The First Step towards Automatic Quality Evaluation of Chinese Vowel Pronunciations for Foreign Learners for Self-training

Junya Shinzawa, Shumei Chen, Jinhua She, Hiroyuki Kameda, and Sumio Ohno

***Abstract*—Nowadays, computer-assisted language learning (CALL) systems are widely used for language education. Since the pronunciation of Chinese is difficult, it is important to build a system to evaluate a learner's pronunciation in a real-time fashion so as to maintain the motivation of learning. This study tried to develop such a system that not only judges pronunciations from the viewpoint of acoustic phonetics, but also provides a learner an advice on improving his/her pronunciations.**

As the first step, we built a system for Chinese monophthong vowels, and analyzed the acoustic features of the pronunciations between Chinese and Japanese. The results show that it is possible to distinguish the characteristics of pronunciation using the formant frequency, which is one of the acoustic features; and it is also possible to distinguish round and unround lips, which has been difficult, by using three kinds of formant frequencies from the first formant to third.

***Index Terms*—CALL (computer-assisted language learning) system, e-learning, formant frequency, Chinese vowel.**

I. Introduction

Nowadays, CALL (Computer Assisted Language Learning) systems [1] are widely used for language education, and smartphones and personal computers are used as terminals to access a virtual learning environment. Since English is practically the first foreign language in Japan, many CALL systems have mainly been built to learn English in Japan. While CALL has also been used for other languages in the last one or two decides, the sources are much limited compared to English.

Chinese is now the second foreign language in Japan. The population of students who learn Chinese has been increasing along with the economic development in China. On the other hand, the pronunciation of Chinese is very difficult, particularly for Japanese. This tends to lead to discomfiture in Chinese learning for many students. To avoid this problem, we developed a system to evaluate a learner's pronunciation in a real-time fashion for self-learning so as to maintain the motivation of learning. This system not only judges pronunciations from the viewpoint of acoustic phonetics, but also provides a learner an advice on improving his pronunciations.

Manuscript received June 25, 2018; revised August 3, 2018. This work was supported in part by the Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science (JSPS) under Grant 17K0294

The authors are with Tokyo University of Technology, Tokyo, Japan (e-mail: g2117013f5@edu.teu.ac.jp, {chin, she, kameda, ohno}@stf.teu.ac.jp).

In this paper, we analyze the acoustic features of Chinse vowels pronounced by native Chinese speakers, and compared them with those collected from Japanese learners. The results show that it is possible to distinguish the characteristics of pronunciation using the formant frequency, which is one of the acoustic features; and it is also possible to distinguish round and unround lips, which has been difficult by using three kinds of formant frequencies from the first formant to third.

II. Preparing of Utterance Material

In this study we first focused on six monophthong vowels in Chinese as the first step to evaluate the pronunciation quality of Chinese vowels.

The IPA (International Phonetic Alphabet) [2], which was established to characterize sounds of all languages, is used to classifies vowels in languages around the world according to the following three aspects;

1) Openness of mouth,
   - Open
   - Open-mid
   - Close-mid
   - Close
2) Tongue position, and
   - Front
   - Central
   - Back
3) Lips rounding.
   - Round
   - Unround

According to the introduction of Chinese monophthong vowels in [3] and based on the IPA, we easily classified the six Chinese monophthong vowels as shown in Table I and arranged them in Fig. 1.

As a speech material, we first collected the pronunciations of Chinese vowels from native Chinese Mandarin speakers (seven males and six females), who all spoke Mandarin Chinese.

TABLE I: Characteristics of Chinese Monophthong Vowels

Vowel	Openness	Backness	Rounded
a	Open	Front	Unround
o	Close-mid	Back	Round
e	Close-mid	Back	Unround
yi	Close	Front	Unround
yu	Close	Front	Round
wu	Close	Back	Round

doi: 10.18178/ijiet.2019.9.1.1176

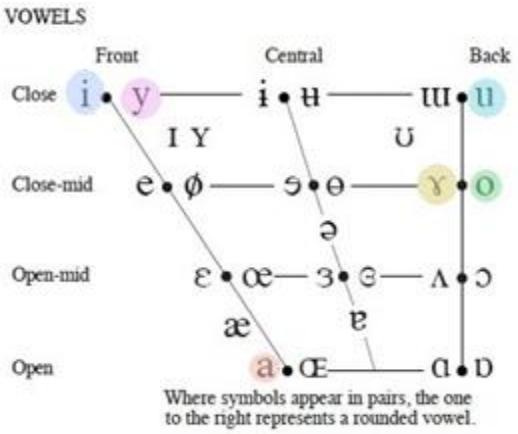

Fig. 1. The arrangement of Chinese vowels based on IPA [2].

Utterances were collected from recorded materials attached with Chinese learning textbooks. For comparison and quality evaluation, utterances from Chinese beginners, who were Japanese native speakers, were also recorded (seven Japanese male university students). They were given a short-term pronunciation lecture before recording by a well-experienced Chinese teacher, and after that they immediately uttered the six monophthong vowels. In the pronunciation lecture, the teacher gave guidance for the pronunciations and carried out utterance practice based on Table II, which is widely used for Japanese learners.

TABLE II: THE CHARACTERISTICS OF THE MONOPHTHONG VOWEL TOLD AT THE TIME OF GUIDANCE

Vowel	Instruction
a	Larger mouth than "a" of the Japanese. Pronounce clearly.
o	Round the mouth than "o" of the Japanese.
e	Mouth of the "e" of the Japanese, they say the "o" in Japanese.
yi	Draw mouth laterally than "i" of the Japanese, Pronounced sharply.
yu	Round and protrude mouth than "u" of the Japanese.
wu	Rolling up your mouth like Japanese "yu", pronounce "i" of the Japanese.

Although the Chinese syllables have different tones, preliminary experiments ensured that the influence of the four tones was small enough in the acoustic features to be noted in this study. For this reason, we only collected and recorded the utterances of tone 1.

III. ANALYSIS OF CHINESE VOWEL PRONUNCIATIONS

Based on the collected and recorded speech materials, we extracted the first and second formant frequencies (F1, F2) as speech features in accordance with Otsuka's study [4]. The formant frequencies are the resonance frequencies (peak of the amplitude spectrum) of the vocal tract, and are sequentially named as the first formant (F1), the second formant (F2), and the third formant (F3) started from the peaks at low frequency. It is known that F1 and F2 correspond to *openness of mouth* and *front-back position of tongue*, respectively [4].

At this time, the formant frequency was extracted using the voice analysis software, *Praat* [5], created by Paul, David et al. The average of the values in all vowel utterances was calculated as the formant frequency of each vowel. The Burg method was used for the formant frequency extraction, where "the maximum formant frequency" was set to be 5000.0 Hz for males and 5500.0 Hz for females according to [5]. "The number of formant" was set to five, and the other parts of extraction settings remained the initial values.

To evaluate the quality of pronunciation, one Chinese professor performed a 5-stage subjective assessment of 1 (poor) to 5 (good) for the vowel utterances of Japanese learners.

The plots of F1-F2 of a monophthong vowel uttered by Chinese male native speakers and Japanese learners are shown in Figs. 2 and 3, respectively. Fig. 2 and 3 show an ellipse of 60% probability for each vowel of Chinese speakers. The numbers plotted in Fig. 3 are the aforementioned subjective evaluation scores.

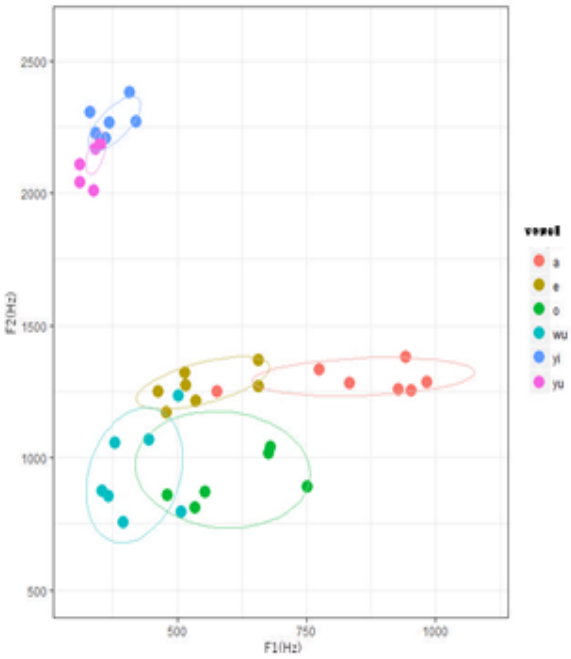

Fig. 2. F1-F2 distribution of monophthong vowel for pronunciation of Chinese male speaker.

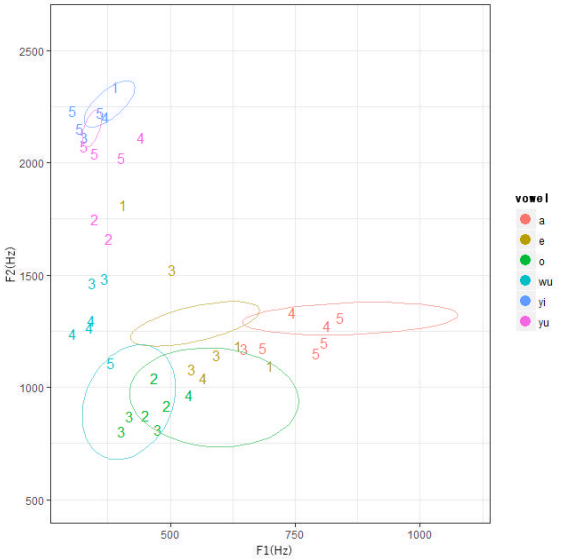

Fig. 3. F1-F2 distribution of monophthong vowel for pronunciation of Japanese male learner (Numbers indicate pronunciation evaluation values).

Most of the pronunciations of Japanese learners were highly evaluated when they were close to the distribution of a Chinese native speaker. However, on the other hand,

- There are some samples with score 5 even outside the probability ellipse;
- There are some samples with score 2 between scores 3 and 4 samples;
- There is a sample with score 1 for "yi" even if it is within an ellipse; and
- Even Chinese speakers do not separated "yi" and "yu" in the F1-F2 space.

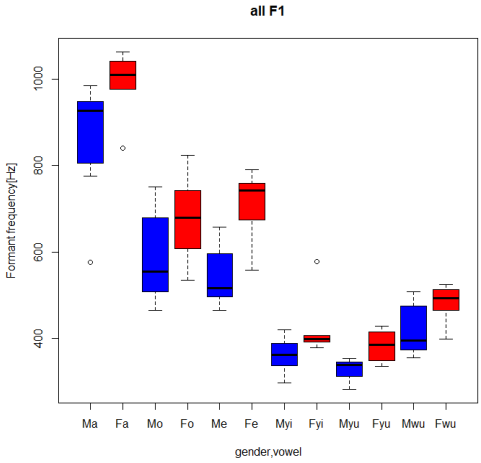

Fig. 4. First formant frequency of Chinese monophthong vowel.

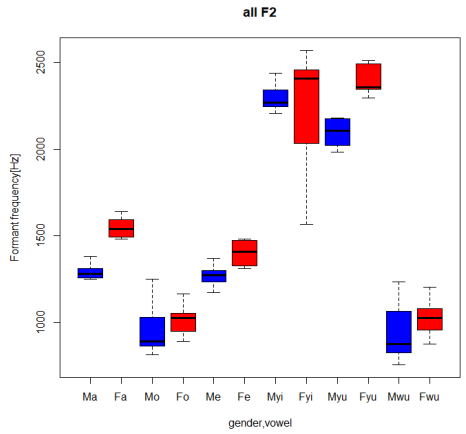

Fig. 5. Second formant frequency of Chinese monophthong vowel.

IV. Distributions of Formant Frequency

Box-plot diagrams were created using the first to fourth formant frequencies for voice collected from Chinese learning materials (see Section II). They were shown in Figs. 4-7, respectively. From the characteristics of Figs. 1, 4, and 5, we observed that

- "a" is a large vowel that makes a mouth widely open and F1 is high.
- "yi", "yu", and "wu" are small vowel utterance that close the mouth (opening small) and F1 is low.
- "yi" and "yu" are front tongue position and F2 is high.
- "o", "e", and "wu" are back tongue position and F2 is low.

Based on these findings, we reached the following conclusions: the first formant has a correlation with the mouth opening and the second formant has a correlation with the position of the tongue. There are two patterns of Chinese vowels, "e" and "o", and "yi" and "yu" that distinguish between round and unround lips.

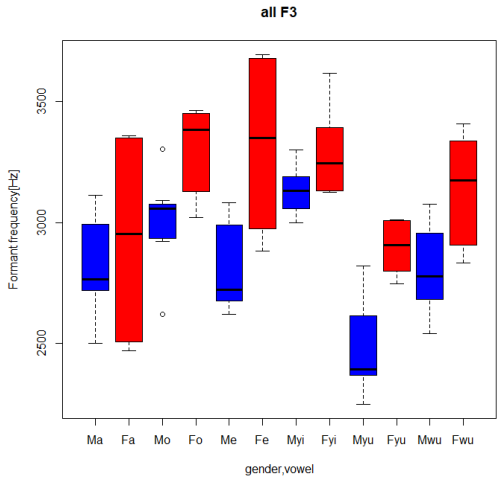

Fig. 6. Third formant frequency of Chinese monophthong vowel.

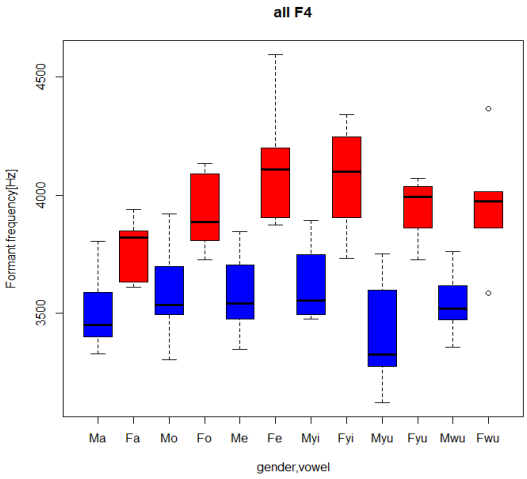

Fig. 7. Forth formant frequency of Chinese monophthong vowel.

As for the first and second formants, we can distinguish them by F2 for the combination of "e" and "o", but there is not a big difference in "yi" and "yu". However, when we looked at the third formant, we saw a tendency that "yi" speech had higher F3 than "yu" did. In "e" and "o", F3 of "o" was higher than that for "e" for male speakers, but there was no big difference for female speakers.

From this fact, it seems that the third formant may be correlated with round/unround lips in pronunciations. This can be said only from limited number of speech material. While it is hard to say it absolutely at this stage, this tendency appears with some speakers. This needs further investigation in the future.

In order to carry out a more precise investigation, we have invited fourteen Japanese university students (five males, nine females) to join our program, and are recording their utterances of Chinese monophthong vowels after their Chinese lessons at the second study term (from September 2018 to January 2019). A total of ten times are recorded for one person, and *Praat* are used to perform formant extraction with the same settings as those used in Section III. We planned to investigate the utterances and to establish rules for learners to improve their pronunciation based on the above

observations. Finally, we try to incorporate the rules in the Chinese course taught at Tokyo University of Technology so as to take full advantage of these results.

V. Conclusion and Future Work

As the first step to build a system that automatically evaluates the pronunciations of Chinese vowels and provides learners an advice to improve their pronunciations, we collected pronunciations of Chinese monophthong vowels from Chinese speakers and Japanese learners, characterized the features of the vowels, and compared the pronunciations between Chinese and Japanese. The results reveal that the evaluation value became higher as the probability ellipse became close, and we need to provide a measure for the distinction. The future work is explained below:

- We do not have enough utterance material of Chinese speakers, and we need to collect and record some to carry out a thorough study.
- Overlaps were observed for some ellipses. It is necessary to find a way to deal with this situation to ensure the recognition processing.
- We used the burg method for formant extraction. It produced unnatural data sometimes for "yi" for females (at the point of "yi" in Figs. 4 and 5). This problem has to be solved.

It is possible to separate speech interval at the stage of analysis and comparison, but it may cause a problem for automatic evaluation. Except for using sound, it is also possible to discriminate between round/unround lips based on image recognition.

References

[1] H.-C. Liao, J.-C. Chen, S.-C. Chang, Y.-H. Guan, and C.-H. Lee, "Decision tree based tone modeling with corrective feedbacks for automatic Mandarin tone assessment," in *Proc. INTERSPEECH*, pp. 602–605, September 2010.

[2] International Phonetic Association. [Online]. Available: https://www.internationalphoneticassociation.org/

[3] S. Chen, *Syabette Iitomo Chinese*, Tokyo: Asahi publishing company, 2016, p. 9.

[4] T. Otsuka, "A trial in utilizing formant values for teaching vowel pronunciation," *Tokyo Women's University Bulletin*, vol. 64, no. 2, pp. 311-333, March 2014.

[5] P. Boersma and D. Weenink. Praat: Doing phonetics by computer [Computer program]. [Online]. Available: http://www.praat.org/

Junya Shinzawa received his B.S. degree in computer science from Tokyo University of Technology, Tokyo, Japan in 2017 and is currently a graduate student working towards his Master degree at the same university. His research interests include speech processing and analysis.

Shumei Chen received her M.A. degree from Meiji University, Tokyo, Japan in 1992. She was a visiting lecturer from 1993 to 1996, and a lecturer from 1997 to 1999 at Keio University, Tokyo, Japan. In 1999, she joined the School of Media Science, Tokyo University of Technology; and in April, 2014, she transferred to the Department of Liberal Arts, where she is currently a professor. Her research interests include Chinese education, Japanese and Chinese culture, and educational methodology.

Jinhua She received his B.S. degree in engineering from Central South University, Changsha, China in 1983, and his M.S. and Ph.D. degrees in engineering from the Tokyo Institute of Technology, Tokyo, Japan in 1990 and 1993, respectively. In 1993, he joined the School of Engineering, Tokyo University of Technology, Tokyo, where he is currently a professor. His research interests include the application of control theory, repetitive control, process control, Internet-based engineering education, and robotics.

Dr. She was the recipient of the International Federation of Automatic Control (IFAC) Control Engineering Practice Paper Prize in 1999 (jointly with M. Wu and M. Nakano).

Hiroyuki Kameda received his Dr. Eng. degree in electronic engineering in 1987 from the University of Tokyo. After that, he was a faculty member of the Department of Engineering of the Tokyo University of Technology, and is now the Dean of the Graduate School of Bionics, Computer & Media Sciences. His main research interests are thought and language, and the science of education for software engineering. He is a member of the Institute of Electronics, Information and Communication Engineers, the Japanese Society for Artificial Intelligence, the Japanese Cognitive Science Society, the Association for Computing Machinery, and INCOSE.

Sumio Ohno received his B.E. degree in electrical engineering in 1988, and his M.E. and Dr. Eng. degrees in electronic engineering in 1990 and 1993, respectively, from the University of Tokyo. From April 1993 to March 1999, he was a faculty member of the Department of Applied Electronics, Science University of Tokyo. Since 1999, he has been with the Department of Information Networks, Tokyo University of Technology, where he is currently a professor in the School of Computer Science. His research interests include speech perception, automatic speech recognition, and prosody. He is a member of the Institute of Electronics, Information and Communication Engineers, and the Acoustical Society of Japan.