Objective measurement of instantaneous intelligibility of L2 utterances based on shadowing

N. Minematsu[1], R. Hakoda[1], C. Zhu[1], N. Nakanishi[2], T. Nishimura[1], D. Saito[1],
[1]Graduate School of Engineering, The University of Tokyo,
[2]Faculty of Global Communication, Kobe Gakuin University

Keywords — L2 speech assessment, instantaneous intelligibility, shadowing and script-shadowing, posteriorgram, DTW

I. INTRODUCTION

Intelligibility of an L2 utterance, which is often a read-aloud script, is usually measured by asking a listener to transcribe the utterance while or after hearing it, and comparing the transcript with the script to calculate the ratio of correctly transcribed words [1]. However, this framework has some drawbacks [2]. Since manual transcription of an utterance always takes a longer time than the utterance and a listener's memory capacity is limited, the script must be short enough. Further, any listener may reconstruct what s/he has heard while transcribing. These drawbacks seem to be inevitable because the conventional framework is based on *post-listening* observation of listeners. In our work [3], another approach was proposed for measuring *instantaneous* intelligibility based on *while-listening* observation, where a native listener is asked not to transcribe but to shadow a given L2 utterance. Instantaneous intelligibility was measured by quantifying inarticulateness in shadowing. In [4], our proposal was refined by imposing another task of script-shadowing (SS) after shadowing (S). Since SS is viewed as the best performance of shadowing, comparison between S and SS via Dynamic Time Warping (DTW) was shown to be more accurate to quantify inarticulateness in shadowing. In this paper, the S-SS-based scores of instantaneous intelligibility are compared with the scores calculated by manually transcribing shadowing utterances. Experiments show that mean correlation between the automatic scores and the manual scores is 0.935, indicating that instantaneous intelligibility scores can be predicted with high accuracy without manual transcription.

II. SHADOWING AND SCRIPT-SHADOWING BY LISTENERS WITH GOOD PROFICIENCY IN ENGLISH

Shadowing is widely used in language classes in Japan, and it is a task of repeating while listening, not after listening. A learner is asked to repeat a given model utterance while hearing it, with a typical delay of about 1 sec. In this paper, a reverse form of shadowing is imposed on listeners with good English proficiency, who are asked to shadow L2 utterances. Here, they are not asked to imitate non-native accents but just to repeat in their own accents what they have heard. In [5,6], shadowing performances were shown to depend on semantic and syntactic difficulty of presented stimuli as well as accents of the stimuli. Shadowing was originally proposed in speech perception studies about a half century ago [7] to analyze the human behaviors of processing spoken words and to examine the process of accessing to the mental lexicon [8] that is assumed to exist in mind. In this paper, shadowing is used not for language learning purposes but for the original purpose. Shadowing performances are viewed as results of *while-listening* observation of listeners' comprehension, and they objectively show when and how listening comprehension fails.

As shown in Fig. 1, a listener shadows an L2 utterance and script-shadows the same utterance. Script-shadowing is a special type of shadowing with the L2 utterance's script presented visually. Since script-shadowing is the best performance of shadowing, comparison of S and SS, termed as |S – SS|, can quantify how smoothly the process of listening comprehension is running. The comparison is made through Dynamic Time Warping (DTW) between two Phonetic PosteriorGrams (PPG, shown in Fig. 2) of S and SS. A PPG is obtained from a spectrogram using a front-end of DNN-based ASR, and it can be viewed as the probabilistic version of a phonemic transcript. In PPG, for a speech frame at time *t* of a given utterance, its phonemic identity is not determined uniquely but it is represented as probabilistic distribution over the entire phonemes. In the ASR community, PPG is used widely as a fundamental speech representation assumed to be independent of extra-linguistic factors such as age and gender.

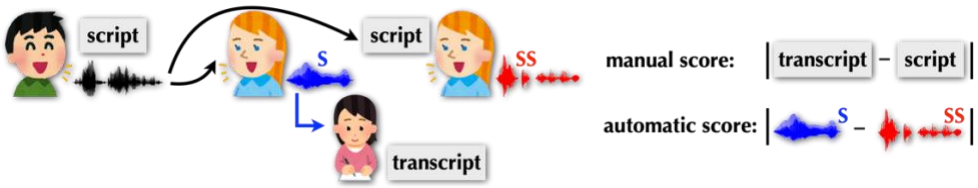

Figure 1: Manual scores and automatic scores of instantaneous intelligibility

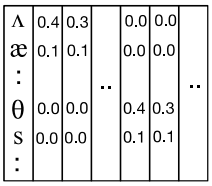

Figure 2: Phonetic PosteriorGram (PPG)

In this paper, |S–SS| is compared with the result of comparing the script of the L2 utterance with the manual transcript of the shadowing. Comparison of the two texts is made with text-based DTW, where two sequences of words are compared. Finally, word-unit accuracy is calculated and used as manual score of instantaneous intelligibility. This manual score can be refined by

using PPGs of the two texts, which are obtained by having Amazon Polly read aloud the two texts and converting the two utterances to PPGs. The two PPGs are compared again by DTW. In the end, we have two kinds of word-unit manual scores of instantaneous intelligibility, i.e. text-based and PPG-based. We compare these manual scores with automatic scores calculated as |S–SS|. Further, another type of automatic scores are introduced here. By replacing a human transcriber in Fig.1 by a machine transcriber, i.e. an ASR system, automatic transcripts are obtained. Here, the Amazon ASR system is used. It should be noted that, in the experiments, shadowers are non-native speakers with good proficiency in English as well as native speakers. The ASR performance will surely depend on the L1 of the shadowers but |S-SS|-based comparison will be independent of their L1.

III. Experiments and Results

A. Experimental Conditions

12 Japanese learners of English were preselected out of 30 based on their Versant Test scores so that they could cover a wide range of proficiency. The 12 learners and additional 2 native speakers wrote original passages and read aloud the passages. The duration of each passage was about 30 seconds. The 14 utterances were used as stimuli, which were presented to 3 groups of shadowers, 2 natives (N1 and N2), 2 Japanese (J1 and J2), and 3 non-natives (NN1, NN2, and NN3) whose L1 are not Japanese. The 2 native shadowers are different from the 2 native speakers who participated in recording. The 5 non-Japanese shadowers did not understand Japanese at all and the 5 non-native shadowers had very good proficiency in English. The 7 shadowers shadowed and script-shadowed the 14 utterances. 7x14 shadowings were carefully transcribed manually. Here four transcribers participated in the experiment. A transcriber transcribed shadowings from the shadowers whose L1 is the same as L1 of the transcriber.

Since SS was always used as reference in DTW-based S-SS comparison, we made comparison using an asymmetric local path for DTW. Since various types of inadequate productions are found in shadowing [9], the local path shown in Fig. 3 was used. Here, the following conditions are satisfied, 1) local distances are always accumulated step by step along with the reference of SS, and 2) insertion of additional words in S are ignored, but replacement as other words or silence in S is penalized.

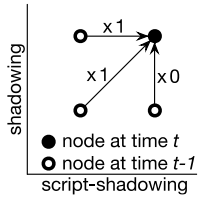

Figure 3: Local path used for DTW

Table 1: Correlations between the automatic scores and the manual scores

automatic vs. manual	N1	N2	J1	J2	NN1	NN2	NN3	mean
ASR vs. text-based	0.889	0.854	0.874	0.690	0.904	0.910	0.847	0.853
|S-SS| vs. text-based	0.869	0.860	0.869	0.898	0.937	0.916	0.901	0.893
|S-SS| vs. PPG-based	**0.961**	**0.883**	**0.921**	**0.899**	**0.956**	**0.946**	**0.980**	**0.935**

B. Results and Discussion

Tab. 1 shows the correlations in the three cases, 1) between text-based automatic scores using ASR and text-based manual scores, 2) between |S-SS|-based automatic scores using PPG and text-based manual scores, and 3) |S-SS|-based automatic scores using PPG and PPG-based manual scores. Here, the correlations are shown for each shadower, and the highest correlations are shown in bold. All the correlations in the second case are negative, but their sings are removed for easy comparison.

The correlations in Table 1 are generally high, which indicates high validity of automatic measurement of listeners' behaviors through shadowing and script-shadowing. However, J1 and NN3 show comparatively smaller correlations in the case of ASR vs. text-based. This is because they had a strong accent although they are proficient users of English. With the strong accent, ASR performance was degraded easily. With |S-SS|, however, their correlations are much improved simply because DTW-based comparison between S and SS is not influenced by the accent found in common between S and SS. In the table, the highest correlations are always found in the case of the automatic scores of |S-SS| vs. the PPG-based manual scores. The mean correlation is so high as 0.935, indicating that automatic prediction of the manual instantaneous intelligibility score is precise enough.

In the proposed framework, human shadowers are needed, and thus it is not fully automatic. Currently, we are collecting a huge number of S and SS from a selected shadower, where read-aloud scripts of hundreds of Japanese learners are used. With this collection of S and SS, we will build a simulator of that shadower, which will be able to indicate in which parts of a new L2 utterance the instantaneous intelligibility becomes lower. With this virtual shadower, no more human shadowers may be needed.

References

[1] T. M. Derwing et al., Accent, comprehensibility and intelligibility: Evidence from four L1s, *J. Studies in Second Language Acquisition*, 19, 1, 1-16, 1997

[2] M. J. Munro et al., Foreign accent, comprehensibility and intelligibility, redux, *J. Second Language Pronuncition*, 6, 3, 283-309, 2020

[3] N. Minematsu et al., "Natives' shadowability as objectively measured comprehensibility of non-native speech," presented at ISAPh2018.

[4] Z. Lin et al., "Shadowability annotation with fine granularity on L2 utterances and its improvement with native listeners' script-shadowing," *Proc. INTERSPEECH*, 3865-3869, 2020

[5] T. Tristichoke et al., "Influence of context variations on smoothness of native speakers' reverse shadowing," *Proc. ICPhS*, 2019

[6] C. Zhu et al., "Analyses on instantaneous perception of Japanese English by listeners with various language profiles," *Proc. PSJ*, 26-31, 2020

[7] W. Marslen-Wilson, "Linguistic structure and speech shadowing at very short latencies," Nature 244, 522-523, 1973

[8] H. Fujisaki et al., "Influence of context and knowledge on the perception of continuous speech," *Proc. ICSLP*, 417-420, 1990

[9] S. Shi et al., "A corpus-based analysis of shadowing speech: case of L2 English by Japanese learners," *Proc. ISAPh*, 34-37, 2016