A WoZ Study for an Incremental Proficiency Scoring Interview Agent Eliciting Ratable Samples

Mao Saeki, Weronika Demkow, Tetsunori Kobayashi, and Yoichi Matsuyama

Abstract To assess the conversational proficiency of language learners, it is essential to elicit ratable samples – speech samples that are representative of the learner's full linguistic ability. This is realized through the adjustment of oral interview questions to the learner's perceived proficiency level. An automatic system eliciting ratable samples must incrementally predict the approximate proficiency from a few turns of dialog, and employ an adaptable question generation strategy according to this prediction. This study investigates the feasibility of such incremental adjustment of oral interview question difficulty during the interaction between a virtual agent and learner. First, we create an interview scenario with questions designed for different levels of proficiency and collect interview data using a Wizard-of-Oz virtual agent. Next, we build an incremental scoring model and analyze the accuracy. Finally, we discuss the future direction of automated adaptive interview system design.

1 Introduction

With a growing demand for language education, there is much need for the automation of assessment for linguistic proficiency. An easily accessible assessment would allow for the monitoring of each individual student's progress and facilitate the tailoring of curriculum for a more effective learning. Although much research has been done on the automatic assessment of written texts and monologues, the valuation of dialogic speech in conversational settings – or oral proficiency – still heavily relies on human-led interviews[1]. Not only are human-led interviews costly, it has been pointed out that behavioral differences among interviewers can lead to unwanted variation in test ratings[2].

Mao Saeki
Waseda University, Tokyo, Japan, e-mail: saeki@pcl.cs.waseda.ac.jp

Given the consistent behavior of dialogue systems, there have been recent attempts on using them for automated oral proficiency assessment[3, 4, 5, 6]. However all studies use a fixed task difficulty throughout the interaction. This is problematic because test takers are composed of highly varying levels of proficiency, and unless they are matched with tasks appropriate to their level, a test may fail to accurately measure their language skill.

To provide tasks with an appropriate level of difficulty, it is necessary to assess test-takers' proficiency incrementally. In this paper, we investigate the feasibility of such incremental assessment. To this end, we first designed an adaptive interview using Wizard-of-Oz (WoZ) system and collected 56 interviews of English learners, scored by human raters. We then used a recurrent neural network (RNN) model to incrementally score the learner at different stages of the interview. To the best of our knowledge, this is the first work on the incremental assessment of oral proficiency in dialogic settings. We demonstrate high agreement to human raters as the validity evidence of our system, promoting the progress for adaptive oral proficiency tests. The rest of the paper is organized as follows. Section 2 reviews previous work on oral proficiency interview frameworks and automated assessments. Section 3 explains the design of the interview test, the development of the WoZ system and the data collection process. Section 4 explains the incremental assessment model. Section 5 reports on findings from the data collection, the performance of the incremental scoring model, and discusses the results. Finally, Section 6 draws conclusions.

2 Related Work

Oral proficiency interviews have long been examined to create fair and reliable tests. One notable framework is the Oral Proficiency Interview (OPI), developed by the American Council on the Teaching of Foreign Languages (ACTFL)[1]. The ACTFL-OPI begins with a "Warm-up", where the examiner eases the candidate into the test by asking questions and making small talk. Through the warm-up, the interviewer makes a brief, or preliminary, assessment of the candidate's proficiency level. The next two stages are part of a crucial "iterative process" in which the examiner alternates between a comfortable and challenging difficulty, in order to provoke loss of linguistic control. Such loss is known as the "signs of breakdown", and may include hesitation, false starts, a lack of response or self-correction. The iterative process is repeated and re-adjusted until sufficient information is gathered to correctly assess the difficulty level at which the speaker experiences breakdown.

To date, only a few studies have used dialogue systems for oral proficiency scoring. The ACTFL Oral Proficiency - computer is a commercially available test which uses a virtual agent for a simulated interview[6]. The interview is simulated in the sense that all system utterances are generated regardless of the user's previous utterance. A self assessment made prior to the interview is used to adjust the question difficulty, but no adjustments are made during the interview itself. [4] used off-the-shelf dialogue systems to have users participate in a task-based conversation. The

interaction was scored automatically using a model for non-interactive speech based on Gaussian process. [5] also collected task-based conversations and scored the interaction aspect using RNN. Other work on dialogue scoring has used RNNs to capture the multi-turn nature of a dialogue, as well as the fusing different modalities. [7] for example, fused features representing the content, delivery and language use, while [8] tried to incorporate visual cues for scoring. The process of narrowing down user level through incremental assessment and question selection (as featured in the ACTFL-OPI) is key to a reliable test. However, no automated assessment has done it so far.

3 Data Collection

3.1 Experimental Design

Since existing interview frameworks are not directly applicable to dialogue systems due to technical limitations, we designed our own task based on the ACTFL-OPI. Our adaptive oral proficiency interview consists of several topics that are set around a main question, and proceeded by follow-up questions. The follow-up questions concerned the same topic as the main question, and served to elicit additional speech sample.

The interview begins with a warm-up, during which all candidates are questioned on the same topic. A preliminary assessment made during this stage is used to branch candidates into three levels of proficiency The proficiency scale used is based on the Common European Framework of Reference for Language (CEFR)[9]. Each of these three levels of proficiency has a pre-prepared subject for discussion as well as corresponding questions. At the closure of each topic, the candidate is reassessed to attune task difficulty. If the candidate either falls behind or goes beyond the criteria for a certain level, they are moved to the respective branch.

3.2 WoZ Interview System

We developed the Intelligent Language Learning Assistant (InteLLA), our virtual agent leading the oral proficiency interview for the data collection. The agent is rendered using the Unity game engine[1], and its motion can be controlled through a list of pre-recorded body movements and vocal responses in WoZ style. The use of WoZ system allows for the collection of human-agent interaction data without the need to build an automated system, making a rapid feasibility study possible. Through an initial study of human-led interviews, 76 utterances and 4 non-verbal behaviors were selected as the options for the Wizard. The utterances include a

[1] https://unity.com/

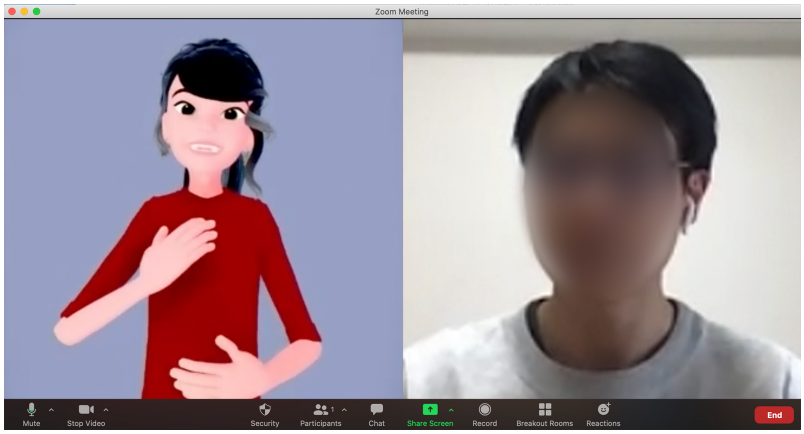

Fig. 1: InteLLA (left) is interviewing with a participant remotely (right) via a video conferencing tool.

greeting, instructions, main questions, follow-up questions and feedback. For the non-verbal behaviors, we included nodding, smiling and surprise, all to be used for active listening. Speech was generated using Text-to-Speech, and the agents motions were recorded using motion and facial expression capture technology.

The agent was controlled by two operators. A main operator managed the utterances, while a sub-operator selected appropriate non-verbal behaviors. The reason for needing two separate operators, is the heavy cognitive and operational load that proved to be too intensive for a single operator. More specifically, the selection of appropriate utterances requires careful listening of the content of a user's speech, which must also be used to make a quick estimation of their language proficiency. This process is done alongside the giving of non-verbal feedback, which in turn requires the monitoring of phonological and visual cues.

3.3 Interview Data Collection and Human Assessment

With the use of InteLLA, we collected interviews from 56 Japanese English-learners. All test subjects were university students with varying levels of English proficiency. Each student discussed 7 different interview topics. The interview was conducted online using the video conferencing tool, Zoom. Though online conversations are different in nature to face-to-face discourse, studies shows that speaking assessment in the two modes show similar results[10]. All test users completed the interview remotely, which lasted 9 minutes on average. After the interview, users were asked to evaluated the interview using a 5-point Likert scale questionnaire to obtain subjective evaluation on firstly, how well the system was able to adapt to each user, and secondly, how well the system was able to measure the user's ability. The former is measured by how appropriate the user thought the question difficul-

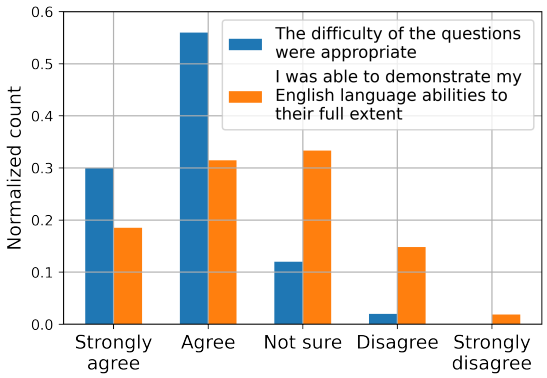

Fig. 2: Subjective evaluation of the WoZ interview on how well the system was able to adapt to each user, and how well the system was able to measure the user's ability.

ties were, and the latter by whether the user was able to demonstrate their language abilities to the full extent. The reasons behind their evaluation were also collected through free-form questionnaires. Figure 1 shows a screenshot of the interview data. Figure 2 shows the results of the questionnaire, which will be discussed in section 5.

Each interview was scored by human raters using the CEFR scale. We adopted the scale for "communicative language competence", consisting of the standard 6 levels: A1, A2, B1, B2, C1 and C2. A1 represents the lowest proficiency, and C2 the highest. The whole dataset was annotated by a single rater with extensive experience in CEFR grading. Since only two students were in the C1 and C2 bandwidth, we excluded them from further analysis. To measure the inter-rater agreement, we asked another rater to annotate a subset of the dataset with 20 students. The inter-rater agreement calculated using quadratic weighted κ (QWκ) was 0.753.

4 Incremental Prediction Model

During the data collection discussed in section 3.3, the operator focused on speech sample elicited in each subsequent topic and updated their assessment based on the observed quality of speech. To capture this incremental decision-making process, we used a LSTM neural network. The input features were chosen from previous studies on monologue and dialogue scoring[11] as well as through the analysis of the annotation process. Features covers aspects such as vocabulary level, fluency and coherence, and a complete list is shown in Table 1. The difficulty of the word used by the student is calculated using the CEFR-J Wordlist[12, 13]. The model was trained after the completion of each topic. For the respective topic labels, we used the score assigned to a user's interview as a whole. A 5-fold cross-validation was conducted, with the Adam optimization algorithm[14] used to minimize the

mean squared error (MSE) loss function over the training data. We oversampled the minority classes to address their imbalance, and trained each fold for 40 epochs.

Feature name	Description
Response length	The length of response for each turn in number of words and seconds.
Word n-gram	Number of unique 1-gram, 2-gram and 3-gram used
Word level	The mean difficulty level of words used.
Speech rate	Number of syllables per seconds.
Pause frequency	Frequency of pauses.
Transition time	The length of time between the end of the system's utterance and the beginning of the user's speech
Discourse marker	Number of discourse markers.

Table 1: List of features for incremental proficiency scoring

5 Results and Discussions

The incremental scoring model was evaluated using accuracy and QWκ after each of the 7 interview topics. The human scores were discrete values while the model predictions were continuous. We therefore rounded the model prediction for evaluation. Figure 3 shows the mean result over 5 runs with an error band. The confusion matrix of the prediction after topic 2 and 7 is shown in Figure 4. The accuracy and correlation was still low after topic 2 (the warm-up), but most predictions were made within one level of error. The accuracy and QWκ increased after each topic, saturating around the fourth topic. Correlation with human scoring exceeded that of the human-human agreement at this point. The highest accuracy and QWκ was achieved at topic 7 (the final prediction), being 0.604 and 0.784 respectively. These are encouraging results because they show that the model can capture the approximate level of proficiency of a user with only a few turns of dialogue. This estimation can then be used for a better adjustment of the task and in turn, a better assessment. One limitation of this study is the lack of advanced level English speakers (C1 and C2 CEFR level). Although beginner to intermediate students are the most common proficiency for Japanese English-learners, we would like to evaluate our model on advanced level speakers in future studies.

Finally, we will discuss the subjective evaluation of the whole interview, shown in Figure 2. 80% of the users found the question difficulty appropriate, which is unsurprising given that human operators were actively adjusting them. Nevertheless, this percentage assures the validity of our test design. On the other hand, only 50% of the users agreed that they were able to demonstrate their English ability to its full potential. We identified two key reasons behind these results from our open-questionnaire. These were a lack of adequate active listening listening strategies from the system, and a lack of sufficient topic development through follow-up questions. Although this work has not focused on dynamic content generation, such

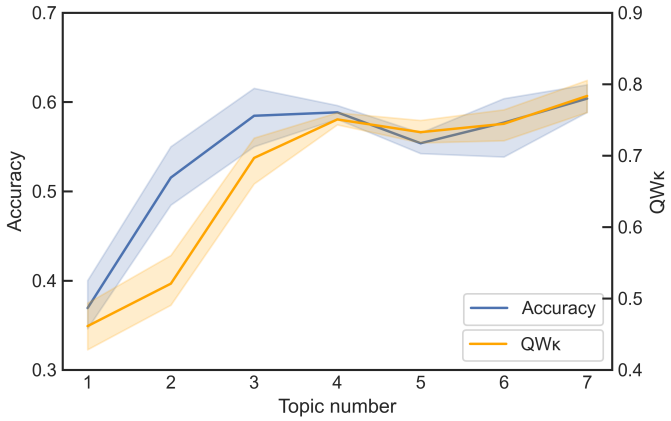

Fig. 3: Accuracy and Quadratic Weighted κ for each interview topic

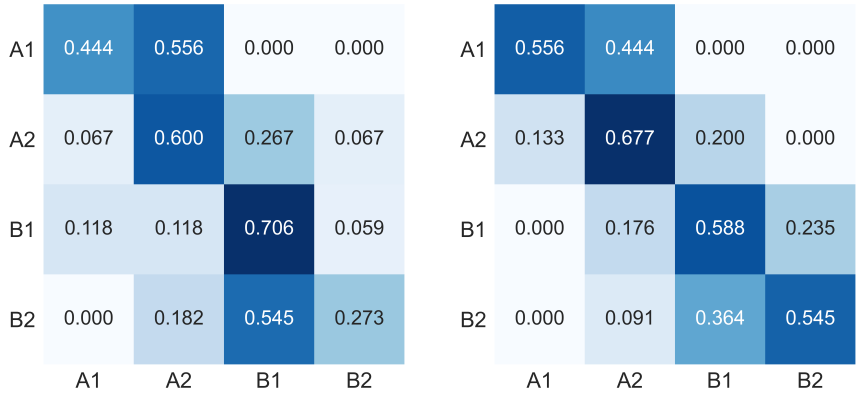

Fig. 4: Normalized Confusion Matrix for incremental prediction after topic 2 (left) and 7 (right)

functionality is important for users to better demonstrate their language abilities. Active listening and question generation strategies have previously been studied for job interview systems[15, 16], and the implementation of such strategies will be considered in future studies.

6 Conclusion

This paper has investigated the feasibility of incremental assessment of oral proficiency using an adaptive test format. First, we designed our own interview protocol for an automated adaptive testing and built a WoZ system to serve as the interviewer. Using the WoZ system, we collected an interview dataset of 56 English learners, annotated using the CEFR scale - an international standard for language proficiency evaluation. We then built a LSTM based incremental assessment model that updates its prediction every few turns of the dialogue. Results showed a moderate agreement with human scoring throughout the beginning of the interview, which increased over time, and finally surpassed human inter-rater agreement. Encouraged by this result, our future direction will be to include the incremental scoring model into dialogue systems for a fully automated adaptive oral proficiency test.

7 Acknowledgements

This paper is based on results obtained from a project, JPNP20006 (”Online Language Learning AI Assistant that Grows with People”), subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

References

1. Judith E. Liskin-Gasparro. The ACTFL Proficiency Guidelines and the Oral Proficiency Interview: A brief history and analysis of their survival. *Foreign Language Annals*, 36(4):483–490, 2003.
2. Annie Brown. Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1):1–25, 2003.
3. Keelan Evanini, Sandeep Singh, Anastassia Loukina, Xinhao Wang, and Chong Min Lee. Content-Based Automated Assessment of Non-Native Spoken Language Proficiency in a Simulated Conversation. In *Machine Learning for SLU & Interaction*, pages 1–7, 2015.
4. Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen, and David Vandyke. Towards Using Conversations with Spoken Dialogue Systems in the Automated Assessment of Non-Native Speakers of English. In *Proceedings of SIGdial*, pages 270–275, 2016.
5. Vikram Ramanarayanan, Matthew Mulholland, and Yao Qian. Scoring Interactional Aspects of Human-Machine Dialog for Language Learning and Assessment using Text Features. In *Proceedings of*, pages 103–109, 2019.
6. Stephanie Dhonau. ACTFL Oral Proficiency - computer. Technical report, 2020.
7. Neural Approaches to Automated Speech Scoring of Monologue and Dialogue Responses. In *ICASSP*, pages 8112–8116, 2019.
8. Mao Saeki, Yoichi Matsuyama, Satoshi Kobashikawa, Tetsuji Ogawa, and Tetsunori Kobayashi. Analysis of Multimodal Features for Speaking Proficiency Scoring in an Interview Dialogue. *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, pages 629–635, 2021.

9. Council of Europe. Common European Framework of Reference For Languages: Learning, Teaching, Assessment. *Cambridge University Press*, 2018.
10. Fumiyo Nakatsuhara, Chihiro Inoue, Vivien Berry, and Evelina Galaczi. Exploring the Use of Video-Conferencing Technology in the Assessment of Spoken Language: A Mixed-Methods Study. *Language Assessment Quarterly*, 14(1):1–18, 2017.
11. Klause Zechner and Keelan Evanini. *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. Routledge, 2020.
12. Yukio Tono. The cefr-j wordlist version 1.6, tokyo university of foreign studies. http://www.cefr-j.org/index.html/, 2021. [Online; accessed 1-June-2021].
13. Masashi Negishi, Tomoko Takada, and Yukio Tono. A progress report on the development of the cefr-j. In *Exploring language frameworks: Proceedings of the ALTE Kraków Conference*, pages 135–163, 2013.
14. Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *Proceedings of ICLR*, pages 1–15, 2015.
15. Koji Inoue, Kohei Hara, Divesh Lala, Shizuka Nakamura, and Katsuya Takanashi. A job interview dialogue system with autonomous android ERICA. In *Proceedings of IWSDS*, pages 1–6, 2019.
16. Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. An Attentive Listening System with Android ERICA: Comparison of Autonomous and WOZ Interactions. In *Proceedings of SIGdial*, pages 118–127, 2020.