InteLLA: Intelligent Language Learning Assistant for Assessing Language Proficiency Through Interviews and Roleplays

Mao Saeki[1,2], Hiroaki Takatsu[1,2], Fuma Kurata[1], Shungo Suzuki[1], Masaki Eguchi[1], Ryuki Matsuura[1], Kotaro Takizawa[1], Sadahiro Yoshikawa[2], and Yoichi Matsuyama[1,2],
[1]Waseda University, [2]Equmenopolis, Inc.
Correspondence: saeki@equ.ai

Abstract

The primary challenge in utilizing dialogue systems for reliable language assessment for interactional skills lies in obtaining ratable speech samples that demonstrate the user's full range of ability. We thus developed a multimodal dialogue system that employs adaptive sampling strategies and enables a mixed initiative interaction through extended interview and roleplay dialogues. The interview is a system-led dialogue aimed at evaluating the user's overall proficiency. The system dynamically adjusts the question difficulty based on a real-time assessment to induce linguistic breakdowns, which provides evidence of the user's upper limits of proficiency. The roleplay, on the other hand, is a mixed-initiative, collaborative conversation intended to assess interactional competence such as turn management skills. Two experiments were conducted to evaluate our system in assessing oral proficiency. In the first experiment, which involved an interview dataset of 152 speakers, our system demonstrated high accuracy in automatically assessing overall proficiency. However, we observed that linguistic breakdowns were less likely to occur among high-proficiency users, indicating some room for further enhancing the ratability of speech samples. In the second experiment based on a role-play dataset of 75 speakers, the speech samples elicited by our system was found to be as ratable for interactional competence as those elicited by experienced teachers, demonstrating our system's capability in conducting interactive conversations. Finally, we report on the deployment of our system with over 10,000 students in two real-world testing scenarios.

1 Introduction

Language testing plays a critical role in ensuring effective language learning, as it provides valuable feedback on learners' proficiency levels and guides instructional planning (Fulcher, 2010). Assessment of oral proficiency is particularly important, as speaking and listening skills are essential for effective communication in a second language. Traditional methods of oral proficiency assessment, however, face several challenges, including the subjectivity of human raters and the difficulty of creating standardized, scalable testing environments (Galaczi and Taylor, 2018).

To address these challenges, several studies have explored automated systems for oral assessment. For example, Ockey and Chukharev-Hudilainen (2021) evaluated the potential of spoken dialogue systems (SDS) for paired oral discussion tasks, concluding that a standardized assessment may favor SDS over human interlocutors due to its systematic behavior. Recent advancements in large language models (LLMs) have further simplified the implementation of such dialogue tasks. However, a significant challenge remains in obtaining "ratable" speech samples that accurately represent the full extent of a learner's language capabilities. Assessment of oral proficiency requires not only measuring linguistic competence, such as grammar and vocabulary, but also evaluating interactional behaviours, including turn-taking, topic management, and repair strategies (McNamara, 1996). Additionally, to provide a reliable assessment, it is crucial to observe the upper linguistic limits of the user (Liskin-Gasparro, 2003). Therefore, an effective dialogue system must be capable of engaging users in a manner that naturally reveals these competencies while also being scalable as a testing tool.

To this end, we developed the Intelligent Language Learning Assistant, *InteLLA*, a multimodal dialogue system designed to elicit spontaneous speech samples from second language learners through a combination of a 15-minute interview and a 10-minute roleplay session. By dynamically adjusting the topic difficulty based on real-time assessments, the system aims to provoke linguistic breakdowns that serve as evidence of a learner's upper proficiency limits. Additionally, the mixed-initiative roleplay component is designed to evalu-

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 385–399
September 18–20, 2024. ©2024 Association for Computational Linguistics

Figure 1: The InteLLA system for oral proficiency assessment. The user connects to an online video call with InteLLA from their web browser on PC, tablet or smartphone.

ate the user's interactional competence in a collaborative setting.

To ensure the functionality and potential limitations of our system for large-scale real-world implementation, this paper reports two experiments: Chapter 4 evaluates how well the system can assess oral proficiency through various experiments designed to test its efficacy; Chapter 5 reports on the field testing results of our system used in real-world testing scenarios with university and high school students in Japan. We also discuss our first year operation of our system in terms of practicality and social impacts.

2 Related Work

2.1 CEFR

The Common European Framework of Reference for Languages (CEFR) serves as a comprehensive foundation for the development of language syllabi and curricula, as well as the evaluation of foreign language proficiency (Council of Europe, 2020). According to the CEFR, the key competencies for effective language communication include range (vocabulary richness), accuracy (grammatical correctness), fluency (smoothness and flow of speech), interaction (ability to engage in conversational exchange), and coherence (Engaging in effective conversational exchange). These competencies are defined across six proficiency levels: A1, A2, B1, B2, C1, and C2, with A1 representing the beginner level and C2 indicating proficient or near-native speaker capabilities.

The CEFR outlines specific communicative activities referred to as "Can-Do" statements, which articulate what learners at each proficiency level should be able to achieve. These "Can-Do" serve as guidelines to determine the appropriate level for a learner based on their demonstrated abilities in a certain social situation. For instance, at the B1 level, learners should be able to handle most situations likely to arise while traveling in an area where the language is spoken.

This standardization is particularly valuable in the development of dialogue systems for language testing, as it offers an established baseline for designing tasks, including the interlocutor's behaviors, and evaluating user performance in a reliable and valid manner.

2.2 Oral Proficiency Interview

In many computerised speaking assessments, the user is given a reading script or situational explanation and is then required to record their speech. Such monologue-based speaking score, however, only have moderate correlations with those elicited in interactive dialogue tasks (Roever and Ikeda, 2022). On the other hand, due to their dynamic and co-constructive nature, dialogic tasks inevitably introduce variability in examiner behaviours and thus affect the learner's performance in the test (Galaczi and Taylor, 2018). This inherent variability poses a challenge for maintaining consistent and reliable assessments in dialogue-based tasks.

To draw out such interactive abilities, interview-based assessments of speaking proficiency conducted by trained professionals have long been considered, a representative implementation being the ACTFL-OPI (Liskin-Gasparro, 2003). The ACTFL-OPI interview consists of several phases. It begins with a "warm-up" where the interviewer asks questions or engages in small talk to familiarize the examinee with the test. Through this warm-up, the interviewer conducts a preliminary evaluation to decide the difficulty level of the first main topic. Next, the main part of the assessment, the "iterative process" takes place. The interviewer alternates between questions that are perceived as comfortably easy and challengingly difficult for the examinee to induce signs of "breakdown". Typically, breakdowns are indicated by hesitation, stumbling, lack of response, or rephrasing. This iterative process continues until sufficient information is obtained to assess the examinee's proficiency accurately. Automated assessment systems that mimic this interview strategy, such as the ACTFL Oral Proficiency Computer, exist. However, these systems do not rely on the user's previous responses but rather output a predefined list of questions sequentially (Isbell and Winke, 2019). Although some measures are taken such as adjusting the difficulty of questions based on self-assessment before the interview, dynamic level adjustments during the interview, as performed by human experts, are not conducted.

Research into systems that conduct interview or counseling-like dialogues has been extensive in the domains other than language testing (Morbini et al., 2014; Inoue et al., 2020). These systems aim to elicit user speech through natural listening and question generation, but few explicitly evaluate user performance. Additionally, there is considerable research on using dialogue systems for speaking proficiency assessment (Ramanarayanan et al., 2019; Litman et al., 2016), but these studies generally assign the same tasks to all users from the perspective of test fairness and avoiding dialogue breakdowns.

2.3 Roleplay Dialogue

While structured interaction tasks such as the ACTFL-OPI have been used extensively to elicit ratable samples to assess linguistic competence (e.g., vocabulary, grammar, pronunciation), it falls short in assessing a full range of interactional competence. As such, language assessment researchers attempt to incorporate roleplay tasks in their tests to simulate authentic social settings for the examinees to demonstrate their abilities to enact simulated social roles by maintaining interpersonal relationships and managing turn-taking in a collaborative and cooperative manner (Kasper and Youn, 2018). By design, such roleplay dialogues should involve mixed-initiative interactions where both the system and the user can take the lead in conversation. This requirement is essential to making it possible to evaluate how well the learner handles unexpected turns and engages in collaborative communication.

2.4 System Requirements

Based on the aforementioned considerations, our system needs to effectively assess oral proficiency through both structured interviews and collaborative roleplay interactions. To achieve this, we have established the following requirements for the conversational agent being developed in this project:

1. **Adaptive speech sampling strategy**: The system should ask relevant questions and provide responses tailored to the user's language level, efficiently sampling ratable speech data for assessment. Multimodal interaction, including non-verbal gestures, is needed to elicit authentic speech, ensuring that scores are generalizable to real-world communication.

2. **Mixed-initiative interaction**: The system should enable collaborative, mixed-initiative dialogues, wherein both the system and the user can dynamically control the conversation. This will allow users to demonstrate their interactional competence, including aspects such as turn-taking and topic development.

3. **Scalability**: To ensure the test is accessible and fair for a diverse user base, the system must be usable across different locations and operable on low-end devices.

3 System Design

The InteLLA system is a multimodal dialogue system where the user connects to an online video call from their personal device, as shown in Figure 1. We adopted a modular architecture, wherein multiple modules, each responsible for specific dialogue capabilities such as ASR, operate concurrently to enable fully-duplex communication (Figure 2). For the ASR module, we employ the Google Text-to-Speech service. The details of the other modules will be discussed in subsequent sections.

3.1 Video Communication Module

To enable users to access the system via video call directly from their web browser, the system is hosted on a server, with agent audio and visuals streamed to the user through a Web Real-Time Communication (WebRTC) solution. This setup leverages server-side GPU resources for machine learning and rendering, ensuring a rich conversational experience even for users with low-end devices. This configuration is crucial for maintaining equitable and consistent testing.

3.2 Dialogue Management Module

LLMs have greatly simplified the design and management of dialogues by enabling the specification of conversation rules through prompts (Brown et al., 2020). However, these models often struggle to maintain coherence when the input (i.e. prompts and dialogue history) becomes excessively long. This poses a particular challenge in our use case,

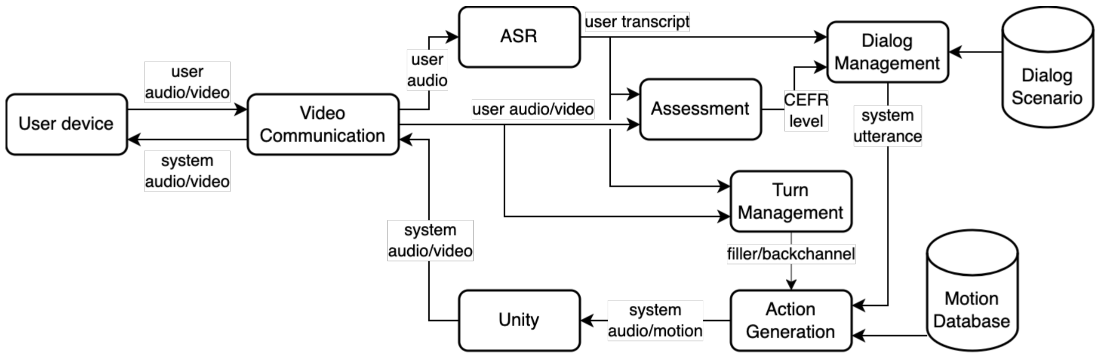

Figure 2: System architecture of InteLLA, comprising video communication, automatic speech recognition, dialogue management, turn management, action generation and assessment modules.

where a single conversation may extend from 20 to 30 minutes. Additionally, altering dialogue content based on real-time assessments for adaptive testing remains an issue.

To address these challenges, we employed a hybrid approach that combines LLMs with scenario-based dialogue management. Specifically, we segmented the interviews and roleplays into multiple topics, with each topic having a sub-goal such as "asking about hobbies" or "conducting a roleplay to borrow a PC from the user," designed to be completed within a 3 to 5-minute timeframe. These topics are tailored for each CEFR level.

Following the OPI framework described in Section 2.2, the conversation initiates with a warm-up topic, designed to make the user comfortable with the system. During the conversation, the assessment module, explained in Section 3.5, evaluates the user's proficiency. Based on this assessment, users are assigned a topic that matches or slightly exceeds their proficiency level. This aims to induce linguistic breakdowns, thereby efficiently observing the user's upper proficiency limits, as shown in Appendix A.1.

When a topic changes, the prompt for the LLM is updated, and the dialogue history is reset. By compartmentalizing conversations in this manner, the LLM can adhere to strict instructions for each individual topic, ensuring coherent and controlled dialogue over a total duration of 20 to 30 minutes. To maintain the memory of previous topics, we summarize earlier dialogue segments and incorporate these summaries into the updated prompts.

To enhance the ratability of speech samples, a panel of applied linguistics researchers and experienced teachers carefully designed and piloted prompts. Following the literature on the correspondence between representative linguistic functions and CEFR levels (O'Sullivan et al., 2002), topics were decided in terms of how likely learners are to use target linguistic functions in response to the system's question. For instance, B2-level learners are expected to have the ability to produce a longer, coherent utterance, and thus the topics, for instance, should require them to compare and contrast multiple ideas. These are combined with generic prompts such as persona of the agent, guideline for the interview and summarized history, and fed to a LLM to generate the next system utterance. We use OpenAI's GPT for the utterance generation.

3.3 Turn Management Module

In spoken dialogue systems, knowing when to speak is as crucial as knowing what to say for maintaining smooth interaction (Skantze, 2021). This is particularly important in the context of oral proficiency testing, where users often produce long pauses between sentences as they formulate their responses, increasing the likelihood of system interruptions.

During these pauses, it is often discernible whether the user intends to continue speaking or has finished based on grammatical completeness, prosody, and eye gaze. To utilize such multimodal cues, we trained an end-of-turn detection model that incorporates text, audio, and image data to predict whether the user has finished their turn, as proposed by Kurata et al. (2023). However, turn overlaps are inevitable, even in human conversation. Not all overlaps are detrimental; for instance, the user may simply be providing backchannel feedback to the system. To determine whether the system should continue speaking or pause when turn overlaps occur, we implemented a barge-in detection system based on the overlap resolution model by Gervits and Scheutz (2018).

This module is also responsible for generating backchannels and fillers. Backchannels are necessary cues to indicate the system is listening to the user, thereby encouraging the user to speak more. Verbal and non-verbal backchannels are generated at the end of clauses. Fillers signify the system's intention to speak and avoid awkward pauses between turns, which may happen due to latency introduced by utterance and action generation. A filler utterance is generated when the user's end-of-turn is detected and and the system's next utterance does not begin immediately after.

3.4 Action Generation Module

While text-to-speech (TTS) has been extensively studied, body and facial motion generation have received comparatively less attention. Although early linguistic-inspired rule-based gesture generation approaches were proposed (Cassell et al., 2001), few end-to-end models exist that use audio and text input to generate body gesture data (Kucherenko et al., 2020). However, the end-to-end models are not fast enough for real time communication. Additionally, while such models can create smooth movements synchronized with speech rhythm, they often struggle to generate semantic gestures that are essential for making conversations engaging.

To achieve real-time generation of natural body facial motions, we employed a database-driven approach. First, we constructed a database of actions performed by a motion actor, with each action mapped to corresponding text descriptions. When generating a motion, the input text is compared to the texts in the database to calculate the cosine similarity of embedded texts. The action most similar to the input text is then selected. Speech is generated using a TTS model, and mouth movements are generated based on vowel sounds estimated from the synthesized speech.

The combined data for speech, body and facial motions are then sent to a game engine for the agent animation to be rendered. Specifically, we used Sentence-BERT (Reimers and Gurevych, 2019) for text embedding, Google Text-to-Speech for TTS, and Unity for rendering the agent.

3.5 Assessment Module

We propose a speaking proficiency assessment model that takes multimodal dialogue data obtained during the conversation with the user, and simultaneously predicts proficiency levels across one holistic criterion (overall) and five analytic criteria: range, accuracy, fluency, phonology, and coherence. The model has multiple encoder modules to consider a wide range of multimodal features theoretically important in language assessment, such as vocabulary richness (Eguchi and Kyle, 2020), grammatical accuracy (Murakami and Ellis, 2022), fluency (Matsuura et al., 2022; Suzuki et al., 2021), goodness of pronunciation (Saito and Plonsky, 2019), and coherence of discourse (Qin, 2022). To capture these linguistic features, each encoder module has a model as a feature extractor that have been pre-trained in various natural language processing tasks such as grammatical error correction (Omelianchuk et al., 2020), coreference resolution (Otmazgin et al., 2023), and pronunciation scoring (Zhang et al., 2021). The inputs of the model are the user's audio and video, speech-recognized text, and the system's utterance text. After various linguistic features are extracted from these input data by the encoder modules, the outputs of each encoder module are blended by the transformer encoder (Vaswani et al., 2017). Then, the vector sequences, in which the influence of the interaction of the various linguistic features is embedded by the transformer encoder, are input to each network specialized for proficiency assessment of each CEFR category. The output layers for each CEFR category with softmax as activation function output the likelihood of each level. The probabilities are converted to a continuous value score x by the following equation: $x = \sum_{c=1}^{6} c \times p_c$ where p_c represents the probability of level c (1:A1, 2:A2, .., 6:C2) in a category ($\sum_{c=1}^{6} p_c = 1$). After computing the discrete level boundaries of A1-C2 so as to maximize Quadratic Weighted Kappa (QWK) in the validation dataset based on x, a normalized score x' is fed back to the learner so that the boundaries of each level are evenly spaced: A1:[0, 1.0], A2:(1.0, 2.0], ..., C2:(5.0, 6.0].

The model was trained on 232 interview dialogues previously collected, and rated for the CEFR score by trained raters. Figure 3 shows an example of the assessment presented to the user. rationales for the assessment are provided for each category and proficiency, based on the CEFR.

4 Experiments

To evaluate the system in terms of its capability of eliciting ratable speech samples, we conducted two experiments. The first experiment was designed to test the system's adaptive speech sampling strategy in system-led interview dialogues in terms of

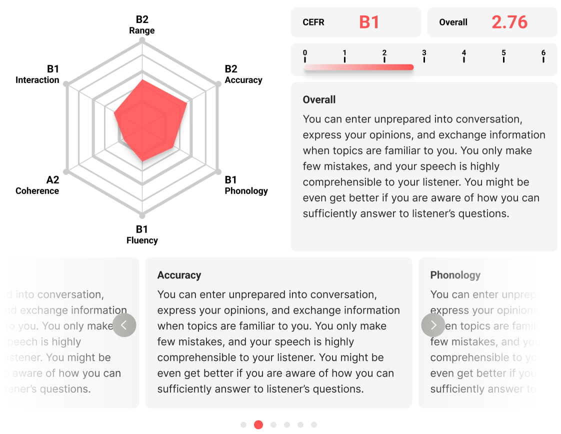

Figure 3: Example of assessment result, including the six core competencies defined by CEFR and the overall score, along with the rationale for these scores.

scoring accuracy as well as the frequency of target phenomenon, that is, linguistic breakdown. The second experiment was set up to gauge the quality of mixed-initiative interaction in roleplay tasks. Given the multifaceted nature of interactional features, the second experiment aims at holistically evaluating the system using human experts' ratings, comparing the scoring reliability between human-interlocutors and the current system.

4.1 Ethical Statement

All data collection for this study, including field testing were reviewed and approved in advance by the ethical review committee ("Ethics Review Procedures concerning Research with Human Subjects") of Waseda University. Prior to all experiments, a consent form outlining the experimental procedures and the use of data (specifically that the recorded audio and video data would be used exclusively for research purposes) was explained to the participants. For high school participants, the procedure was explained to both them and their parents or guardians. Consent was obtained through a detailed consent form, ensuring all parties were fully informed before participation.

4.2 Interview Experiment

We recruited 152 university students with varying levels of oral proficiency to participate in an interview session with our system. Among the participants, 94 participants were female, and 58 were male, with an average age of 20. Each user were given 4 topics, and the whole interview lasted around 15 to 20 minutes. The recordings from these interviews were assessed for CEFR levels by the three trained raters, all of whom hold MA degrees in TESOL or equivalent as well as more than 5 years of teaching experience, and completed a rater training program conducted by researchers in Applied Linguistics. The inter-rater reliability for the CEFR assessment was measured using QWK, which ranged from 0.800 to 0.835, indicating high consistency among raters. In instances of disagreement between raters, the true label was determined through discussion. We then compared the final scores from the assessment module to the human raters' scores (gold standard). The QWK between our system and the gold standard was 0.929, demonstrating very high reliability. The confusion matrix, comparing the model's predictions with human ratings, is shown in Table 1. As evident from the confusion matrix, all model predictions were within one level of the human scoring.

True \ Pred	A1	A2	B1	B2	C1	C2
A1	4	3	0	0	0	0
A2	0	25	6	0	0	0
B1	0	2	32	7	0	0
B2	0	0	3	49	1	0
C1	0	0	0	2	11	3
C2	0	0	0	0	0	4

Table 1: Confusion Matrix of automatic assessment ("Pred") and the gold standard by the trained human raters ("True").

Next, recordings were evaluated for linguistic breakdowns by the same raters. A breakdown was defined as "failure to manage to maintain their speech or respond to the question sufficiently," following the criteria established in (Isbell and Winke, 2019). The occurrence of breakdowns observed in a recording for each proficiency level was observed as follows: A1 and A2 – 100%, B1 – 79.4%, B2 – 42.9%, C1 – 20.9%. C2 proficiency level participants were excluded from this analysis since, theoretically, they would not exhibit breakdowns. These results indicate that students with higher proficiency experienced fewer breakdowns. This trend is expected, as higher proficiency learners, particularly those at B2 or higher levels, may employ a range of linguistic repertoires to strategically navigate around breakdowns (Council of Europe, 2020). However, such strategic behavior can influence other aspects of utterances, including lexical richness and circumlocution. Therefore, the system's adaptive sampling strategy should be evaluated with these considerations in mind. Given the consistency of ratings across levels, it is plausible to argue that despite some room for improvement especially for advanced learners, the current adaptive sampling strategy can elicit ratable speech samples from learners at various proficiency levels.

4.3 Roleplay Experiment

We recruited a total of 75 university students for the roleplay data collection. Among the participants, 54 were female, 20 were male, and one participant did not answer, with an average age of 20. Each participant completed two roleplay sessions with a one week interval: one with a human examiner and one with our system. The order of interlocutor conditions was counterbalanced across participants. Five experienced English tutors were randomly assigned to each student to complete the roleplay in the human session. We adapted a roleplay task used previously in the context of second language assessment literature (Al-Gahtani and Roever, 2018), shown in Appendix A.2. Upon completion of the data collection, four experienced tutors (recruited from the same pool of the examiners) rated each session recording in terms of interactional competence (IC) (Galaczi and Taylor, 2018). Since there was no established rating scale for the assessment of IC, we developed our own CEFR-inspired IC scale. Given our focus on mixed initiatives in interaction, we decided to include two relevant components of IC: **Turn-management** and **Topic-management**. Turn-management is defined as the ability to sustain a cooperative and collaborative conversation through appropriate turn-taking, Topic-management pertains to developing ideas collaboratively toward the intended interactional outcome. The detailed descriptors are shown on Table 7 and 6 in Appendix B.3.

Using a spiral rating design (Eckes, 2015), students' performances were evaluated by alternating pairs of two raters, and each rater assessed only one of the student's videos to mitigate bias such as halo effects. This resulted in a total of 528 raw data points in a 6-level ordinal scale from A1 to C2 (i.e., 66 students $\times$ 2 interlocutor types $\times$ 2 raters $\times$ 2 rating criteria). IC dimensions that could not be observed in the video were marked as unratable.

To evaluate the extent to which our system elicited speech samples that are informative for IC assessment (i.e., ratability), we compared the scoring reliability of IC ratings between the interlocutor conditions (human tutors vs. the system). To systematically control for the effects of rater severity and examinees' proficiency levels, the reliability index was estimated through a series of Many-Facet Rasch Modeling (MFRM; for details, see Appendix B) (Eckes, 2015). Results revealed the comparable level of reliability between the interlocutor conditions of human tutors (0.767) and our system (0.771). See Appendix B.1 for details.

The infit/outfit statistics based on the Rasch model indicates that the AI-based roleplay followed more closely with the assumption of the Rasch measurement model (see Table 4 in Appendix B.2). Taken together, these findings suggest that speech samples elicited through our system are as ratable as human interlocutors for IC assessment, and the system yields psychometrically more consistent data for assessing IC components related to mixed initiatives than human tutors.

4.4 Discussion

The interview experiment demonstrated that the InteLLA system can elicit ratable speech samples for oral proficiency assessment, evidenced by high inter-rater reliability both among human raters and between human and our system. However, we also found the low rate of linguistic breakdowns among high-level participants. This could be attributed to their problem-solving strategies. This suggests that there should be some room for enhancing the ratability of speech samples. Future work, for instance, will need to engage with the accuracy of real-time assessment mechanisms that can operate effectively with fewer samples.

Conversely, the roleplay experiment showed that our system can sufficiently elicit interactional competence for human ratings, specifically turn management and topic management, on par with human interlocutors. Future work includes extending the assessment model with the capability to automatically evaluate interactional competence.

5 Field Testing

We report on the system's performance and stakeholders' satisfactions in the real-world scenarios with university and high school students in Japan.

5.1 Field Testing with University Students

Over the past year (AY2023), the InteLLA system has been deployed to provide assessments to over 10,000 Japanese university students. The system served as a middle-stakes test, where the results were used to determine the appropriate English class level for each student. Tests were administered remotely, with students using their personal computers from home.

To evaluate the system's performance, we randomly selected 300 recordings for detailed analysis. These recordings were scored according to the CEFR level by three trained raters, with the final score determined by majority vote in cases of disagreement. The reliability of the automatic assessment, when compared to human ratings, was found to be 0.869, demonstrating a high level of reliability. However, three out of 300 recordings were deemed unratable, indicating they could not be scored reliably due to technical problems. The user's audio input was too small for ASR to recognize and for the dialogue management to keep the conversation coherently. These issues should have caused significant delays of the system responses due to network problems and consequently interfere with ratable speech elicitation. The results of this field test demonstrate that our system can provide accurate oral proficiency assessments even in uncontrolled, real-world scenarios.

5.2 Field Testing with High School Students

As another field study, a total of 97 students in Chiba prefecture in Japan, all aged 16, participated in eight English conversation sessions over a period of one month. The first and last sessions served as a pretest and posttest and were conducted using using the interview scenarios. The other sessions in between engage them with daily conversations similar to the roleplay format in the second experiment. After each session, students completed a brief questionnaire assessing their learning motivation.

The pretest and posttest scores were compared using a linear mixed-effects model to estimate the group-level improvement, including the random-effect variable of participants to controlling for individual variability in the pretest scores. The analysis revealed a significant increase of 0.30 points ($p < 0.001$) out of 6.0. Among various patterns of score changes, we found A2-level students at the pretest significantly improved and reached B1-levels at the time of the posttest. Students who exhibited notable improvement in this category also showed a positive trend in survey responses over time. These responses included "Enjoyment of the conversations", "Feeling of being able to express themselves", "Comfort and relaxation while speaking", "Desire to speak more in English." Notably, we adopted intact classes for this field-testing study, meaning that these improvements may not solely be attributed to interactions with InteLLA but also to the students' regular English classes during the period of the study. these findings may indicate the potential of using our system as learning materials for English speaking skills. However, this study demonstrates the potential of using multimodal dialogue systems such as InteLLA for developing English speaking skills and language learning motivation.

6 Conclusions and Future Directions

In this paper, we presented InteLLA, a multimodal dialogue system designed for the assessment of oral proficiency. InteLLA is designed to elicit ratable speech samples that display the user's full range of interactional skills. To enhance the ratability of speech elicited, the system is required to adaptively change the difficulty levels of questions to collect learners' linguistic breakdowns as the evidence of their upper limit of proficiency. To capture learners' ability to maintain collaborative conversations, the system is expected to enable mixed initiative interaction where learners need to engage with turn-taking management and topic development. To evaluate InteLLA's usefulness in oral proficiency assessment, we conducted two experiments using interview and roleplay conversations. The results from the interview conversations demonstrated that our system consistently elicited ratable samples especially for lower-level learners, and automatically estimated scores based on those samples exhibited a high agreement with experts' ratings. In roleplay conversations, the ratability of the speech samples elicited through InteLLA were comparable to experienced English teachers in terms of scoring reliability by human raters. The current study takes an interdisciplinary approach to integrating research on multimodal dialogue systems into real-world problems in language learning and testing. We demonstrated that InteLLA is suitable for middle-stake assessments, effectively scaling to accommodate a large number of users. The scalability and accuracy position InteLLA as a valuable tool for oral proficiency assessment in varied contexts.

Although our system showed the potential for reliable standardised assessment of oral proficiency, several challenges should be acknowledged to minimize possible negative consequences of its further real-world implementations (Dai et al., 2024). In the large-scale deployment of InteLLA, Quality of Service (QoS) (Aurrecoechea et al., 1998) factors, such as latency derived from the inference model and communication architecture, as well as the resolution of InteLLA's rendering engine, are critical from the perspectives of Quality of Experience (QoE) (Möller and Raake, 2014) and cost. Future work is expected to evaluate how the configuration of the developed large-scale dialogue system impacts user engagement (Kurata et al., 2024), considering a dialogue quality management system that can efficiently facilitate system improvements.

7 Acknowledgement

The research presented in this study was achieved through funding and support from the project "Technological Development for Next-Generation Artificial Intelligence Evolving with Humans (JPNP20006) / Development of an Online Language Learning Support AI System that Grows with Humans" by New Energy and Industrial Technology Development Organization (NEDO), and "Beyond 5G Seed Creation Program / Research and Development of an XR Communication Infrastructure for Realizing High-Immersion Interaction Experiences with Conversational AI Agents (JPJ012368C06301)" by the National Institute of Information and Communications Technology (NICT).

References

Saad Al-Gahtani and Carsten Roever. 2018. Proficiency and preference organization in second language refusals. *Journal of Pragmatics*, 129:140–153.

Cristina Aurrecoechea, Andrew T Campbell, and Linda Hauw. 1998. A survey of qos architectures. *Multimedia systems*, 6:138–151.

Trevor G Bond and Christine M Fox. 2013. *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486.

Council of Europe, editor. 2020. *Common European framework of reference for languages: learning, teaching, assessment ; companion volume*. Council of Europe Publishing, Strasbourg.

David Wei Dai, Shungo Suzuki, and Guanliang Chen. 2024. Generative ai for professional communication training in intercultural contexts: where are we now and where are we heading? *Applied Linguistics Review*.

R.J. de Ayala. 2022. *The Theory and Practice of Item Response Theory*. Methodology in the Social Sciences Series. Guilford Publications.

Thomas Eckes. 2015. *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*.

Masaki Eguchi and Kristopher Kyle. 2020. Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *Modern Language Journal*, 104(2):381–400.

Glenn Fulcher. 2010. Practical language testing. *Practical Language Testing*, pages 1–352.

Evelina Galaczi and Lynda Taylor. 2018. Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3):219–236.

Felix Gervits and Matthias Scheutz. 2018. Pardon the interruption: Managing turn-taking through overlap resolution in embodied artificial agents. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 99–109, Melbourne, Australia. Association for Computational Linguistics.

Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. A Job Interview Dialogue System That Asks Follow-up Questions: Implemen-tation and Evaluation with an Autonomous Android. *Transactions of the Japanese Society for Artificial Intelligence*, 35(5):D–K43 1–10.

Dan Isbell and Paula Winke. 2019. Actfl oral proficiency interview–computer (opic). *Language Testing*, 36(3):467–477.

Gabriele Kasper and Soo Youn. 2018. Transforming instruction to activity: Roleplay in language assessment. *Applied Linguistics Review*, 9:589–616.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*.

Fuma Kurata, Mao Saeki, Masaki Eguchi, Shungo Suzuki, Hiroaki Takatsu, and Yoichi Matsuyama. 2024. Development and validation of engagement and rapport scales for evaluating user experience in multimodal dialogue systems. In *Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology, Hokkaido, Japan*, pages 1–14.

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal Turn-Taking Model Using Visual Cues for End-of-Utterance Prediction in Spoken Dialogue Systems. In *Proc. INTERSPEECH 2023*, pages 2658–2662.

Judith E. Liskin-Gasparro. 2003. The ACTFL Proficiency Guidelines and the Oral Proficiency Interview: A brief history and analysis of their survival. *Foreign Language Annals*, 36(4):483–490.

Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen, and David Vandyke. 2016. Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of english. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 270–275.

Ryuki Matsuura, Shungo Suzuki, Mao Saeki, Tetsuji Ogawa, and Yoichi Matsuyama. 2022. Refinement of utterance fluency feature extraction and automated scoring of L2 oral fluency with dialogic features. In *Proceedings of the 14th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, page 1312–1320.

Tim McNamara. 1996. Measuring second language performance.

Sebastian Möller and Alexander Raake. 2014. *Quality of experience: advanced concepts, applications and methods*. Springer.

Fabrizio Morbini, David Devault, Kallirroi Georgila, Ron Artstein, David Traum, and Louis-Philippe Morency. 2014. A Demonstration of Dialogue Processing in SimSensei Kiosk. In *SigDial*, pages 254–256. Association for Computational Linguistics.

Akira Murakami and Nick C. Ellis. 2022. Effects of availability, contingency, and formulaicity on the accuracy of english grammatical morphemes in second language writing. *Language Learning*, 72(4):899–940.

Gary J Ockey and Evgeny Chukharev-Hudilainen. 2021. Human versus Computer Partner in the Paired Oral Discussion Test. *Applied Linguistics*, 42(5):924–944.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, page 163–170.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 2752–2760.

Barry O'Sullivan, Cyril J Weir, and Nick Saville. 2002. Using observation checklists to validate speaking-test tasks. *Language testing*, 19(1):33–56.

Jie Qin. 2022. Potential contribution of SFL to task-based research: An examination of planning effects using genre-based theme analysis. *System*, 104:1–12.

Vikram Ramanarayanan, Matthew Mulholland, and Yao Qian. 2019. Scoring interactional aspects of human-machine dialog for language learning and assessment using text features. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 103–109.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Carsten Roever and Naoki Ikeda. 2022. What scores from monologic speaking tests can (not) tell us about interactional competence. *Language Testing*, 39(1):7–29.

Kazuya Saito and Luke Plonsky. 2019. Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, 69(3):652–708.

Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.

Shungo Suzuki, Judit Kormos, and Takumi Uchihara. 2021. The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *Modern Language Journal*, 105(2):435–463.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Proceedings of the 22rd Annual Conference of the International Speech Communication Association*, page 3710–3714.

A Dialogue Samples

A.1 Interview Example

An example of an interview dialogue is shown below. The first to third parts of the adjacency pairs (Schegloff and Sacks, 1973) are referred to as 1st, 2nd, and 3rd, respectively.

Warm-up phase : First, in the warm-up phase, relatively simple conversations such as greetings and small talk are conducted to ease tension and get a general sense of the user's level. In this example, the proficiency assessment model estimates the user to be at the A1 level, and then it moves on to the level-check phase to confirm this estimation.

InteLLA: What is your favorite season? (1st - **topic initiation**)

User: My favorite season is winter. (2nd)

InteLLA: Are there any activities you like to do in winter? (1st - **additional question**)

User: Uh ... Ski and making snowman. (2nd)

InteLLA: That sounds like a lot of fun. (3rd - **acknowledgement**) Could you tell me more about it? (1st - **continuation request)**

User: I like skiing with family. I go every year. (2nd)

Level Check : In this example, the level check was deemed successful because fluency was maintained during the A1 level conversation. Next, the process moves to the "probe" phase, where it is tested whether the user can maintain fluency with a slightly more difficult topic.

InteLLA: Alright. What did you eat for breakfast this morning? (1st - **topic initiation**)

User: I ate uh... Sandwich it is chicken and salad it is very delicious.(2nd)

InteLLA: Do you usually eat breakfast? (1st - **additional question**)

User: Uh yes I always eat breakfast. (2nd)

InteLLA: Oh, that's a healthy diet. (3rd - **acknowledgement**)

User: Yeah. (3rd)

InteLLA: I see what time do you usually eat breakfast. (1st - **additional question; A1 level**)

User: Uh seven A.M. I wake up and I go to kitchen and I eat breakfast. (2nd)

Probe Phase : In this example, a breakdown in fluency occurred during the A2 level conversation, indicating that the user's English conversation ability is estimated to be at the A1 level. In actual dialogues, this probe and level-check routine is repeated several times. Finally, the user's level is confirmed, and a closing dialogue is conducted during the cool-down phase.

InteLLA: Have you ever been to a foreign country? (1st - **topic initiation**)

User: Uh no. I never go to foreign country. (2nd)

InteLLA: Ok. which country would you like to visit in the future? (1st - **additional question**)

User: I would like visit ... Singapore. (2nd)

InteLLA: Why is that? (1st - **continuation request**)

User: Because I want visit ... I like go to nice ... ah nice ... (2nd - **breakdown detected**)

InteLLA: That's ok. Let's move on. (1st - **topic transition**)

A.2 Roleplay Example

A conversation-analytic (CA) transcription of a roleplay dialogue from the experiment is shown in table 2 ("Lending a computer" scenario). Table 3 shows CA notations used in the transcript. In this scenario, InteLLA plays a role of a university friend of the user, asking to borrow a PC from them because hers broke down while she was working on the term paper. This roleplay scenario is an adapted version from previous study on second language learning and assessment (Al-Gahtani and Roever, 2018). This roleplay attempts to assess the following Can-Do statements for the user role:

- **(Conversation; B2+)** Can indicate reservations and reluctance, state conditions when agreeing to requests or granting permission, and ask for understanding of their own position.
- **(Conversation; B2)** Can sustain relationships with users of the target language without unintentionally amusing or irritating them or requiring them to behave other than they would with another proficient language user.

The example roleplay card based on (Al-Gahtani and Roever, 2018) is as follows.

Roleplay card:

```
Read the following instructions carefully.
You have 3 mins at maximum to prepare for
this role play.

Situation You are a student. It's 11
pm. You're working on a term paper that
is due tomorrow morning at 8 am. You are
planning to spend all night finishing the
paper. Now, you decided to take a quick
10-min break. You opened your phone and
noticed a text message from A. Friend A lives
two floors above you, but you have only known
for a month. So you wondered what happened
to A. The text message says that he/she wants
to borrow your PC because theirs broke down.
Because you have ONLY one computer (which
you are using for writing the paper), you
think it is inconvenient for you to lend it
to another person. Now, friends A rings your
doorbell and you are answering it.

Task

  • Explain your situation and first try to
    decline the request.

  • Then negotiate for a solution that works
    for both of you.

  • You can lend it to them but make sure
    that you secure enough time to finish
    your term paper.

  • Do NOT show irritation or annoyance to
    the friend A.
```

B Many-Facet Rasch Modeling

Many-facet Rasch Modeling is a psychometric approach often used in performance assessment (i.e., type of assessment involving a set of raters evaluating performances of the test-takers on predetermined criteria for their skill mastery) (Eckes, 2015). In performance assessments such as the roleplay in Experiment 2 (Section 4.3), multiple factors can add "noises" to the raw score, including but not limited to choice of raters, tasks used, and interlocutors. Many-facet Rasch Model attempts to account for the different sources of variation in the raw score (i.e., facets) and to transform the raw score into a latent logit score. In so doing, it simultaneously computes the harshness of rater and the difficulty of interlocutors on the same logit scale.

The design of experiment 2 yielded the following facets:

- Examinee Interactional Competence (66 persons)
- Rater severity (4 raters)
- Criteria (Turn and Topic-management)
- Interlocutor difficulty (AI and 5 human tutors)

Following Eckes (Eckes, 2015), a full MFRM can be expressed in the following formula:

$$\ln\left[\frac{p_{nlijk}}{p_{nlijk-1}}\right] = \theta_n - \beta_l - \eta_v - \alpha_j - \tau_k$$

where

- p_{nijk} = probability of person n receiving a rating of k on criteria l from rater j when the interlocutor is v,
- p_{nijk-1} = probability of person n receiving a rating of $k-1$ on criteria l from rater j when the interlocutor is v,
- θ_n = ability (= IC) of person n,
- β_l = difficulty of criteria l,
- η_v = difficulty of interlocutor v,
- α_j = severity of rater j,
- τ_k = difficulty of receiving a rating of k relative to $k-1$.

This allows the estimation of locations of each constituent from each facet on a latent logit scale.

B.1 Person separation reliability

After fitting a Rasch model, the reliability of ratings can be calculated to indicate the consistency of person's ability estimate (i.e., location on the latent logit scale) that is beyond the influence of other facets. This is calculated by dividing the amount of variation in Expected A Posteriori (EAP) estimates of person's abilities based on the Rasch model after considering other facets over the total amount of variation in persons' abilities. This reliability estimate ranges from 0 to 1, a high score indicating a high level of reliability. As presented in Section 4.3, the person separation reliability was .767 for human-based roleplays and .771 for AI-based roleplays. For more information about person separation reliability estimate see (Eckes, 2015; Bond and Fox, 2013).

Table 2: Conversation-analytic transcript of a roleplay dialogue ("Lending a computer" scenario)

Speaker	CA transcription
InteLLA:	Ok in this roleplay, I will start talking. Are you ready?
User:	Yes I am.
InteLLA:	Alright three two one.
	(0.6)
InteLLA:	Oh hi.
	(0.5)
User:	[Hi:
	[Sorry to bother you.
	(0.3)
User:	°Hmm°,=
InteLLA:	=Did you look at my message?
User:	.hh yes, I just opened it. I was actually working on my paper and
	I didn't have time tuh check [your message.
InteLLA:	[Hmm.
	(2.8)
InteLLA:	I see. (0.6) I'm in a bit of bind. [My computer died and I need=
User:	[Hmm,
InteLLA:	=to finish my paper.
	(0.5)
User:	Mm-huh=
InteLLA:	=Could I borrow your PC?
	(0.8)
User:	.Hhh I really wish I could. but it's just that I'm still working
	on my term paper, a:nd [it's due] tomorrow mornin' at eight. a::nd
InteLLA:	[hmm,]
User:	.hh I was planning to spend all day:: finishing the pa- all night
	sorry (.) finishing the paper a::nd .hh I'm not really sure if
	I could lend it to you before that.
	(4.0)
InteLLA:	Oh (.) I understand. (0.5) Your paper is important too. (0.6)
	Could I borrow it after you finish?
	(1.1)
User:	°U::m,° we:ll I wish I could if I could finish it by eight
	[.hhh] >I could lend it to you bu'< how many hours do you=
InteLLA:	[Hmm,]
User:	=need (.) for that.
	(2.5)
InteLLA:	I think need about three hours.
	(0.8)
User:	[three hours. °right°]
InteLLA:	[Could I have it from eight] to eleven?
	(0.8)
User:	Oh from eight to eleven? Sure that's after my (.) uh: term paper
	is done so::=
InteLLA:	=hmm=
User:	=I:::m >hoping I could finish it by the deadline,< so: I will work
	on it all night, a:nd >I'll finish it by eight< so: (0.3) is that
	fine? Is your- when's your [paper due?
InteLLA:	[Hmm,
	(4.2)
InteLLA:	My paper is due at 9 in the morning. (0.7) But I need some time to
	finish it.
	(0.7)
User:	Hmmm right but you're not >goin' tuh be able to turn it in< by
	(0.9) [the deadline, right?
InteLLA:	[hmm,
	(4.0)
InteLLA:	I understand your concern.
	(0.6)
InteLLA:	Could I have it a bit earlier (.) maybe at 7?
	(0.4)
User:	Yeah, sure I could finish- try to finish it by seven and .hhh maybe
	you could work on [that] fo::r two hours and then you could=
InteLLA:	[Hmm,
User:	=hopefully turn it in by the deadline?
	(2.5)
InteLLA:	Oh wow! (0.5) That would be great. (0.6) Thank you so much.
	((Roleplay continues))

Table 3: CA notations used in the example transcript

CA notation	Description
Period (.)	Falling intonation
Comma (,)	Slightly rising intonation
Question (?)	Rising intonation
(ss.mm)	Silence in seconds
(.)	A brief silence (shorter than 250ms)
Colon (:)	Lengthening of previous sounds
Dash (-)	A cut-off of speech
Equal sign (=)	Latching (i.e., transition spaces minimized between turns)
Opening bracket ([)	Overlap onset
Closing bracket (])	Overlap offset (optional)
.hh	Inhalation
Hhh	Exhalation
Degree sign (°)	whispering; smaller voice
>WORD<	words pronounced at a faster pace
<WORD>	words pronounced at a slower pace

B.2 Infit/Outfit statistics

In the context of educational measurement, a good assessment instrument should be able to "discriminate" among persons with different ability levels. One important assumption made by Rasch-family models is that score distributions from a good measurement instrument roughly follows a logistic regression with a slope of 1 (de Ayala, 2022). With such a strong assumption on the underlying patterns of data, it is impossible to obtain a perfect fit to the empirical data (Bond and Fox, 2013). Put differently, it is possible to obtain statistics on how well each constituent from each facet performs in relation to this model assumption. Two fit statistics (Infit and Outfit statistics) are commonly used to assess the amount of deviations of persons, raters, interlocutors, etc.

Outfit statistics is an unweighted average of squared standardized residuals (de Ayala, 2022; Bond and Fox, 2013). As such, it tends to emphasize the unexpected scoring patterns that are located far from the person's (or rater's) estimated scores. On the other hand, Infit statistics is a weighted average, which underscores misfit that are close to the persons' (, raters', or interlocutors') location estimates.

An ideal infit and outfit statistics is considered to be close to 1 (Bond and Fox, 2013). Infit/outfit statistics over 1.3 may indicate underfitting, suggesting some erratic scoring patterns. Infit/outfit statistics smaller than 0.7 may indicate overfitting and too deterministic pattern of rating scores. As shown in Table 4, our system showed a good fit to the data according to both infit and outfit statistics. Some variations in misfit patterns were observed for individual human tutors. Two of them (A and B) slightly underfit (although mostly acceptable range) while the other three tutors tended to overfit (which was less problematic in this context).

Table 4: Fit statistics for roleplay interlocutors.

Interlocutor	Outfit	Infit
AI	0.980	0.986
Tutor A	1.253	1.327
Tutor B	1.231	1.144
Tutor C	0.823	0.813
Tutor D	0.884	0.882
Tutor E	0.762	0.774

B.3 CEFR Descriptors

In this section, we introduce the descriptors we adopted for the rating described in Section 4. Table 5 shows the descriptors of the overall oral interaction defined by (Council of Europe, 2020).

Based on the definition of the interactional competence by (Galaczi and Taylor, 2018) describing "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event, " Table 6 and 7 shows our extended descriptors of turn management and topic management respectively.

Table 5: **Overall Oral Interaction**: The ability to engage in spoken communication, managing and participating in conversations with fluency and spontaneity, while effectively responding to and understanding various contexts.

Level	Descriptor
C2	- Can take part effortlessly in any conversation or discussion and have a good familiarity with idiomatic expressions and colloquialisms. - Can express fluently and convey finer shades of meaning precisely. If a problem arises, can backtrack and restructure around the difficulty so smoothly that other people are hardly aware of it.
C1	- Can express fluently and spontaneously without much obvious searching for expressions. - Can use language flexibly and effectively for social and professional purposes. - Can formulate ideas and opinions with precision and relate contributions skilfully to those of others.
B2	- Can interact with a degree of fluency and spontaneity that makes regular interaction with users of the target language quite possible. - Can take an active part in discussion in familiar contexts, accounting for and sustaining views.
B1	- Can deal with most situations likely to arise while travelling in an area where the language is spoken. - Can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).
A2	- Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. - Can handle very short social exchanges, even though understanding enough to keep the conversation going oneself is not usually possible.
A1	- Can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate and help formulate what is being tried to express. - Can ask and answer simple questions in areas of immediate need or on very familiar topics.

Table 6: **Turn Management**: The ability to keep the conversation cooperative and collaborative, in relation to the expected balance of contributions to the interaction among participants by means of socioculturally and pragmatically appropriate turn-taking.

Level	Descriptor
C2	- Can interact with ease by (skillfuly) interweaving his/her contributions into the conversation.
C1	- Can initiate, respond appropriately, and balance conversations, linking contributions to those of other speakers.
B2	- Can initiate discourse appropriately, actively invite the partner, take their turn when appropriate, and end conversation when they need to, though they may not always do this elegantly. - Can gain time and keep the turn while formulating what they want to express (e.g. "That's a difficult question to answer"). - Can maintain and balance a natural and colloaborative flow to the interaction (no long pauses within/between turns, no dominating interruptions). - Can make prompt and relevant responds appropriately, linking contributions to those of other speakers.
B1	- Can start up a conversation and help keep it going by asking people relatively spontaneous questions about a special experience or event, expressing reactions and opinions on familiar subjects. - Can intervene in a discussion on a familiar topic, using a suitable phrase to get the floor.
A2	- Can ask and answer questions about habits and routines, pastimes and past activities, and plans and intentions. - Can participate in short conversations in routine contexts on topics of interest.
A1	- Can ask and answer simple questions, initiate and respond to simple statements in areas of immediate need or on very familiar topics, including the factual information of themselves and other people (e.g. their home country, family, school).

Table 7: **Topic Management**: The ability to develop ideas collaboratively, as opposed to extending their own speech, in relation to the communicative purpose and outcome and the topic of the interaction

Level	Descriptor
C2	- Can advise on or discuss sensitive issues without awkwardness, understanding colloquial references, dealing diplomatically with disagreement and criticism. - Can link contributions skilfully to those of others, widen the scope of the interaction and help steer it towards an outcome.
C1	- Can develop others'/own ideas and relate own contribution skillfully to that of others.
B2	- Can take the initiative to introduce and contribute relevant new ideas in a discussion, extending the partner's thoughts and working towards joint decisions. Can effectively summarize the discussion at key stages, evaluate the main points within their area of expertise, and propose the next steps to advance the interaction. Can enhance the interaction by providing comments and asking questions that deepen collective understanding.
B1	- Can ask others to explain their ideas, give or seek personal views and opinions, and summarize the opinions or the points reached in an interaction. - Can help focus the argument and keep the development of ideas on course.
A2	- Can exchange what to do in the evening or at the weekend / what to do, where to go and make arrangements to meet.
A1	- Can exchange likes and dislikes for sports, foods, etc., using a limited repertoire of expressions, when addressed clearly, slowly and directly.