Pronunciation Clinic

Which part of your pronunciation to correct first to become like your model speaker?

N. Minematsu[1], K. Kamata[1], M. Takazawa[1], S. Asakawa[1], T. Makino[2], K. Takeuchi[1], Y. Yamauchi[3], T. Nishimura[1], K. Hirose[1]
([1]Univ. of Tokyo, [2]Chuo Univ., [3]Tokyo Int. Univ.)

Which aspect in teachers' voices to imitate?

Two kinds of vocal imitation

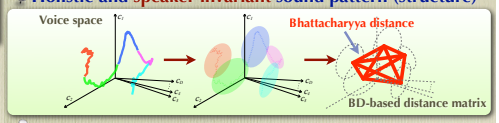
- Vocal imitation performed by myna birds
  - Imitation of absolute properties of voices (sounds)
  - Hearing an adept myna bird say something, one can guess its owner.
- Vocal imitation performed by learners and children
  - Imitation of **not** absolute properties of voices (sounds)
  - **Phoneme-based (string-based) imitation is difficult for young children.**

What in voices to imitate acoustically?

- The shape of the vocal tract differs among speakers.

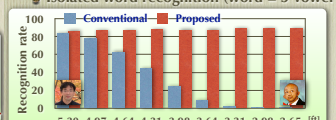
Our answer to the hard question

Holistic and speaker-invariant sound pattern (structure)



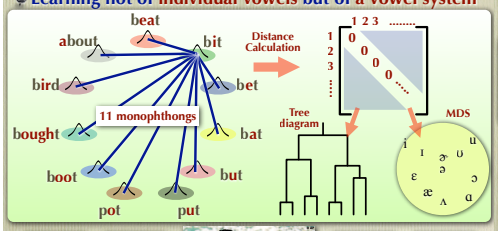
Use of structures for automatic speech recognition

- Isolated word recognition (word = 5 vowel sequence, e.g. /aeoui/)



A vowel training system for everybody!!

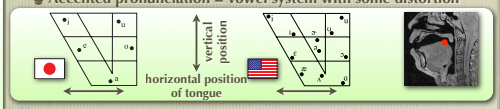
Learning not of individual vowels but of a vowel system



Structural representation of the vowel system

Vowel system and vowel chart

- Accented pronunciation = vowel system with some distortion

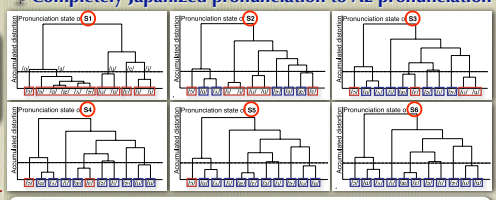


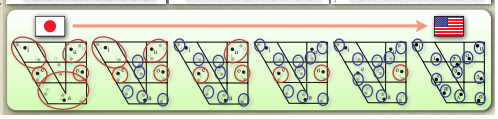
What's possible in the proposed demo system

- 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

Completely Japanized pronunciation to AE pronunciation





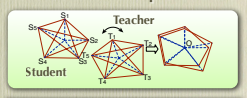
Which vowels to correct at first in your case?

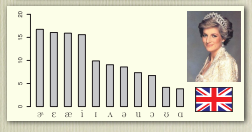
Who is your model speaker?

- A famous phonetician, a movie star (character) or a sport player??

Which vowels to correct at first to become like him/her?

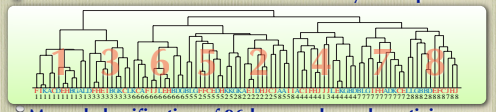
- The system can show the shortest cut to the model pronunciation.



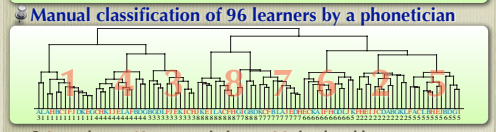


Classification of learners

Automatic classification of 96 learners by a computer



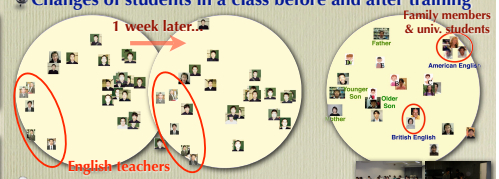
Manual classification of 96 learners by a phonetician



- 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 all the learners on earth!?

Changes of students in a class before and after training



Classification of more than 600 learners

Automatic Pronunciation Evaluation of Japanese EFL Learners' Utterances Generated through Shadowing

By Dean Luo, Nobuaki Minematsu (The University of Tokyo)
Yutaka Yamauchi (Tokyo International University)
and Keikichi Hirose (The University of Tokyo)

1

Background

- What is shadowing?
  - "Repeat-after-me" type exercises that require learners to reproduce nearly at the same time.
  - Improve both listening and speaking skills in language learning. (Tamai 2001)
- Challenges in evaluating utterances in shadowing
  - Speaking style is very different from read speech.
    - Pronunciation often becomes very inarticulate and unintelligible especially in case of beginners.
  - Manual scoring is very time-consuming for teachers

2

Automatic Scoring based on HMMs

- GOP (Goodness of Pronunciation) scoring
  - Based on HMM likelihood ratio.
  - Requires acoustic models of the target language and transcription.

$$GOP(p) = \frac{1}{D_p}\log\left(P(p \mid O^{(p)})\right) \quad (1)$$

$$= \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)P(p)}{\sum_{q\in Q} P(O^{(p)} \mid q)P(q)}\right) \quad (2)$$

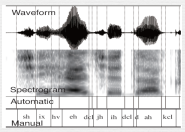
$$\approx \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)}{\max_{q\in Q} P(O^{(p)} \mid q)}\right) \quad (3)$$

$P(p|O^{(p)})$ is the posterior probability that the speaker uttered phoneme p given $O^{(p)}$
Q is the full set of phonemes.
D_p is the duration of phoneme p.

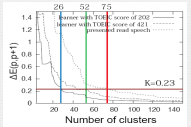
3

Automatic Scoring Based on Clustering

- Clustering-based scoring
  - Unsupervised phoneme segmentation based on time-constrained bottom-up clustering
    - Number of clusters can be used as an indicator of articuratory efforts (how articulate the pronunciation is)
  - No acoustic models or transcription is needed
  - Language independent



An example of unsupervised phoneme segmentation

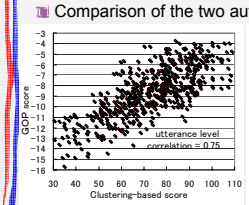


Unsupervised phoneme segmentation on shadowing productions and presented native speech

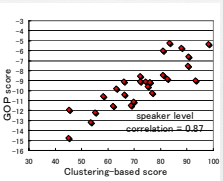
4

Correlation between two automatic scores

- Comparison of the two automatic scores



Correlation at utterance level

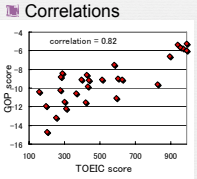


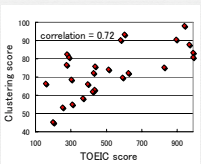
Correlation at speaker level

5

Correlations between automatic scores and TOEIC scores (27speakers)

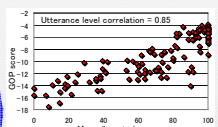
- TOEIC score
  - Test of English as International Communication (including listening and reading part)
- Correlations

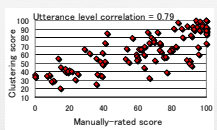


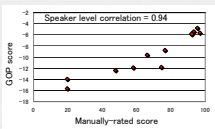


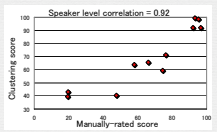
6

Correlations between automatic scores and manually-rated scores (11 speakers)









7

Conclusions

- High correlations between automatic scores and TOEIC scores or manually-rated scores in shadowing.
  - Much higher than correlations in recently reported works on read speech evaluation (Chandel et al,2007)
    - Shadowing might pose a cognitive load on learners adequately
- High correlations between unsupervised clustering-based scores and supervised GOP scores
  - Language-independent clustering-based method is still available for evaluation
    - Any languages, any speaking styles
    - Low cost, high availability

8

Future Works

- Compare shadowed speech with read speech of the same speakers.
- Integrate both techniques
  - Develop a hybrid system with more reliability
- Compare various shadowing tasks
  - Text contents
  - Speed of presented speech
  - Accent and speaking style of presented speech

9