Natives' shadowability as objectively measured comprehensibility of non-native speech

N. Minematsu[1], Y. Inoue[1], S. Kabashima[1], D. Saito[1], Y. Yamauchi[2], and K. Kanamura[3]
[1]The University of Tokyo, [2]Soka University, [3]Nagoya University of Economics

Keywords—shadowability, speech comprehensibility, measurement, posteriogram, goodness of pronunciation

I. INTRODUCTION

This paper introduces a novel scheme to measure *speech comprehensibility* of L2 utterances objectively by using *native listeners*' shadowings for the L2 utterances. Smoothness of the shadowings is automatically measured as phoneme-based posterior probabilities calculated by applying Deep Neural Network (DNN)-based acoustic models to native listeners' shadowings. The posterior scores have been widely used as Goodness Of Pronunciation (GOP) in CALL studies and they are often applied to L2 utterances including learners' shadowings. In this paper, however, GOP is applied to *native listeners*' shadowings. After shadowing a given L2 utterance, the native shadowers are asked to rate speech comprehensibility subjectively. Correlation analysis shows that GOP of the native shadowings is more highly correlated (R=0.74) to perceived comprehensibility than GOP of the L2 utterances (R=0.58). This result indicates that native shadowings of L2 utterances have higher usability to predict speech comprehensibility objectively than the L2 utterances themselves. Natives' shadowability can be used as objective measurement of speech comprehensibility of L2 utterances.

II. BACKGROUND AND OBJECTIVES

In applied linguistics, intelligibility of an utterance is defined as how accurately linguistic units such as words can be identified in the utterance. Degree of intelligibility of an utterance can be measured objectively by asking native listeners to write down or repeat that utterance word by word. Correct identification rate represents intelligibility of that utterance [1,2]. Comprehensibility of an utterance means how easily and smoothly listeners can understand the content of that utterance, often quantified using questionnaires imposed on listeners [1,2]. Since correct comprehension of an utterance often requires correct identification of words, the authors consider that comprehensibility covers intelligibility and represents more. In [3], 800 Japanese English read-aloud sentences and 600 American English utterances were presented to 173 American listeners who had never talked with Japanese people before. Their oral repetitions were transcribed and word-based correct identification rate, i.e. intelligibility, was calculated. The average rate was approximately 50 % for JE. In the present study, by following [3], a different style of repetition is imposed on native speakers (listeners) to automatically calculate scores that can approximate well speech comprehensibility of L2 utterances.

III. PROPOSED METHOD

A. Natives' shadowing of non-native utterances

In [3], listen-and-repeat was performed for a total of 1,400 utterances. It is highly speculated that efforts of listening and delay of repetition depended on listeners. If delay is minimized, repetition becomes shadowing, where only small listening efforts are allowed and easily comprehensible utterances only will be shadowed smoothly. The authors consider that results of *repetition* indicate how *intelligible* a given utterance is and that results of *shadowing* indicate how *comprehensible* it is.

B. Phoneme-based posterior probabilities as Goodness Of Pronunciation

In this study, as smoothness of natives' shadowing or accuracy of articulation, phoneme-based posterior probabilities are calculated. For speech feature frame at time t, o_t, phoneme-posterior $P(c|o_t)$ can be predicted for every phonemic class c by using DNN acoustic models. After forced alignment, the phoneme intended at time t, p_t, is obtained. Then, $P(p_t|o_t)$ is accumulated during the duration of each phoneme and, for a given utterance x, its GOP score is calculated as

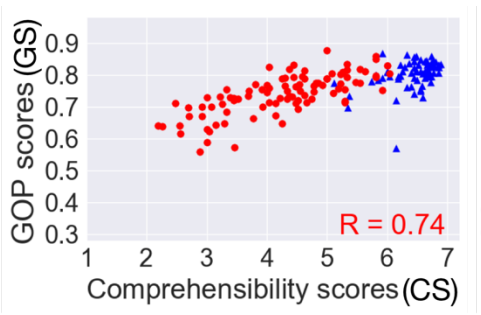
$$GOP(x) = \frac{1}{N_x}\sum_i \frac{1}{D_i} \sum_{a_i \le t \le b_i} P(p_t|o_t)$$

where N_x is the number of phonemes in x and D_i is the duration of the i-th phoneme. a_i and b_i are start and end of the i-th phoneme.

IV. EXPERIMENTS

A. Learners and native shadowers

In this study, the target language of learning is set to Japanese, and learners are Vietnamese. For experiments, their Japanese



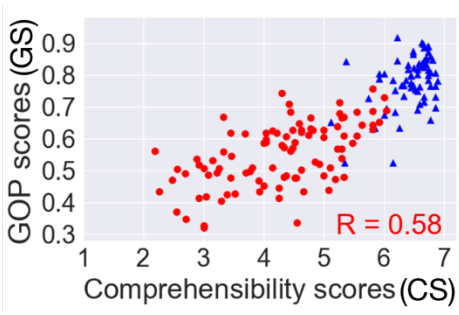


Fig. 1 Correlation between GOP of natives' shadowings and comprehensibility (left) and correlation between GOP of learners' utterances and comprehensibility (right)

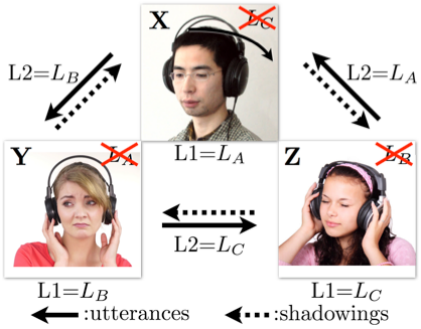


Fig.2 Inter-learner shadowing

utterances were collected and, as reference, native Japanese utterances were also collected. Ten paragraphs were selected from a Japanese textbook for intermediate learners. Each phrase in the ten paragraphs was read aloud by six Vietnamese learners and six native speakers. If a reader stammered, s/he was allowed to read as many times as s/he wanted. Among the six Vietnamese learners, three were at an intermediate level, whose length of learning is shorter than three years (2.7 years on average) and the other three were at an advanced level, who had learned Japanese longer than three years (5.8 years on average). Finally, 96 Vietnamese Japanese (VJ) utterances and 68 native Japanese (NJ) utterances were selected and used for the following experiments.

27 native Japanese, who are different from the above six Japanese, were asked to shadow the above utterances and their shadowing utterances were recorded in such a way that the presented VJ or NJ utterances were not leaked and recorded into a microphone. After shadowing each utterance, the native Japanese were asked to rate how easily they understood the presented utterance, which corresponds to perceived comprehensibility score, CS. Here, a seven-degree scale was used. Before the shadowing experiments, 15-min shadowing practices were made with utterances not used in the experiments.

B. GOP scores calculated for learners' utterances and natives' shadowings

The GOP score, GS, is calculated from each of learners' utterances and from each of natives' shadowings. It should be noted that each VJ utterance has one GS, calculated from that utterance, and the utterance also has 27 CSs and 27 native shadowings, which give us 27 GSs for that VJ utterance, calculated from natives' shadowings. The left figure of Figure 1 shows correlation between averaged GS over natives' shadowings and averaged CS over the shadowers. Red dots and blue dots correspond to VJ and NJ, respectively. The right figure shows correlation between GS of the VJ utterances and averaged CS over the shadowers. It is clearly shown that GS of natives' shadowings is much more highly correlated with CS than GS of learners' utterances. This result indicates very high usability of natives' shadowings to predict CS of a given L2 utterance.

C. Toward inter-learner shadowing

The authors can claim that the proposed framework is very promising but have to confess one critical issue. How can a sufficient number of native shadowers be prepared always for learners? A possible and feasible solution is inter-learner shadowing, shown in Figure 2. Learner X, who speaks L_A as L_1, is learning L_B. He is shadowed by learner Y, who speaks L_B as L_1 and is learning L_C. She is shadowed by learner Z, who speaks L_C as L_1 and is learning L_A. This is a speech version of Lang-8 [5], where any learner can support other learners and can be supported by other learners.

V. Conclusions

In this paper, objective measurement of comprehensibility of learners' utterances was examined based on acoustic analysis and GOP calculation of native listeners' shadowings. Experiments showed much higher usability of natives' shadowings compared to learners' utterances. Further, inter-learner shadowing is also explained. In the new future, the authors are going to test the proposed framework using three groups of learners of (L_A, L_B, L_C) = (Japanese, American English, Chinese).

References

[1] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, 45, 1, 73-97, 1995.

[2] N. Minematsu, K. Okabe, K. Ogaki, and K. Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ database," *Proc. INTERSPEECH*, 1481-1484, 2011.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, 1, 95-108, 2000.

[4] J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, and N. Minematsu, "Automatic scoring of shadowing speech based on DNN posteriors and their DTW", Proc. *INTERSPEECH*, 1422-1426, 2017.

[5] Lang-8, http://lang-8.com