Mao Saeki

Waseda University
Wasedamachi-27, Shinjuku-ku
Tokyo, Japan

saeki@pcl.cs.waseda.ac.jp
https://saeki-m.github.io/

1 Research interests

My research interests lie in **multimodal dialog systems**, especially in **turn-taking** and the understanding and generation of **non-verbal cues**. I am also interested in bringing dialog system research into **industry**, and making virtual agents practical in real world setting.

I have been working on the Intelligent Language Learning Assistant (InteLLA) system, a virtual agent designed to provide fully automated English proficiency assessments through oral conversations[1]. This project is driven by the practical need to address the lack of opportunities for second-language learners to assess and practice their conversation skills.

While recent advancements in large language models (LLMs) have enabled natural conversation, effective assessment requires several components that current LLMs cannot fully achieve. Based on interviewer guidelines for oral proficiency assessment (Liskin-Gasparro, 2003), the following functionalities have been identified as necessary for the agent to possess for an effective assessment:

1. Automated Proficiency Assessment
2. Confusion Detection
3. Multimodal Turn-Taking Prediction

My past research has focused on solving each of these problems, which will be explained in subsequent sections. Additionally, these research outcomes have been integrated into the InteLLA agent and evaluated for its effectiveness in end-to-end assessment.

1.1 Automated Proficiency Assessment

For scalable and reliable evaluations, an automated assessment model is essential. While handcrafted lexical and acoustic features have been extensively investigated for assessment, end-to-end approaches, particularly those utilizing visual features like facial expressions and eye gaze—critical components of interaction—have not been thoroughly explored. I proposed a multimodal oral proficiency assessment model incorporating lexical, prosodic, and visual cues (Saeki et al., 2021). The results demonstrated that end-to-end approaches using deep neural networks achieve a higher correlation with human scoring compared to those employing handcrafted features. Furthermore, the effectiveness of the modalities was found to be in the order of lexical, acoustic, and visual features.

Assessment is also important during the interview. For the user to demonstrate their full range of ability, they must be challenged with appropriate levels of questions, according to assessment during the interview. Saeki et al. (2022a) investigated the feasibility of incremental assessment of oral proficiency using an adaptive test format.

1.2 Confusion Detection

Language learners often face confusion, where they fail to understand what the system has said and may be unable to respond, leading to a conversational breakdown. Detecting such states and keeping the conversation moving forward by repeating or rephrasing system utterances is crucial. In Saeki et al. (2022b) we collected a dataset of user confusion using a psycholinguistic experimental approach and identified seven multimodal signs of confusion, some unique to online conversations. We trained a classification model of user confusion using these features. An ablation study showed that features related to self-talk and gaze direction were most predictive.

1.3 Multimodal Turn-Taking Prediction

Language learners often produce long silences while formulating their responses. Such pauses should not be interrupted, however, the system should promptly take its turn when the user has finished speaking to achieve a natural conversation flow. While the effectiveness of visual cues—such as gaze, mouth, and head movements—has been suggested, few studies have fully incorporated them into turn-taking models. We proposed a multimodal model for predicting the end-of-turn probability in spoken dialogue systems(Kurata et al., 2023). An ablation study on visual features showed that eye movements contributed more significantly than mouth and head movements. Additionally, an end-to-end visual feature extraction model utilizing 3D-CNN was employed to comprehensively capture these visual cues. Combining visual features with acoustic and verbal information, the AUC score for end-of-turn prediction improved from 0.896 to 0.920, demonstrating the effectiveness of these visual cues.

[1] https://youtu.be/RzCq5Z4cDBk?feature=shared

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 18–20
September 16–17, 2024.

1.4 InteLLA System Evaluation

The primary challenge in using dialogue systems for reliable language assessment of interactional skills lies in obtaining ratable speech samples that demonstrate the user's full range of abilities. We developed a multimodal dialogue system that employs adaptive sampling strategies and enables mixed-initiative interaction through extended interviews and role-play dialogues (Saeki et al., 2024). The interview is a system-led dialogue aimed at evaluating the user's overall proficiency. The system dynamically adjusts question difficulty based on real-time assessment to induce linguistic breakdowns, providing evidence of the user's upper proficiency limits. The role-play, 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, involving an interview dataset of 152 speakers, our system demonstrated high accuracy in automatically assessing overall proficiency. However, linguistic breakdowns were less likely to occur among high-proficiency users, indicating room for improving 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 were as ratable for interactional competence as those elicited by experienced teachers, demonstrating our system's capability in conducting interactive conversations. Finally, we reported on the deployment of our system with over 10,000 students in two real-world testing scenarios.

1.5 Future Planned Work

Using the InteLLA system, a future direction I am planning is to automatically evaluate interactional competence the user is able to demonstrate. Interactional competence is an important metric in the context of language assessment; however, I believe it could also benefit Spoken Dialogue Systems (SDS). For example, interactional competence has been identified in the field of language assessment to include functions such as turn management strategy, which consists of timing, turn-allocation, overlap resolution, and preference organization. Measuring and closing the gap of interactional competence of turn management strategy between an SDS and a human interlocutor would mean the SDS is recognized more similarly to a human interlocutor, which is a measure of improvement for the SDS. Furthermore, if the relationship between the interlocutor's and user's turn management strategies is identified, we can effectively improve the system to achieve a more authentic spoken dialog experience.

2 Spoken dialogue system (SDS) research

In light of recent advancements in large language models (LLMs) and multimodal LLMs, I believe that this generation will witness the widespread usage of spoken dialogue systems (SDS) in everyday life. Currently, many SDS frameworks depend on multiple models, external APIs, and networks. An imperative field of research in the coming years will be the continuous monitoring and quality assurance of these complex systems. Automatic detection of bugs and conversation experience issues will be crucial for the widespread usage of SDS. Additionally, it will be essential to ensure that subsequent changes do not introduce new problems or degrade overall system performance.

Another exciting advancement for SDS would be the development of fully end-to-end models. For instance, models like GPT-4o exhibit impressive expressiveness; however, they are likely not yet capable of fully interactive conversations as they process user utterances in chunks. Such models would struggle with complex turn-taking phenomena like allowing users time to think, backchanneling, and handling overlapping responses without external modules. Research on incremental models that process and output audio in real-time will be essential to overcome these limitations and achieve more natural and fully duplex interactions.

3 Suggested topics for discussion

- Quality Assurance and Testing of SDS: With the rapid deployment of Spoken Dialogue Systems (SDS) in real-world applications, similar to other industrial software, automated testing is becoming crucial. How can we ensure that updates to the system do not degrade performance and genuinely improve the conversational experience? What methodologies or frameworks can be employed for objective and automated testing?

- Identifying and Implementing Improvements: Virtual agents in SDS have several aspects that can be enhanced, such as speech content, Text-to-Speech (TTS) quality, turn-taking mechanisms, and motion. How can we efficiently pinpoint bottlenecks in user experience to prioritize and implement improvements effectively?

- Leveraging Larger and Multimodal Models: The advent of large language models (LLMs) with multimodal capabilities suggests significant potential for handling spoken dialogue with natural speech. Is this the future direction for SDS development? What roles can university researchers and young professionals without huge computational resource play in the race for bigger models?

References

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. https://doi.org/10.21437/Interspeech.2023-578.

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. https://doi.org/10.1111/j.1944-9720.2003.tb02137.x.

Mao Saeki, Weronika Demkow, Tetsunori Kobayashi, and Yoichi Matsuyama. 2022a. A woz study for an incremental proficiency scoring interview agent eliciting ratable samples. In *Conversational AI for Natural Human-Centric Interaction*. Springer Nature Singapore, Singapore, pages 193–201.

Mao Saeki, Masaki Eguchi, Hiroaki Takatsu, Shungo Suzuki, Shungo Suzuki, Fuma Kurata, Ryuki Matsuura, Kotaro Takizawa, Sadahiro Yoshikawa, and Yoichi Matsuyama. 2024. Intella: Intelligent language learning assistant for assessing language proficiency through interviews and roleplays. *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* page to appear.

Mao Saeki, Yoichi Matsuyama, Satoshi Kobashikawa, Tetsuji Ogawa, and Tetsunori Kobayashi. 2021. Analysis of multimodal features for speaking proficiency scoring in an interview dialogue. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. pages 629–635. https://doi.org/10.1109/SLT48900.2021.9383590.

Mao Saeki, Kotoka Miyagi, Shinya Fujie, Shungo Suzuki, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoichi Matsuyama. 2022b. Confusion detection for adaptive conversational strategies of an oral proficiency assessment interview agent. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2022-September:3988–3992. https://doi.org/10.21437/Interspeech.2022-10075.

Biographical sketch

Mao Saeki is a Ph.D. student in Computer Science at Waseda University. His research interests focus on multimodal conversational AI, particularly in the understanding and generation of non-verbal cues. He developed the InteLLA system, a virtual agent designed to automatically assess the English proficiency of language learners. Additionally, he is a founding member of Equmenopolis Inc., a company dedicated to integrating SDS research into societal applications.