Sadahiro Yoshikawa

Equmenopolis, Inc.
Japan Advanced Institute of Science and Technology
yoshikawa@equ.ai
s2230033@jaist.ac.jp

1 Research interests

My research interests lie in the area of **how users feel when using spoken dialogue systems (SDSs)**, including measuring **user satisfaction** and scoring **naturalness of voice conversation** such as speed, tone, response timing, and turn-taking events. In my ongoing master's thesis, I am working on response timing estimation.

As a research engineer in a company, I have been working on quality assurance for the Intelligent Language Learning Assistant (InteLLA) system, a virtual agent providing English proficiency assessments through oral conversations (Matsuyama et al. (2023)). The quality assurance for this system is towards consistently making users feel "Wow!". In the conversation with a virtual agent, the quality of the animation as well as the voice is important. I am trying to define the metrics for each critical point one-by-one based on the user satisfaction. There are also the viewpoint of **cost efficiency** when building **SDSs on a large-scale**. Building a framework that optimizes costs while maintaining user satisfaction is critical to long-term SDS operation.

1.1 Response Timing Prediction

Response timing has important role for SDS for not only the impression but also the intention of the utterance. For instance, the experience by Roberts and Francis (2013) showed the perceived willingness begins to drop after 600ms, and then clearly and significant steps down from 700 to 800 ms, and the corpus analysis by Kendrick and Torreira (2015) suggests the proportion of dispreferred actions is significantly greater than that of preferreds in case of the responses after approximately 700 ms and the gaps longer than the norm (>300 ms) decrease the likelihood of an unqualified acceptance.

Researchers built models to predict the actual response timing using LTSM (Roddy and Harte (2020)), with syntactic completeness prediction model (Sakuma et al. (2023)). Although most response timing estimation models are regression models, even if the error is the same at 200 ms, the influence of the error at 400 ms and 1500 ms is different. Furthermore, it is hard to confirm how much the error will affect human perceptions. I would like to deal with the difficulty of the response timing perception of humans utilizing deep neural network models.

Besides, several previous studies have indicated that the distribution of response timing varies depending on the conversation situation, such as the nature of conversations such as task-oriented or not (Levinson and Torreira (2015)) and the speaker's language (Stivers et al. (2009)). Therefore, when applying timing estimation models to SDS, we must also consider where the application will be located.

1.2 Future Turn-taking Prediction

The faster turn-taking event prediction with high accuracy, the more inference cost can be used for the quality for the actions of SDSs at the turn-take. If enough time can be used for inference, the inference cost may also be used not only for the actions but also for user adaptation or adjustment of response timing. Therefore, future turn-taking prediction is crucial for SDSs.

Ekstedt and Skantze (2022) proposed Voice Activity Projection (VAP) model forward to predicting future voice activity. The predictive task of VAP uses VAP window, which is discretized into a fixed number of bins as each bin indicates the probability whether the voice is active or not. When a VAP window is set to predict future voice activities, the performance of the task indicates the ability for future prediction. In the paper, the VAP model (referred to as the Discrete model in the paper) enumerates each possible configuration of a VAP window as separate states. In the model, a VAP window can be viewed as sequence of bits where the total number of states grows exponentially as 2^{n_bits}. For instance, the number of bins to 4 for each speaker was in 8 total bits in the paper, thus the output dimension of the model is 256, indicating 256 different possible states. This discrete method resulted in high performance on the task related to turn-taking events in near future (S-pred).

There are extended researches for the VAP model, such as the CPU inference (Inoue et al. (2024)) and multimodal VAP (Onishi et al. (2023)). However, there is some challenges at real-time inference of turn-taking events in practice. Raux and Eskenazi (2009) indicates the lower latency, the more user interruption is caused in an exper-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 32–34
September 16–17, 2024. ©2024 Association for Computational Linguistics

iment with the users of an automated call system for bus information. Moreover, SDSs need an algorithm to actually trigger a turn-taking cue using the predicted probability. Ruede et al. (2017) proposed an implementation using local maximum value within a window of an user utterance to trigger a backchannel. Although this is one of the solution for this issue, the window is only used for the model producing the local maximum value curve such as LSTM, not VAP, and a delay occurs because there is a margin between the maximum value in the window and the end of the window. Lala et al. (2019) showed an implementation using consecutive positive predictions as a turn-taking cue. However, this model aims at turn-taking that combines filler and eye-gaze, it thus needs to be verified whether it can be applied to VAP.

1.3 Allowable Threshold for Overlap

Skantze (2021) explains overlap has two types: cooperative and competitive overlap. There is so far very little work on how to produce cooperative overlapping speech, and there is a system regarding overlap in DeVault et al. (2009), in order to help the user to complete the sentence, possibly overlapping with the user's speech. However, the system often resulted in the agent being perceived as barging in and interrupting the user's speech. Unlike cooperative overlaps, competitive overlaps need some kind of resolution mechanism (to determine who should get the floor).

I wondered how much the competitive overlap of SDSs is bad. How much does overlap affect SDS user experience (UX)? Is it enough to add resolution mechanism even if the overlap occurs many times? Is there a way to get users to allow the competitive overlap? In a large-scale usage of SDS, unexpected competitive overlap is inevitable. Therefore, I would like to make the metrics how the effect of the overlap for SDS UX compared with other violations. Besides, I'm exploring a turn-taking strategy that get users to allow overlaps.

1.4 Allowable Threshold for Disturbed Video

InteLLA, which is our virtual agent communicates with the users through video streaming for easy usage to the person who has less computing resources, so we are managing the computing resources for the animation and the network resources. Moreover, InteLLA provides English proficiency assessments, thus the stable conversation is required. In a large-scale usage of SDS, if the excess optimization of computing resources or shortage of network resources is occurred, the agent turn-taking becomes unstable such as disturbed, choppy, delayed, etc. Therefore, optimizing the costs also requires managing computing resources and network latency to ensure video quality for the stable conversation.

However, it is unclear to what extent animation quality affects proficiency ratings and user satisfaction. Moreover, to the best of my knowledge, there is no solid indicators for the video quality of SDSs. If we measure the relationship between the resources and the quality precisely, the users can be use InteLLA at less network resources and InteLLA can work with less computing resources. Therefore, I'm exploring the metrics for precisely measure of the animation to ensure proficiency ratings and user satisfaction.

There is a reference for quality control indicators regarding the video itself. Min et al. (2024) indicates there is full-reference (FR) or no-reference (NR) analysis. For assessing corrupted video quality, my research will be started with FR.

2 Spoken dialogue system (SDS) research

I think the field of dialogue system will become closer to front-end or game engineering if SDSs are easy to custom and publish to internet like a cloud service by an individual in 5 to 10 years. Software for SDSs will be more complex, and be OS-dependent like browsers, and well-known software will be utilized without many people understanding the detailed mechanisms, and some SDSs will be designed by designers. Over the next 5 to 10 years, the number of software that can use for SDSs is expected to increase explosively. At that time, I think what researchers of SDSs should do is understanding the mechanisms of each modules of SDSs and defining the evaluation criteria for SDSs to guide engineers to build stable, secure, and reproducible SDSs. After the next 10 years, the core technologies for SDSs will be expanded by other modalities such as virtual reality and sensing technologies. Therefore, even if the SDSs we called today will be generalized, researchers will be required to extend another modalities for SDSs.

3 Suggested topics for discussion

- How to evaluate the user satisfaction regarding turn-taking events?
- How to implement future turn-taking prediction in practice?
- How to evaluate the quality of animation behaviors in conversations?

References

David DeVault, Kenji Sagae, and David Traum. 2009. Can i finish? learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, USA, SIGDIAL '09, page 11–20.

Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*. pages 5190–5194. https://doi.org/10.21437/Interspeech.2022-10955.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using voice activity projection. In *ArXiv*. https://arxiv.org/abs/2401.04868.

Kobin H. Kendrick and Francisco Torreira. 2015. The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes* 52(4):255–289. https://doi.org/10.1080/0163853X.2014.955997.

Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, ICMI '19, page 226–234. https://doi.org/10.1145/3340555.3353727.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6:731. https://doi.org/10.3389/fpsyg.2015.00731.

Yoichi Matsuyama, Mao Saeki, Hiroaki Takatsu, Ryuki Matsuura, Fuma Kurata, and Shungo Suzuki. 2023. Intella: Dialog-based english speaking assessment agent that elicits learnerapos;s language ability. *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN* 79(3):162–169. https://doi.org/10.20697/jasj.79.3$_1$62.

Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. 2024. Perceptual video quality assessment: A survey. https://arxiv.org/abs/2402.03413.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *Proceedings of the 11th International Conference on Human-Agent Interaction*. Association for Computing Machinery, New York, NY, USA, HAI '23, page 13–21. https://doi.org/10.1145/3623809.3623837.

Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*. Association for Computational Linguistics, Boulder, Colorado, page 629. https://doi.org/10.3115/1620754.1620846.

Felicia Roberts and Alexander L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America* 133(6):EL471–EL477. https://doi.org/10.1121/1.4802900.

Matthew Roddy and Naomi Harte. 2020. Neural Generation of Dialogue Response Timings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 2442–2452. https://doi.org/10.18653/v1/2020.acl-main.221.

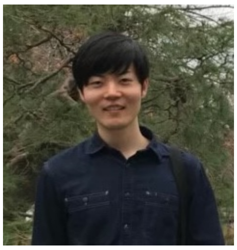
Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings. In *Proc. Interspeech 2017*. pages 879–883. https://doi.org/10.21437/Interspeech.2017-1606.

Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2023. Response Timing Estimation for Spoken Dialog Systems Based on Syntactic Completeness Prediction. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. pages 369–374. https://doi.org/10.1109/SLT54892.2023.10023458.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67:101178. https://doi.org/10.1016/j.csl.2020.101178.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106(26):10587–10592. https://doi.org/10.1073/pnas.0903616106.

Biographical sketch



Sadahiro Yoshikawa is a Research Engineer at Equmenopolis. He is also a master's mature student at the Graduate School of Computer Science, Japan Advanced Institute of Science and Technology. Formerly, he was a freelancer as a Data Engineer. His interest is voice response quality of SDSs and defining metrics for SDSs towards quality assurance.