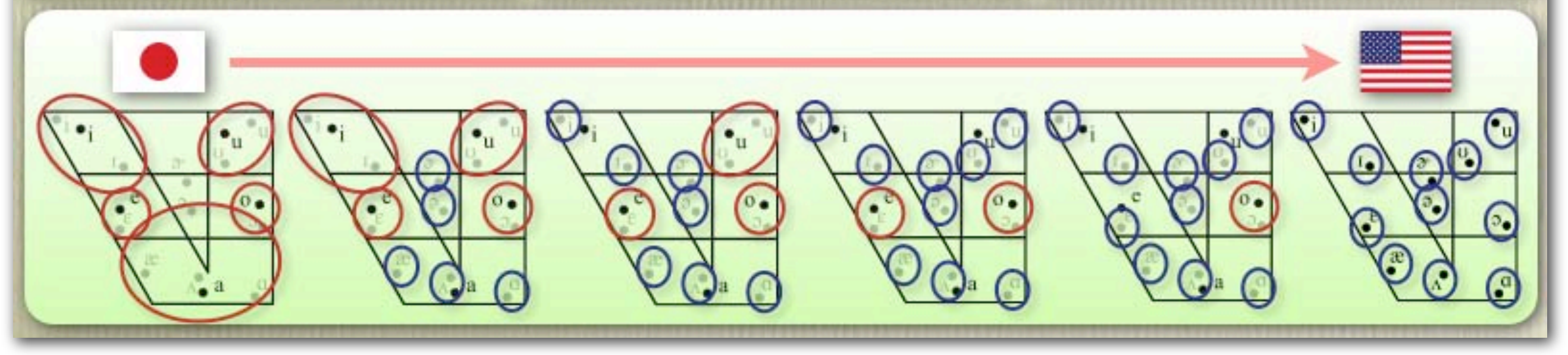


- What's possible in the proposed demo system
  - The demo system can
    - 1. record or log a history of vowel pronunciation training of each learner.
    - 2. provide for learners a window of "favorite teacher selection".
    - 3. show which vowel to correct first to become like the selected teacher.
    - 4. classify all the registered learners only wrt pronunciation proficiency by ignoring gender, age, etc. very effectively.
    - 5. give a very motivating user-interface for pronunciation training.

Developmental changes in vowel training ①

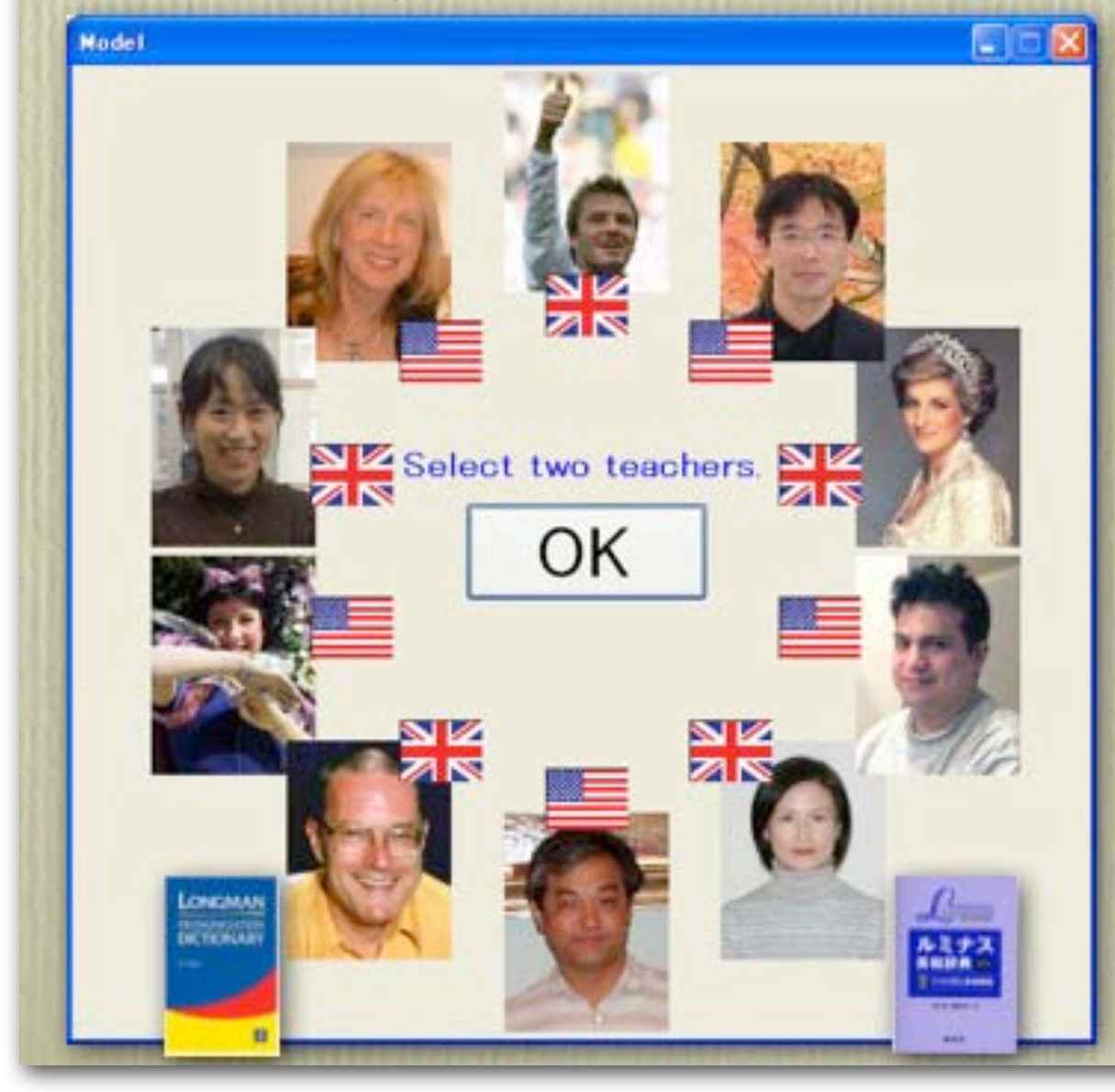
- Completely Japanized pronunciation to AE pronunciation

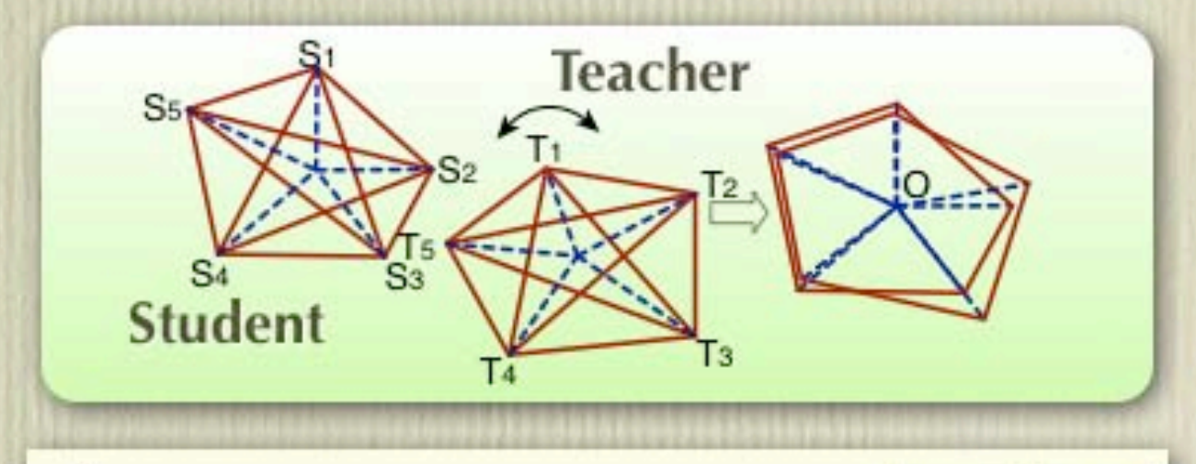


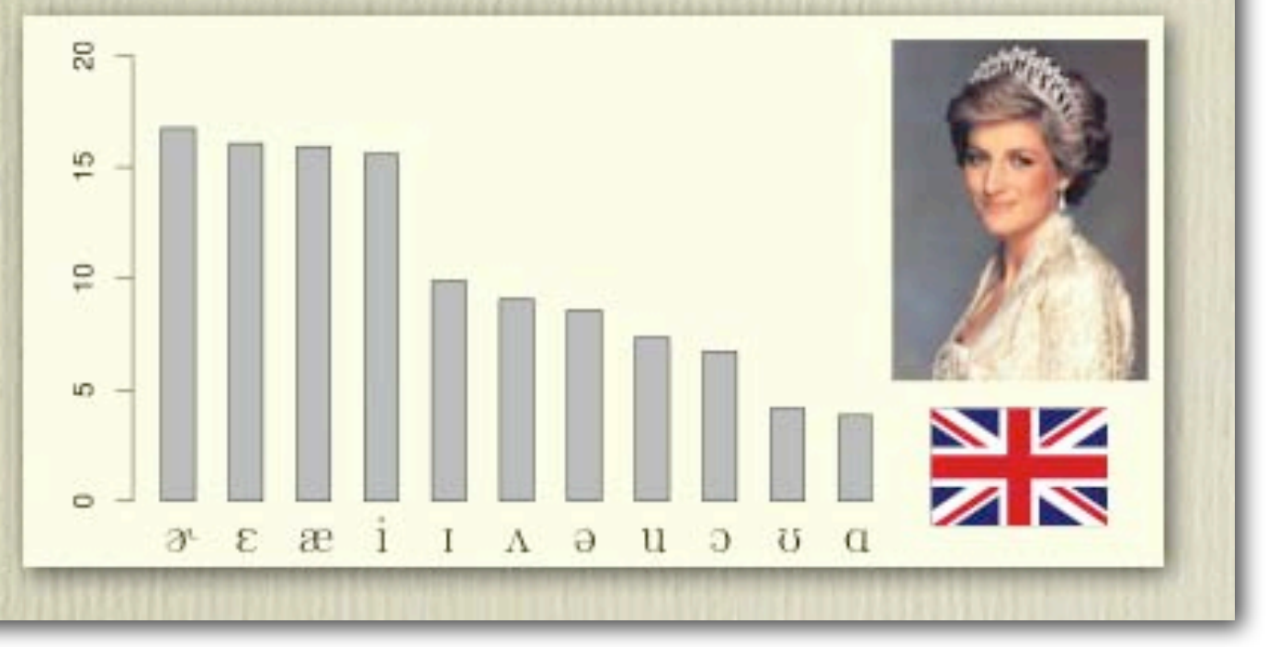


Which vowels to correct at first in your case? ②③

- Who is your model speaker?
  - A famous phonetician, a movie star (character) or a sport player??
- Which vowels to correct at first to become like him/her?
  - The system can show the shortest cut to the model pronunciation.

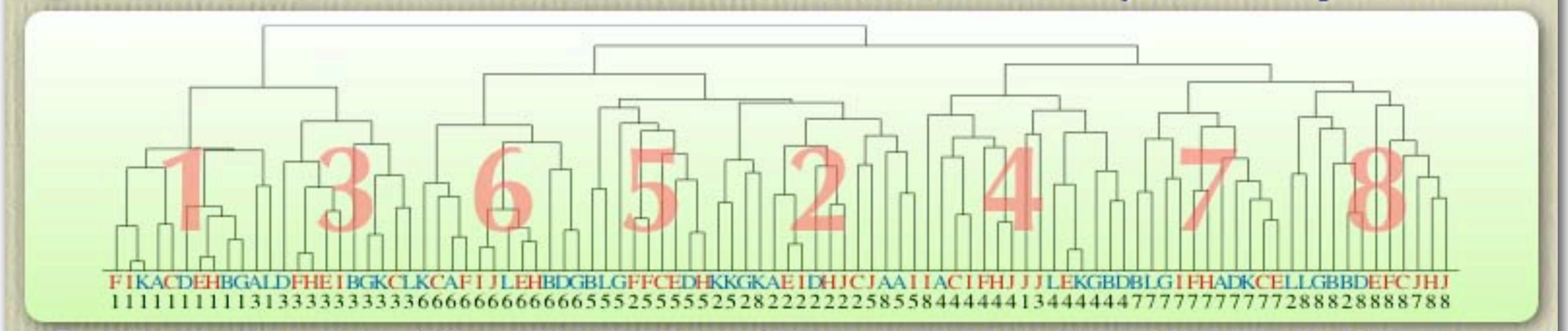




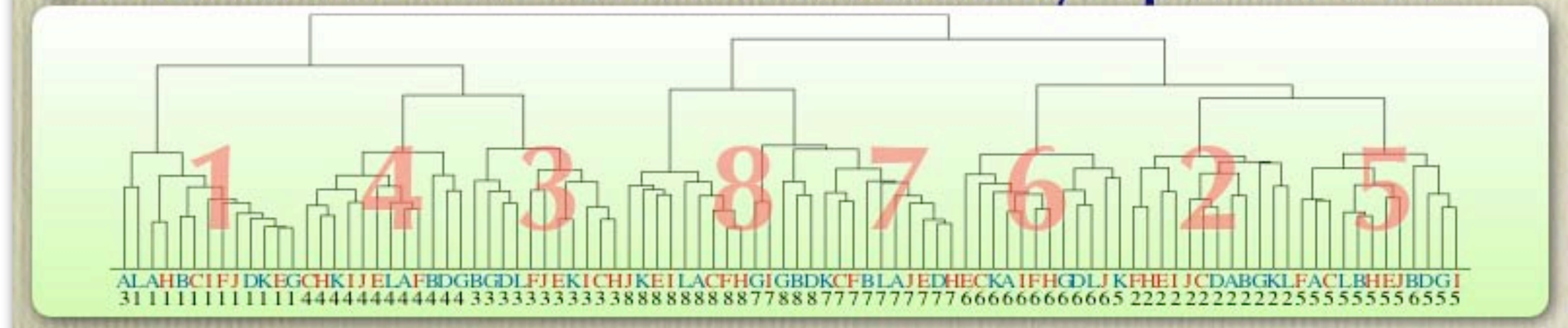


Classification of learners ④

- Automatic classification of 96 learners by a computer



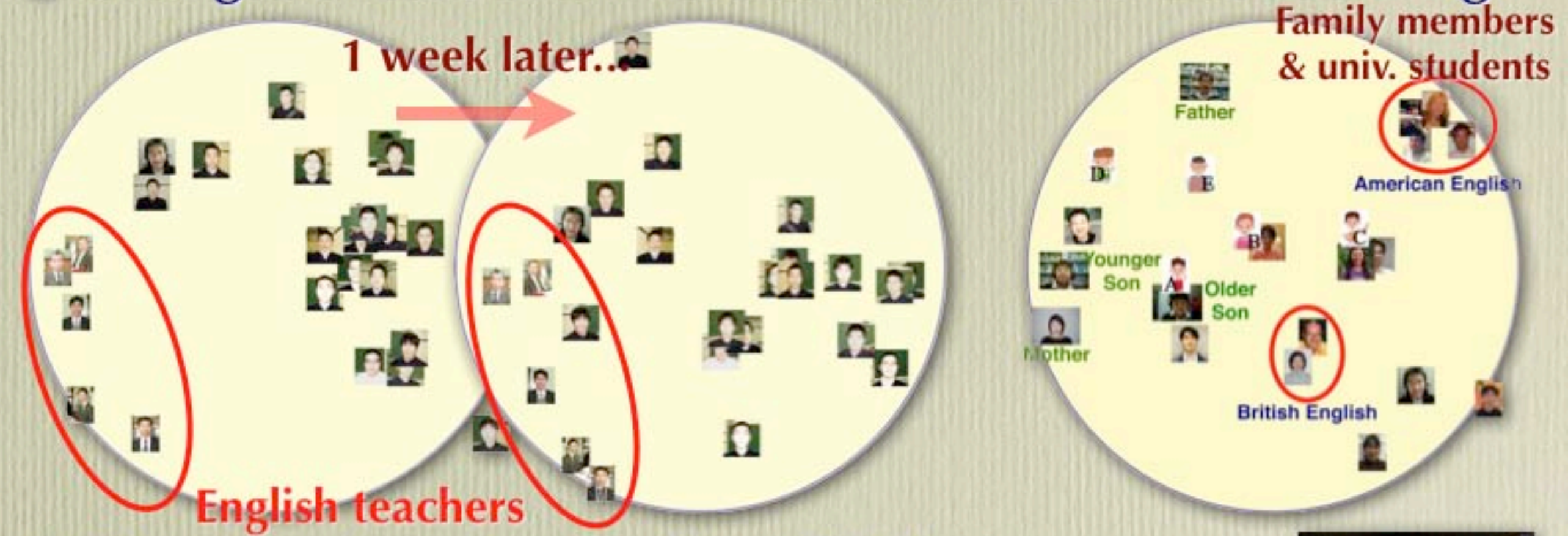
- Manual classification of 96 learners by a phonetician



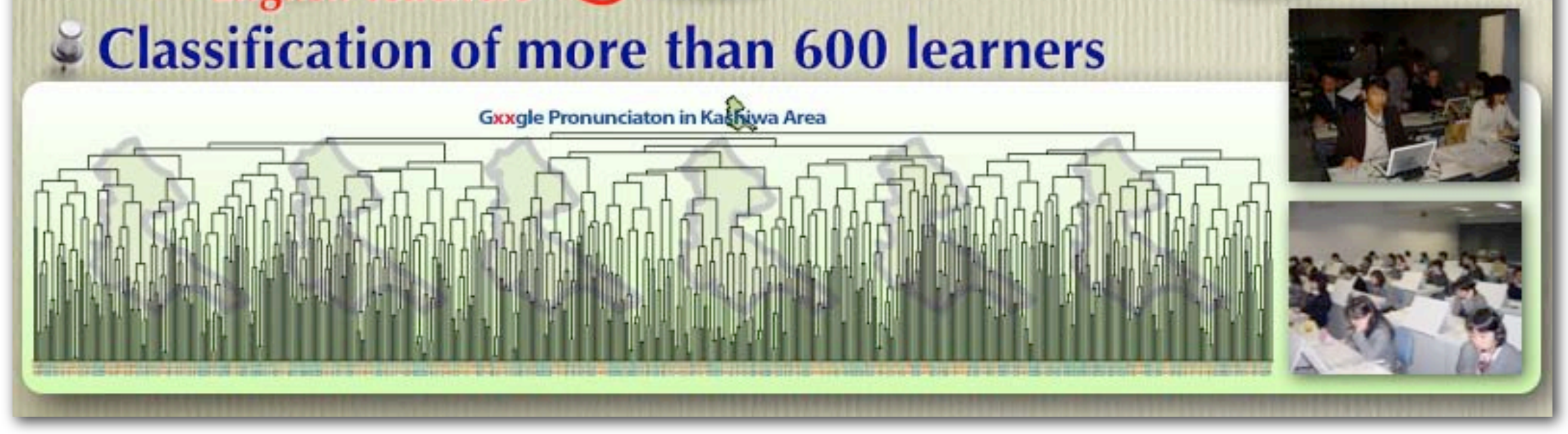
- 8 speakers x 12 pronunciations = 96 simulated learners
  - 1 to 8 = pronunciations, A to L = speakers
  - By substituting J vowels for some E ones, 12 v-systems are defined.

Classification of learners ④⑤

- Changes of students in a class before and after training

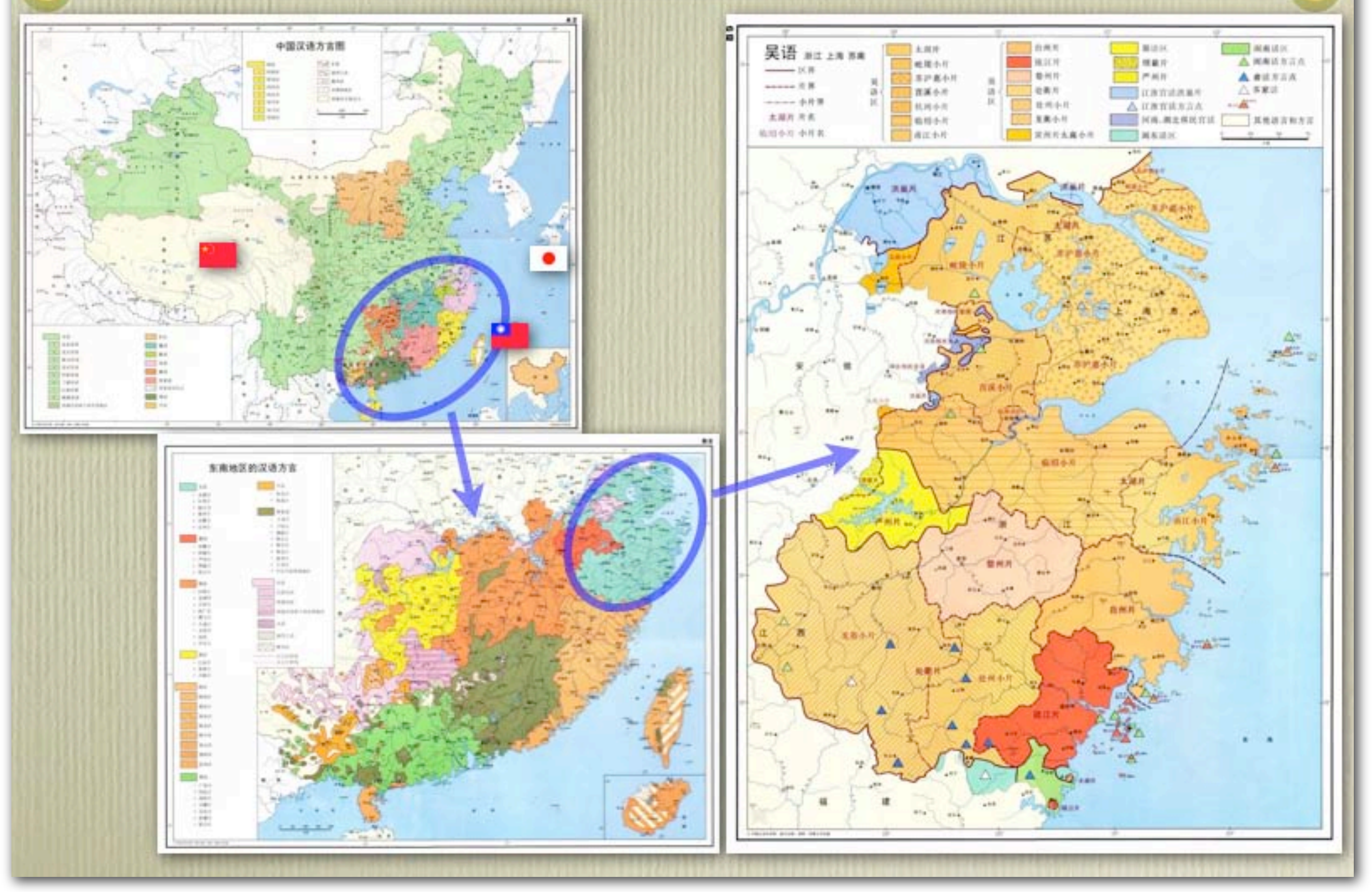


- Classification of more than 600 learners



DIALECT-BASED SPEAKER CLASSIFICATION

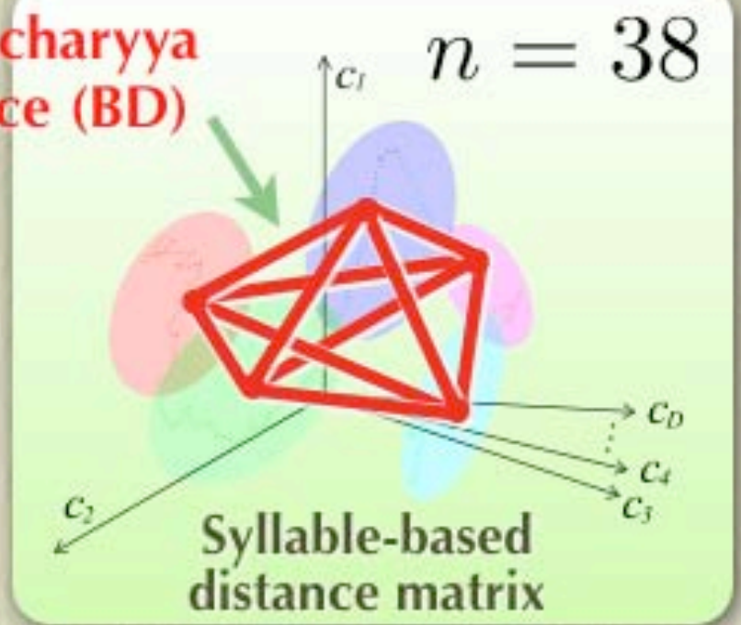
Chinese dialects & sub-dialects



Pronunciation structure of the 38 syllables

- 38 syllables covering all the 38 Mandarin finals

Characters	笔,思,十,耳,五,鱼,阿, 波,饿,哀,悲,早,肉,左, 鸭,哇,别,月,歪,对,腰,牛, 安,烟,弯,捐,恩,彬,温,君, 央,帮,汪,崩,冰,翁,宗,用
Mandarin Pronunciation	/bi/,/ci/,/shi/,/er/,/wu/,/yü/,/a/,/bo/,/e/, /ai/,/bei/,/zao/,/rou/,/zuo/,/ya/,/wa/,/bie/, /yue/,/uai/,/dui/,/yao/,/niu/,/an/,/yan/,/wan/, /juan/,/en/,/bin/,/wen/,/jun/,/ang/,/yang/, /wang/,/beng/,/bing/,/weng/,/zong/,/yong/



- Two definitions of structure-to-structure distance
  - Contrast-based (proposed)
  - Substance-based (conventional)

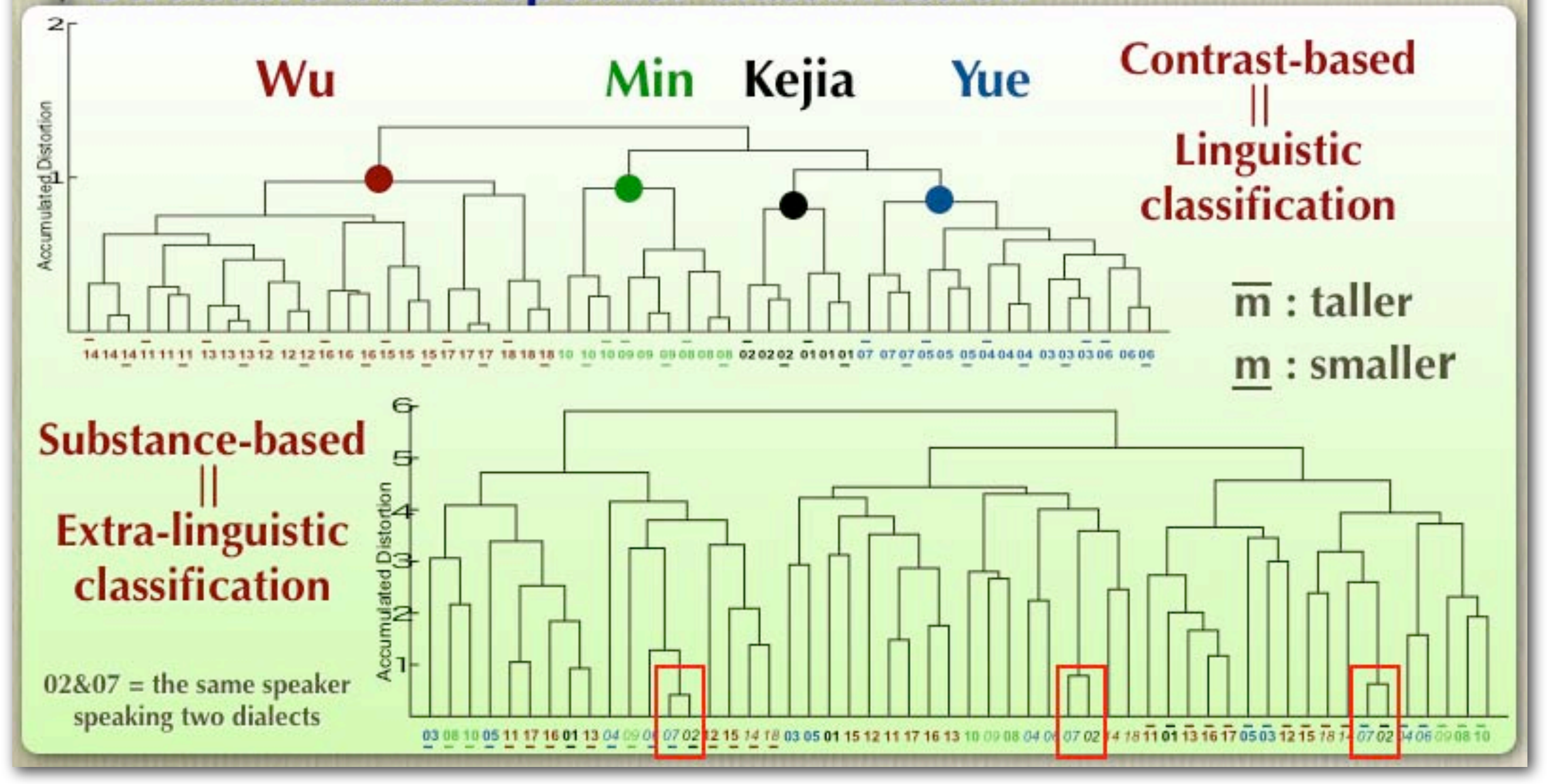
$$D_1(A,B) = \sqrt{\frac{1}{38}\sum_{i<j}(A_{ij}-B_{ij})^2}$$

$$D_2(A,B) = \sqrt{\frac{1}{38}\sum_{i} BD(S_i^A, S_i^B)}$$

A, B = speakers, A_{ij} = distance matrix for A
S = syllable, i, j = syllable index

Classification of Chinese speakers

- Spectrum warping done to increase speaker variability
  - Original speakers(m=18) → taller version(18) + smaller version(18)
- Structure-based speaker classification



The ultimate goal of these studies