Development of an Automatic Evaluation System of Reading-aloud and Shadowing Performances

Yutaka Yamauchi
(Tokyo International University)

Nobuaki Minematsu
(University of Tokyo)

Outline

1. Assessing reading-aloud performances
2. Evaluating shadowing performances
   A) Manual evaluation method
   B) Automatic evaluation system
3. Demonstration (Volunteers are welcome!)
4. Future tasks

Difficulty of Reading aloud

- L2 reading aloud is difficult for L2 learners.
- Learners are required to do two things almost simultaneously
  - ① decode and comprehend written information
  - ② produce sounds with accurate pronunciation, rhythm and intonation suitable to the text
  - – using orthographical, phonological, semantic and syntactic knowledge

Role of repetition

- If learners repeat the same passage again and again, their reading-aloud performances become smoother and more fluent.
- L2 reading-aloud performance makes progress as repetition times increase.
- How can this progress be assessed?

Assessing reading-aloud performances

1. Overall impression of the rater
2. Discrete points
   - Pronunciation
   - Rhythm/Intonation
   - Suitable segmentation (taking breaths)
   - Fluency
   - Others

Aims & Research Questions

1. Does repetition improve both fluency and pronunciation of reading-aloud performances?

2. Are any differences observed on the improvement as proficiency levels of readers increase?

Method - Procedure

- Participants were requested to record the target passage three times in as accurate and fast a way as possible.
- The native speaker's model reading was not presented to the participants while recording so that they could read at their own pace.
- The first and last reading-aloud recordings were compared in terms of pronunciation accuracy and fluency.

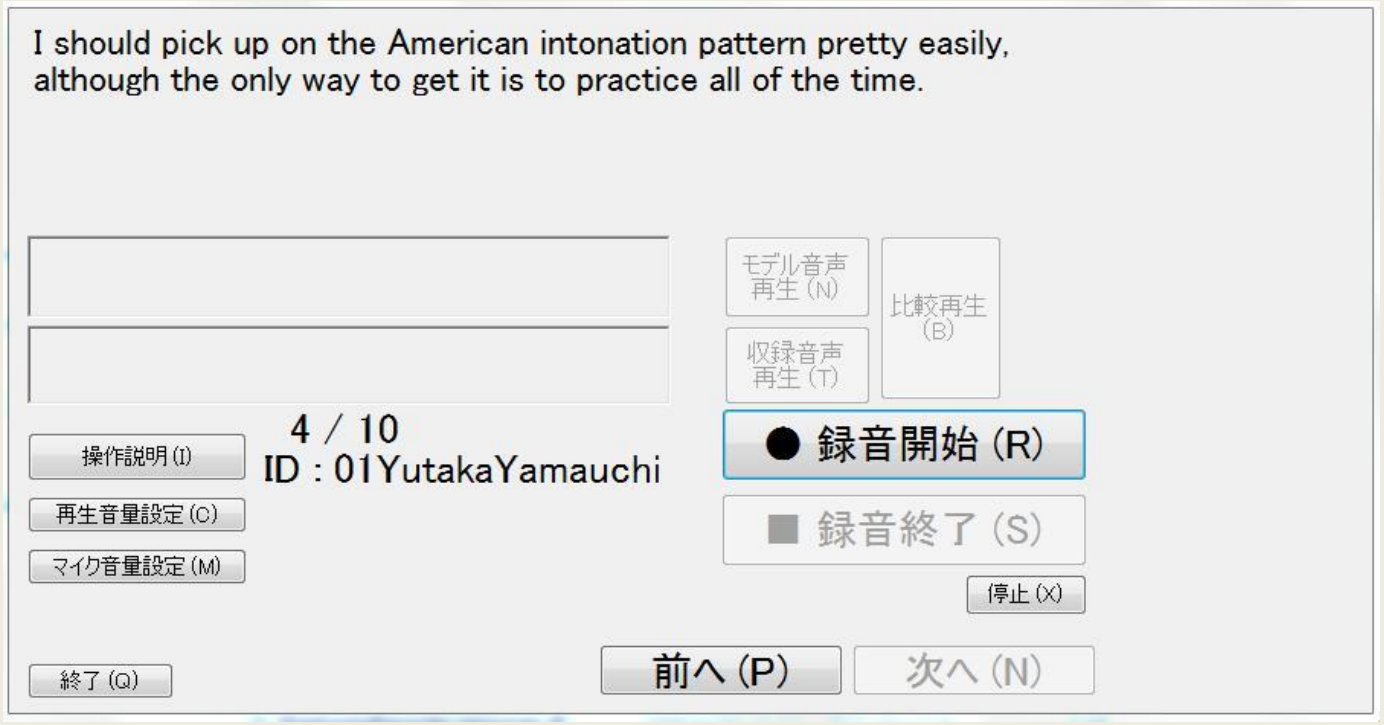
Material (high low)

- English passage of 112 words, the readability of which is 76.2 in Flesch Reading Ease and 5.6 in Flesch-Kincaid Grade Level 5.6

- 1*Hello, my name is Ann.* 2 *I'm taking American Accent Training.* 3 *There's a lot to learn, but I hope to make it as enjoyable as possible.* 4*I should pick up on the American intonation pattern pretty easily, although the only way to get it is to practice all of the time.* 5 *I use the up and down, or peaks and valleys, intonation more than I used to.* 6*I've been paying attention to pitch, too.* 7 *It's like walking down a staircase.* 8 *I've been talking to a lot of Americans lately, and they tell me that I'm easier to understand.* 9*Anyway, I could go on and on, but the important thing is to listen well and sound good.* 10*Well, what do you think? Do I?*

Sound recording system developed by Minematsu Lab

- Learners can record their reading while looking at the script on the PC monitor.



Proficiency test

- TOEIC: Test of English as International Communication
- Listening section of 100 questions (Full score 495)
- Reading section of 100 questions (Full score 495)
- Total full score 990

IELTS	TOEIC
9.0	-
7.5~8.0	990
7.0	880
6.5	800
6.0	730
5.5	645
5.0	590
4.5	500
4.0	450
3.5	300

Participants

- Subjects: 36 Japanese EFL learners with TOEIC scores ranging from the 200s to the 900s

	High-proficiency group	Middle-proficiency group	Low-proficiency group
TOEIC score	900–730	720–590	580–220
Average score	875	662	385
Number	12	12	12

Method – pronunciation

- Pronunciation accuracy was measured by GOP (Goodness of Pronunciation)

(Witt & Young 2000)

$$GOP(p) = \frac{1}{D} \log(P(p \mid O))$$

GOP (Goodness of Pronunciation) scoring

Based on HMM likelihood ratio (Witt and Young, 2000)

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

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

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

$P(p \mid O^{(p)})$: the posterior probability that the speaker uttered phoneme p given

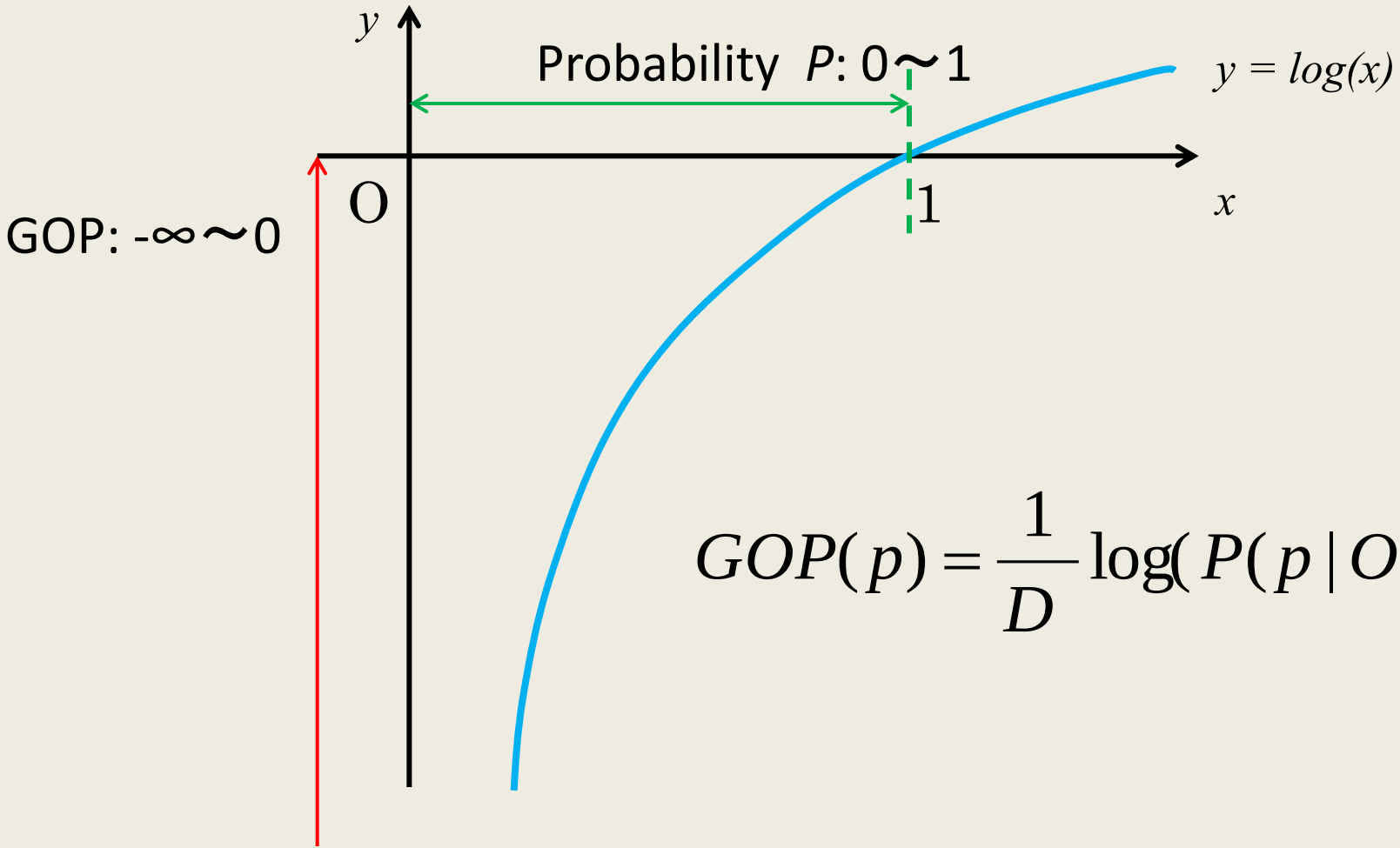
Q : the full set of phonemes

D_p : the duration of phoneme p

Numerator of Eq.3 : calculated by scores generated during forced alignment

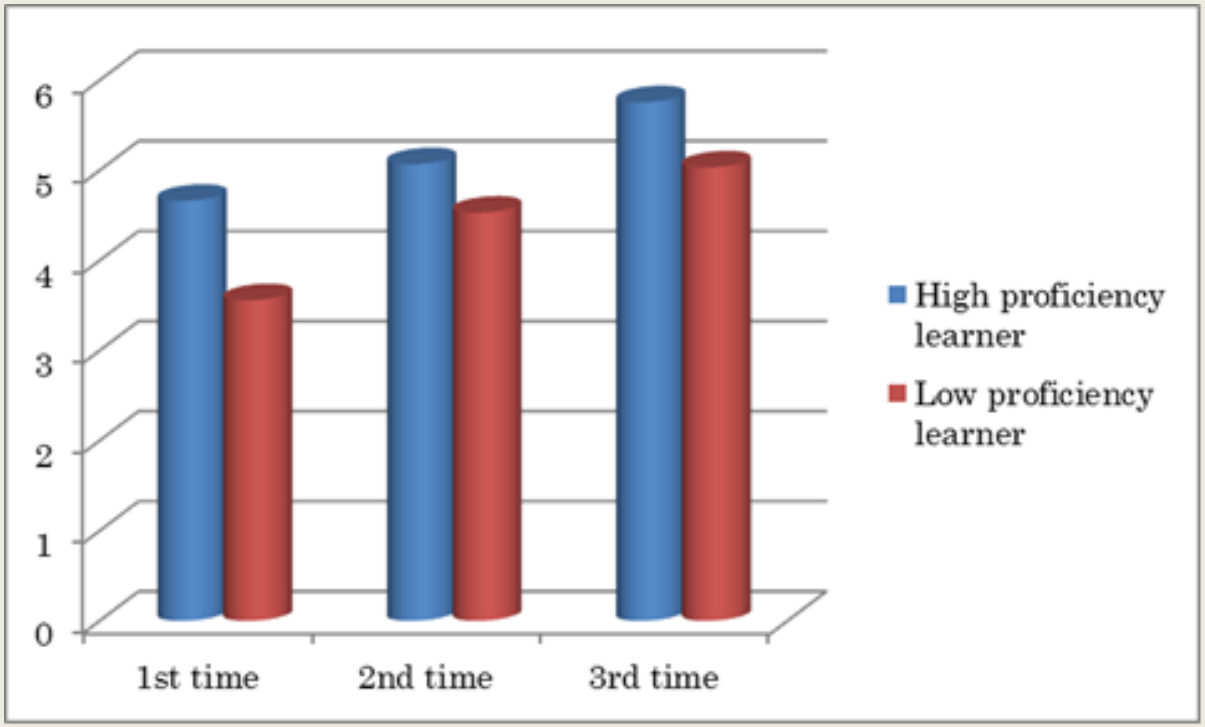
Denominator: approximately attained by continuous phoneme recognition

GOP (Goodness of Pronunciation)



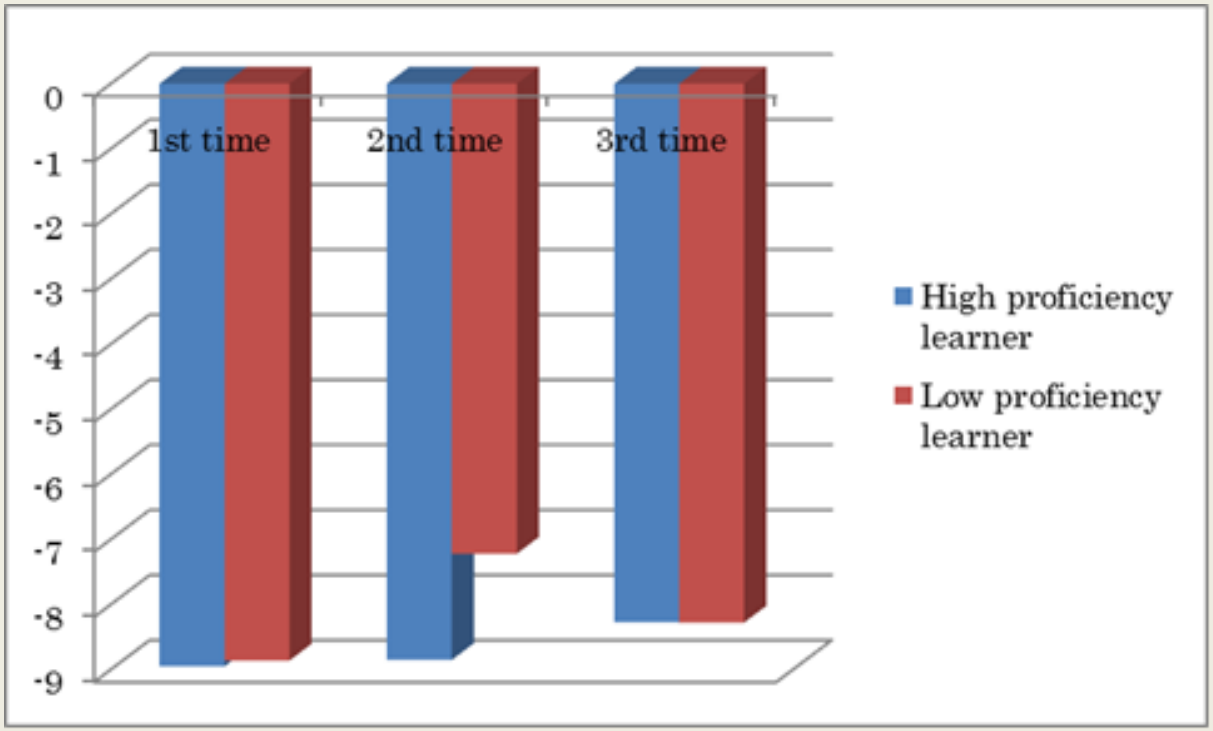
$$GOP(p) = \frac{1}{D}\log(P(p \mid O))$$

How ROU changes according to times of repetition between two proficiency learners



- Repetition of reading aloud the same passage tended to increase the ROU (rate of utterance) values. This shows the reader's fluency tended to go up as repetition times increased in the cases of both the high and low proficiency learners.

How GOP changes according to times of repetition between two proficiency learners



- Repetition of reading aloud the same passage tended to see the GOP values get close to zero. This means the reader's pronunciation tended to get more accurate as times of repetition increased, especially in the case of the high proficiency learner.
- The GOP difference between the two proficiency groups was not so large as the ROU difference.

Results and discussion (1)

(1) Repetition seems to be more effective for the high proficiency learners.

- They can use their knowledge, and their sentence processing gets faster and more automatized as repetition times increase.

(2) The ROU difference between the two proficiency learners was larger than the GOP difference.

① ROU seems to be more valid as a measurement of reading-aloud performances than GOP.

② Pronunciation quality seems to remain more stable than speech rate.

③ Explicit explanation and practice may be more necessary to improve pronunciation quality.

Results and discussion (2)

- Reading aloud without model reading is difficult for low proficiency learners.
- Their vocabulary size is very limited. Even in the case of known words, the triangular linkage between spelling, meaning and pronunciation is not completely established.
- Pronunciation model or model reading should be presented to low proficiency learners in an effective way.
- Shadowing while looking at text may be one of the effective ways to present model reading to low proficiency learners.
- Why is shadowing effective in improving reading-aloud performances?
- How can shadowing performance be assessed using the latest speech technology?

Shadowing

- Online task of repeating heard speech (Marslen-Wilson 1973)
- Originally used for simultaneous interpretation training
- In simultaneous interpretation training learners are required to input the message in L2 , change and output it in L1 simultaneously

How to use shadowing in classroom

- In ESL/EFL classrooms
  learners are required to input the message in L2 , comprehend and output it in L2 simultaneously while imitating the model's pronunciation, rhythm and intonation as much as possible

- Shadowing training is expected to improve learners' pronunciation and reading –aloud performances as well as listening comprehension and speaking skills

Types of shadowing

- Shadowing practice has been widely used in EFL classrooms in Japan from junior high school to university levels

- There are two types:
  ① prosodic shadowing
  ② content shadowing

Prosodic shadowing

- Learners repeat and imitate the model

  but don't have to understand its message
- It is good in improving pronunciation and prosodic aspects
- This practice lacks understanding messages
- Parrots or kids before the critical period might perform better than adult learners
- So it is not enough for ESL/EFL practice aimed at improving communication skills with message interaction

Content shadowing

- learners repeat and imitate the model while understanding the message and retaining the main points in their memory
- It is good as an ESL/EFL practice aimed at improving communication skills
- But it is very difficult and demanding because its cognitive load is very high
- Content shadowing using a passage never read before can reveal learner’s problem-solving skills with their overall skills = overall proficiency

Shadowing & Working memory

- Content shadowing requires learners to effectively use their working memory (WM).
- WM enables learners to decode and store the auditory information.
- Effective use of WM is reported to correlate with learners' proficiency levels (Gathercole and Baddeley 1993) .
- Content shadowing is expected to reflect their overall proficiency.
- Shadowing may be used as an overall proficiency test in terms of WM.

Assessing shadowing performances

Shadowing performances have been assessed based mainly on:

1 Overall impression of the rater

2 Ratio of words correctly reproduced in the target passage

Overall impressionistic method

- Procedure
  - The rater listens to the recorded speech again and again
  - Assesses it on a five-scale basis
- Merit:
  - It reflects veteran instructors' experiences
- Demerit:
  - Subjective → Large discrepancy between raters

Reproduction ratio method

- Procedure
  - The rater listens to shadowing performances again and again while looking at the original script
  - Counts the number of words correctly reproduced
  - Sums up the number of words in the passage and calculates the ratio
- Merit:
  - Relatively objective
- Demerit:
  - Energy-and time-consuming
  - Smaller discrepancy between raters
  - Focuses only on words and ignores prosodic features (rhythm, intonation, etc.)

Automatic scoring method

- To solve the manual assessing problems
- Automatic assessing system has been newly developed
  - Collaborative project with an engineering researcher named N. Minematsu (University of Tokyo)
  - Uses the latest speech information processing technology: GOP (Goodness of Pronunciation)
  - Compares actual pronunciation in the shadowed speech with pronunciation based on the HMM (Hidden Markov Model) stored in the program on a word-by-word basis
  - Automatically computes the distance between the two
  - Presents numerical scores
- Merits:
  - Highly objective, time-and energy saving
- Demerits:
  - Ignores prosodic features (rhythm, intonation, etc.)

Aim of this study

To investigate

① the relationship between manual and automatic scores of shadowing performances

② the relationship between shadowing scores and overall proficiency scores

③ the possibility of content shadowing as an overall proficiency test

④ how much GOP scores can cover prosodic features of shadowing performances

Research questions of this study

① Are manual scores by veteran language instructors highly correlated with automatic scores using GOP?

② Are shadowing scores highly correlated with overall proficiency scores ?

③ Are almost the same results obtained from experiments with different passages and learners?

④ Are GOP scores highly correlated with manual scores on prosodic features by veteran language instructors?

Materials (passage)

- Consists of 143 words
- Read by an American native speaker at 130 words per minute.
- Never presented to the participants before

- The MacDonald's house has been broken into. A policeman has come to check it out. He finds a boy standing nearby. The policeman is now talking to the boy. He wants to know how the door of the MacDonald's house was broken open. The boy said that it had already been broken before he and his friend went to the house. He said that they simply walked into the house. The police officer asked, "Why were your fingerprints found all over the door? And why were your boots scratched? It was you who kicked the door open, wasn't it? Why did you steal the stereo and the CDs? Did you just want to have a bit of fun, or were you trying to get some money? Now then, tell me the truth. I don't want to hear any more of your lies."

Materials (Questions)

- Seven comprehension questions were given after shadowing
  - Related to the main points of the passage
  - Given in a multiple-choice form (receptive form) to reduce their memory load
  - Given in L1 to reduce reading difficulty in L2
- (Ex.) What happened at the MacDonald's house?
  a. A fire broke out in her house.
  b. Thieves entered her house.
  c. The police inspected her house.
  d. The boys were invited to her house.

Participants

- Subjects: 36 Japanese EFL learners with TOEIC scores ranging from the 200s to the 900s

	High-proficiency group	Middle-proficiency group	Low-proficiency group
TOEIC score	900–730	720–590	580–220
average score	875	662	385
number	12	12	12

Data collection procedure

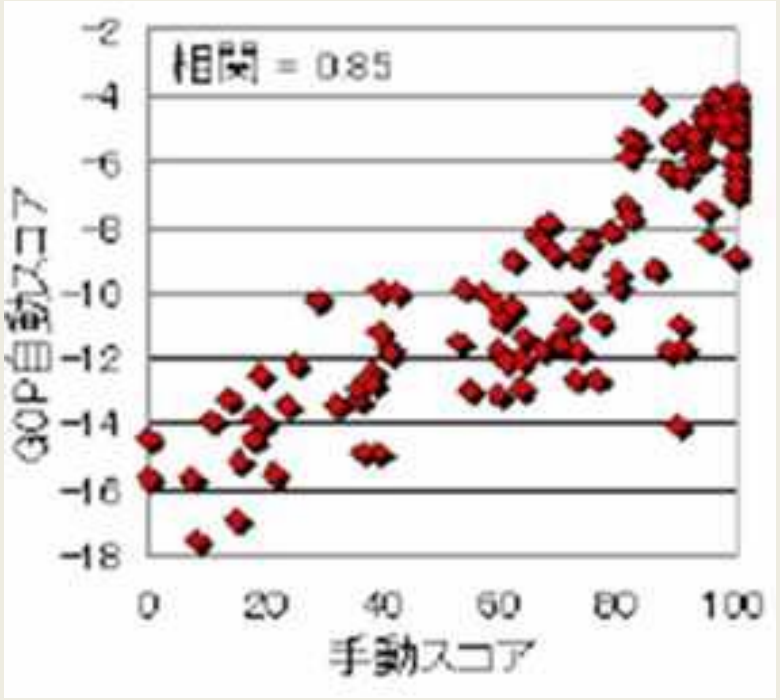
- Headset test and warm-up activities in the CALL classroom
- Practice shadowing using an easier and shorter passage with fewer multiple-choice questions given afterward
- Recording session
- 36 participants recorded their shadowing into the PC while listening to the passage
- They answered seven multiple-choice questions related to the main points of the passage

Assessment procedure

- Four veteran instructors

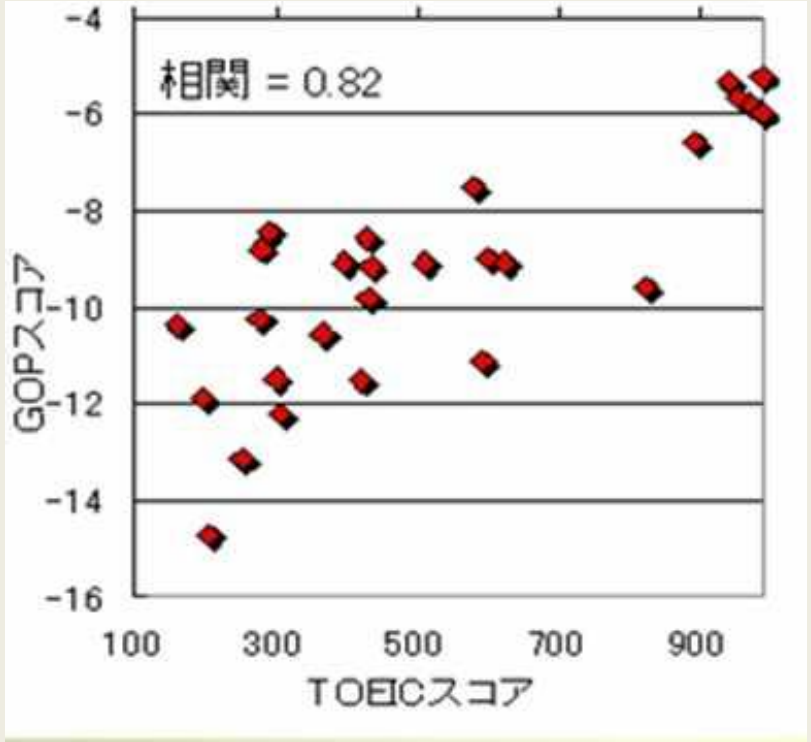
  listened to the shadowed speech
- Firstly, assessed them based on their overall impression on a five-scale basis
- Secondly, assessed the same speech based on the reproduction ratio method
- Pearson correlation coefficient was computed between the three scores:
  - Proficiency scores by TOEIC
  - Shadowing scores by overall impression
  - Those by correctly reproduced ratio

RQ1 manual vs automatic scoring



Manual and automatic scores

r=0.85



TOEIC and automatic scores

r=0.82

RQ2 Correlation

- Shadowing scores measured by impression and correctly reproduced ratio were highly correlated with proficiency scores by TOEIC ($p < .01$)

	Impression	Repro Ratio	TOEIC	Comprehension
Impression		.905	.897	.768
Reproduction Ratio	.905		.769	.682
TOEIC	.897	.769		.806
Comprehension	.768	.682	.806	

RQ 3 Reliability

Results of experiments with different materials and EFL learners

	Correlation between manual and automatic scores	**Correlation between TOEIC and automatic scores**
Experiment 1	r=.85	r=.82
Experiment 2	r=.72	r=.68
Experiment 3	r=.84	r=.90

RQ4 Coverage of GOP on prosodic features

Reproduction ratio & prosodic features

Variables	Correlation
Reproduction ratio & manual scores	.717
Reproduction ratio & automatic scores	.716
Prosodic features & reproduction ratio	.578
Prosodic features & automatic scores	.722

Coverage of automatic assessing toward prosodic features = 52.13 % (square of r)

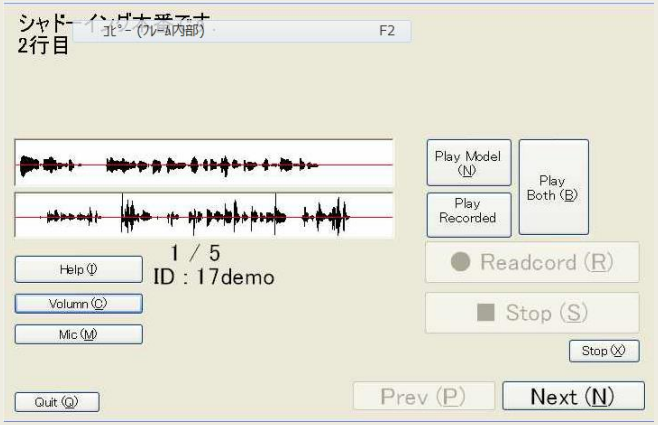
Practicality of automatic assessing

- Ease of use
  - An average PC and a headset
  - User-friendly interface
- Flexibility in choosing suitable materials
  - Text files & audio files (from Web sites, DVDs, etc)
  - Choice, change and adjustment of materials suitable for learners' levels, interests and concerns

e.g. From elementary or junior high textbooks to the USA presidential speech

Practicality of automatic assessing

- The learner's self-evaluation
  - Comparison of sound waves and actual sounds between the learner's shadowing and the model speech
  - Self-recognition of the learners' own shortcomings



Practicality of automatic assessing

- Capability of a proficiency test or placement test
  - High correlation between shadowing and TOEIC scores
  - Prediction of TOEIC scores based on shadowing performances
  - Use of content shadowing as a proficiency test
    → placement test
    → pass-or-fail test = entrance examination

Validity, Reliability＆Practicality

- Confirmation of automatic assessing in three respects:
  - ① Validity (Comparison with manual and TOEIC scores)
  - ② Reliability (Comparison with the results of a number of experiments)
  - ③ Practicality (Ease of use and flexibility in choosing materials suitable for the learner’s levels , interests and concerns)

Summary and Conclusion

- Assessing shadowing manually contains several problems
  → Subjective, time- and energy-consuming
- Shadowing scores assessed manually are highly correlated with proficiency scores measured by TOEIC
- Shadowing scores reflect learners' proficiency and could be used as a proficiency test
- Automatic assessing system using GOP
  - High correlation with manual and TOEIC scores
  - Almost the same results in a number of experiments
  - Ease of use and flexibility in choosing materials

  → Confirmation of validity, reliability and practicality

Demonstration of this system

- Automatic scores are highly correlated with manual and TOEIC scores
- Overall proficiency scores can be predicted by shadowing scores based on the regression formula: $Y = a X + b$

  (Y: Proficiency scores X: shadowing scores)
- Shadowing can be used as a placement test

Future tasks

- Adding functions of assessing prosodic features directly
- Applying the basic concepts to automatic evaluation system for other languages like Japanese, Chinese and so on
- Developing tablet or smart phone versions
- Investigating acquisition process of prosodic features through shadowing
- Examining transforming process from controlled to automatized processing through shadowing

Acknowledgements

- Co-researchers:
  - Kay Husky (Tokyo International University)
  - Akemi Kawamura (Tokyo International University)
  - Megumi Nishikawa (Tokai University)
  - Shuhei Kato (Hoya Service Corporation)
  - Masaya Fujita (Hoya Service Corporation)
- MEXT/JSPS KAKENHI Grant Number 25580135

Main references

- D. Luo, N. Shimomura, N. Minematsu, Y. Yamauchi, and K. Hirose (2008) “Automatic pronunciation evaluation of language learners' utterances generated through shadowing," *Proc. INTERSPEECH*, 2807-2810.

- Y. Yamauchi, N. Minematsu, D. Luo, A. Kawamura, and M. Nishikawa (2011) “A study on the validity of automatic evaluation of shadowing performances by Japanese EFL learners in comparison with discrete-point evaluation,” *Proc. The 51st Annual Conference of the Japan Association for Language Education and Technology,* 62-63.

- Y. Yamauchi, N. Minematsu, D. Luo (2009) “Automatic assessing system of shadowing performances: Can it predict TOEIC scores?” *Extensive listening and reading magazine,* Vol. 17, 30-32.

- Y. Yamauchi, N. Minematsu, A. Kawamura, M. Nishikawa, M. Fujita, and S. Kato (2013). Effects of repetition on reading-aloud performances of EFL Japanese learners. *Proceedings of the Conference of the Japan Association for Language Education and Technology (the Kanto Chapter)*, 22-23.

- M. Witt and S.J. Young (2000) "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication,* 30, 95-108.