Simulating Turn-Taking in Conversations with Delayed Transmission

Thilo Michael
Quality and Usability Lab
Technische Universität Berlin
thilo.michael@tu-berlin.de

Sebastian Möller
Quality and Usability Lab
Technische Universität Berlin
German Research Center for AI (DFKI)
sebastian.moeller@tu-berlin.de

Abstract

Conversations over the telephone require timely turn-taking cues that signal the participants when to speak and when to listen. When a two-way transmission delay is introduced into such conversations, the immediate feedback is delayed, and the interactivity of the conversation is impaired. With delayed speech on each side of the transmission, different *conversation realities* emerge on both ends, which alters the way the participants interact with each other. Simulating conversations can give insights on turn-taking and spoken interactions between humans but can also used for analyzing and even predicting human behavior in conversations. In this paper, we simulate two types of conversations with distinct levels of interactivity. We then introduce three levels of two-way transmission delay between the agents and compare the resulting interaction-patterns with human-to-human dialog from an empirical study. We show how the turn-taking mechanisms modeled for conversations without delay perform in scenarios with delay and identify to which extend the simulation is able to model the delayed turn-taking observed in human conversation.

1 Introduction

Turn-taking in human conversations has proven to be influenced by many auditory, visual, and contextual cues. Especially in telephone conversations, where no visual cues are present, people rely on the immediacy of signals in prosody and content to perform smooth and uninterrupted turn-taking. Investigating the influence of delay on conversations has been a focus in telephone quality research for a long time, where the goal is to study how degradations of packet-switched VoIP-transmissions influence the conversation structure and thus, the perceived quality (ITU-T Recommendation P.805, 2007; ITU-T Recommandation G.107, 2011). But also in the field of human-computer interaction, where Spoken Dialogue Systems (SDS) with realistic turn-taking have become feasible, it is of interest to study how humans interact and react to delayed voice transmission.

It has been shown that the perception of changes in transmission time not only depends on the duration of the delay but that the effects on the conversations also vary with the type of conversation itself (Hammer et al., 2005). Concretely, conversations with lower interactivity, i.e., slower speaker alternation rate and less turn-taking, are not as prone to be affected by transmission delay than conversations with higher interactivity. Simulating a conversation does not only give insights into the interactivity patterns that arise during a conversation but can also used to predict events and behaviors. In such a simulation, two dialog systems exchange information through a speech channel. Information is processed in increments to allow for a turn-taking mechanism and structured dialog (Michael and Möller, 2020).

In this paper, we present a simulation with different levels of interactivity and evaluate how a probability-based turn-taking function models the behavior in conversations under the influence of transmission delay. For this, we simulate two different goal-oriented conversation scenarios standardized by the ITU, namely the Short Conversation Test (SCT) with a low conversational interactivity and the Random Number Verification test (RNV) with a high conversational interactivity (ITU-T Recommendation P.805, 2007). We simulate conversations with $0ms$, $800ms$, and $1600ms$ delay and compare metrics of interactivity like speaker alternation rate, gaps, overlaps, and pauses, as well as unintended interruption rates to human-to-human conversations with the same delay conditions.

2 Related Work

Turn-taking in conversations is a long-studied phenomenon (Sacks et al., 1974), with recent work focusing on human turn-taking behavior in conversations (Lunsford et al., 2016), end-of-turn prediction (Liu et al., 2017; Skantze, 2017) and rule-based turn-taking models (Selfridge and Heeman, 2012; Baumann, 2008; Michael and Möller, 2020). While the effects of transmission delay on turn-taking conversations have been studied in the field of speech transmission quality (Kitawaki and Itoh, 1991; Egger et al., 2010), it has to the best of our knowledge not been modeled. However, the influence of delay on the perception of the conversational quality has been modeled by the E-model (ITU-T Recommandation G.107, 2011).

Due to the delayed arrival of turn-taking signals, transmission delay affects the flow of a conversation (Hammer, 2006). However, the degree to which turn-taking and the interactivity of a conversation is degraded also depends on the interactivity of the conversation itself (Raake et al., 2013; Egger et al., 2012). To evaluate those dependencies, conversation tests with distinct levels of conversational interactivity (CI) have been standardized, during which participants perform goal-oriented tasks with an interlocutor. One prominent conversation test with a high CI is the RNV test, where participants alternatingly exchange a list of numbers organized in 4 blocks (Kitawaki and Itoh, 1991). An example of a conversation test with low CI is the SCT, where participants solve real-world tasks like ordering pizza or booking a flight.

Parametric Conversation Analysis (P-CA) is a framework to assess the structure of conversations programmatically (Hammer, 2006). With an independent voice activity detection of the two speakers, four conversation states can be derived: M ("mutual silence"), D ("double talk"), A ("speaker A") and B ("speaker B") (Lee and Un, 1986; ITU-T Recommendation P.59, 1993). Based on these four states interactivity metrics like the speaker alternation rate (SAR), interruption rate (IR), as well as turn-taking information like gaps and overlaps between speaker turns, can be calculated (Hammer et al., 2005; Lunsford et al., 2016). For delayed conversations, the unintended interruption rates (UIR) measures the number of interruptions that were caused by the delay and were not intended to be interrupting the interlocutor (Egger et al., 2010).

As conversation simulations focus on turn-taking, they need to respond to incoming signals in a timely matter and thus need to process data *incrementally*. The incremental processing on the scale of a complete dialogue system has been proposed by Skantze and Schlangen (Schlangen and Skantze, 2011) and implemented in InproTK (Baumann and Schlangen, 2012) and Retico (Michael and Möller, 2019).

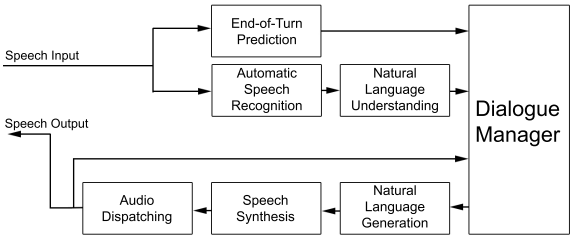

Figure 1: Schematic of an incremental spoken dialogue network containing parts for speech understanding and end-of-turn prediction on the top, the and dialogue managing unit on the right and the speech generation and audio dispatching on the bottom.

3 Simulation Setup and Turn-Taking

The simulation is based on a set of conversation tests carried out with 58 untrained participants who were 18 to 71 of age (M: 32, SD: 13.48), of which 48.2 percent identified as female. During the experiments, each pair of participants carried out SCT and RNV conversations with end-to-end one-way transmission delays of $0ms$, $800ms$, and $1600ms$, resulting in 174 recorded conversations. For the simulation, one scenario was selected from each conversation type, and 20 SCT conversations and 20 RNV conversations at $0ms$ delay were annotated with dialogue acts, transcripts, and turn-taking information. 20 different conversations from each conversation type were used to evaluate the simulation.

The simulation was implemented using the incremental processing pipeline of the retico framework (Michael and Möller, 2019). It consists of two spoken dialogue systems (agents) that are connected through a simulated transmission network that is able to introduce delay to both agents. A schematic view of the incremental modules of one agent in the simulation is shown in Figure 1. The speech input and output, as well as natural language understanding modules, are created by specifically recognizing the annotated empirical conversations. Language generation and synthesis is handled by

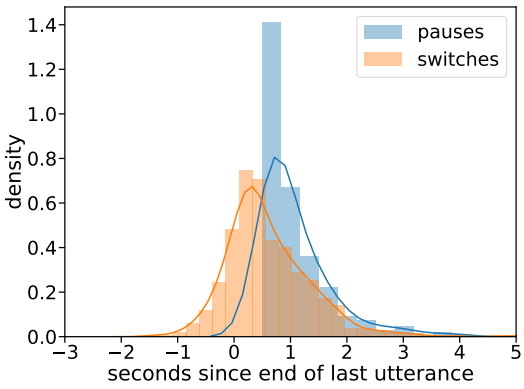

Figure 2: Distribution of speaker switches (orange) and pauses/turn-keeps (blue) in SCT conversations without delay as measured by the seconds since or until the end of the last utterance.

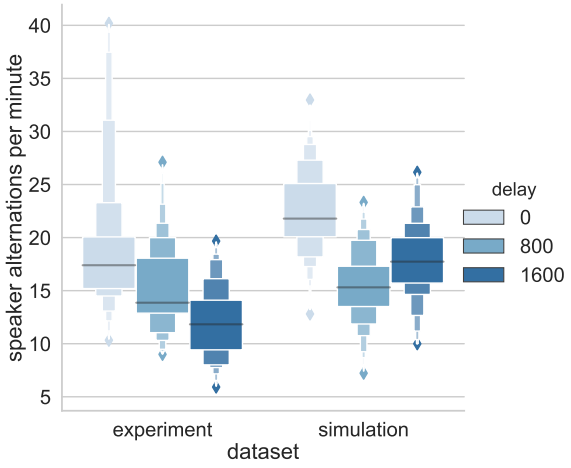

Figure 3: Speaker alternation rate for empirical and simulated SCT conversations at 0, 800, and 1600 ms delay.

transmitting utterances cut from the empirical data so that the length and content of the utterances match. An end-of-turn prediction module predicts the time until the end of the utterance, and an audio dispatching module reports the progress of the current utterance to the dialogue manager of the same agent. The dialogue manager uses agenda-based dialog management to fulfill the goal-oriented tasks of the SCT and RNV scenarios, and it also handles turn-taking.

The turn-taking of the agents in the simulation is modeled by probability distributions that are based on the work by Lunsford et al. (Lunsford et al., 2016). We calculated the distributions of turn-switches (*gaps* and *overlaps*) as well as turn-keeping (*pauses*) as shown in Figure 2. These distributions are measured respective to the end of the last utterance so that negative values correspond to double-talk, and positive values correspond to mutual silence. The cumulative distribution of the pauses and switches were fitted with a logistic regression and inverted to form a model for turn-switches (Equation 1) and turn-keeping (Equation 2).

$$0.27 - 0.322581 \log(\frac{1}{r} - 1) \quad (1)$$

$$1.10641 - 0.161705 \log(\frac{1}{r} - 1) \quad (2)$$

By selecting $r \in [0, 1]$ randomly from a uniform distribution and treating switches and pauses as equal alternatives, the agent in the simulation can perform turn-taking in the simulation. Depending on which agent is currently speaking, the dialogue manager decides when to make a pause or a speaker switch. This way, the models of pauses and switches compete at every end of a turn. To prevent prolonged interruption (e.g., when both agents start speaking at the same time), the dialogue manager stops the speech production when double talk occurs in the middle of utterances.

4 Results and Discussion

To evaluate the simulation approach, we simulated 100 RNV and 100 SCT conversations, each with 0, 800, and 1600 ms transmission delay. This results in 600 simulated conversations that we compare to the 174 conversations recorded in the experiment.

The comparison of the states of the SCT conversations (Figure 4) and RNV conversations (Figure 5) shows that the distinct levels of interactivity between these two types of conversations are also visible in the simulated conversation. When introducing delay, the state probabilities of the empirical data and the simulated conversation for mutual silence, speaker a and speaker b show similar changes. However, these effects stagnate for the simulated conversations at 1600 ms. This can also be seen when comparing the speaker alternation rate (Figure 3) of the simulated SCT conversations. There, the drop in speaker alternations due to increased delay is visible for $800ms$ but increases again for $1600ms$, contrary to the behavior of the empirical conversations. This seems to indicate changes in the turn-taking behavior with an increased level of transmission delay.

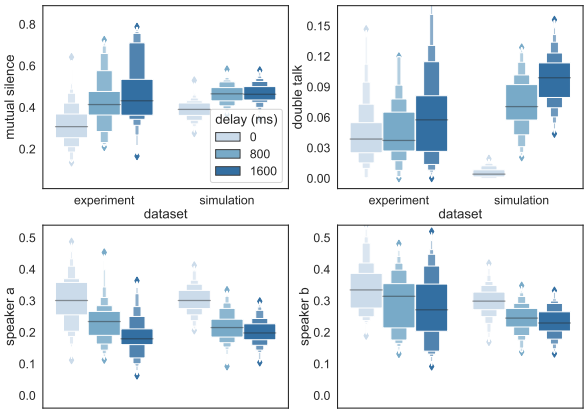

Figure 4: Comparison of states probabilities mutual silence, double talk, speaker a and speaker b between the empirical and simulated SCT conversations at 0, 800, and 1600 ms delay.

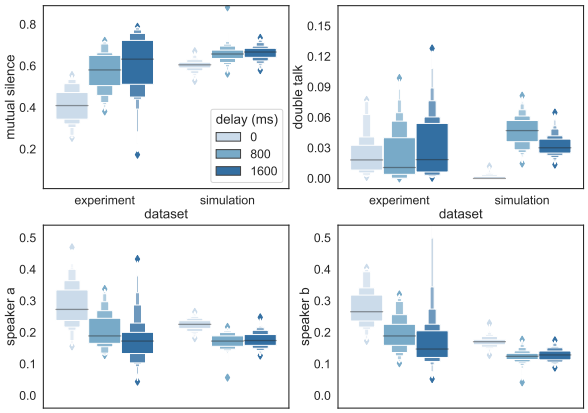

Figure 5: Comparison of states probabilities mutual silence, double talk, speaker a and speaker b between the empirical and simulated RNV conversations at 0, 800, and 1600 ms delay.

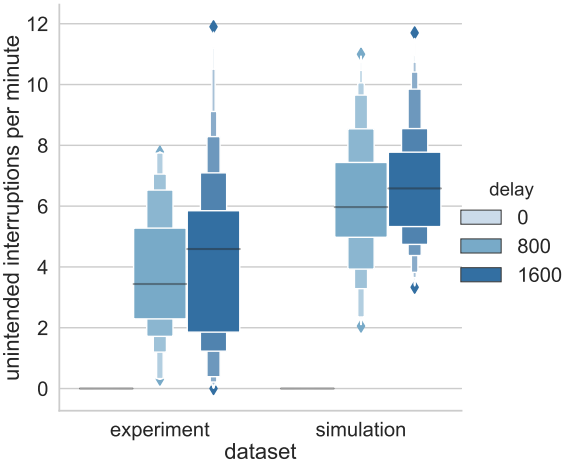

Figure 6: Unintended interruption rate for empirical and simulated SCT conversations at 0, 800, and 1600 ms delay.

While the state probabilities for double talk stay almost constant for SCT and RNV empirical conversations, it increases strongly in the simulations. It also stagnates at $1600ms$ delay for RNV conversations (Figure 5). The mismatch in double talk in conversations without delay might stem from errors in the end-of-turn prediction, where the model is too pessimistic in the prediction of the end of an utterance.

In general, the simulations seem to have less variance in almost all metrics (state probabilities, speaker alternation rate, interruption rates). One reason for that might be the limited amount of possible utterances that are available in the simulation, resulting in less variance.

Figure 6 shows the unintended interruption rate (UIR), i.e., the interruptions that are caused by delay and were not intended by the interrupting participant. While the increase in UIR is visible for empirical as well as simulated conversations, the number of unintended interruptions in the simulations is generally higher.

5 Conclusion

In this work, we modeled human turn-taking based on the distribution of turn-switches and -pauses. We applied this model in a conversation simulation. We evaluated how well the interactivity of real-world conversations with distinct levels of interactivity and different transmission delay can be modeled with this approach. The simulated conversations show the distinction between the interactivity of RNV and SCT scenarios as well as differences in speaker alternations and interruptions when introducing transmission delay. However, the influence of delay on turn-taking in the simulations seems to saturate with high delay levels. This might hint to a change in turn-taking behavior when large amounts of delay are present.

In future work, we plan to identify the changes in turn-taking behavior and model them based on events in the conversation (e.g., continued interruptions). We also plan to evaluate the proposed turn-taking model for the use in spoken dialogue systems.

Acknowledgments

This work was financially supported by the German Research Foundation DFG (grant number MO 1038/23-1).

References

Timo Baumann. 2008. Simulating Spoken Dialogue With A Focus on Realistic Turn-Taking. *13th ESSLLI Student Session*, pages 17–25.

Timo Baumann and David Schlangen. 2012. The INPROTK 2012 release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 29–32. Association for Computational Linguistic.

Sebastian Egger, Raimund Schatz, and Stefan Scherer. 2010. It Takes Two to Tango - Assessing the Impact of Delay on Conversational Interactivity on Perceived Speech Qualit. In *Eleventh Annual Conference of the International Speech Communication Association*, pages 1321–1324. ISCA.

Sebastian Egger, Raimund Schatz, Katrin Schoenenberg, Alexander Raake, and Gernot Kubin. 2012. Same but different? — Using speech signal features for comparing conversational VoIP quality studies. In *Communications (ICC), 2012 IEEE International Conference on*, pages 1320–1324. IEEE.

Florian Hammer. 2006. *Quality Aspects of Packet-Based Interactive Speech Communication*. Forschungszentrum Telekommunikation Wien.

Florian Hammer, Peter Reichl, and Alexander Raake. 2005. The well-tempered conversation: interactivity, delay and perceptual VoIP quality. In *IEEE International Conference on Communications*, volume 1, pages 244–249. Institute of Electrical and Electronics Engineers (IEEE).

ITU-T Recommandation G.107. 2011. *The E-model: a computational model for use in transmission planning*. International Telecommunication Union, Geneva.

ITU-T Recommendation P.59. 1993. *Artificial Conversational Speech*. International Telecommunication Union.

ITU-T Recommendation P.805. 2007. *Subjective Evaluation of Conversational Quality*. International Telecommunication Union, Geneva.

Nobuhiko Kitawaki and Kenzo Itoh. 1991. Pure delay effects on speech quality in telecommunications. *IEEE Journal on selected Areas in Communications*, 9(4):586–593.

H Lee and C Un. 1986. A study of on-off characteristics of conversational speech. *IEEE Transactions on Communications*, 34(6):630–637.

Chaoran Liu, Carlos Ishi, and Hiroshi Ishiguro. 2017. Turn-Taking Estimation Model Based on Joint Embedding of Lexical and Prosodic Contents. In *Proc. Interspeech 2017*, pages 1686–1690.

Rebecca Lunsford, Peter A Heeman, and Emma Rennie. 2016. Measuring Turn-Taking Offsets in Human-Human Dialogues. In *Proceedings of INTERSPEECH*, pages 2895–2899.

Thilo Michael and Sebastian Möller. 2019. Retico: An open-source framework for modeling real-time conversations in spoken dialogue systems. In *30th Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pages 238–245, Dresden. TUDpress.

Thilo Michael and Sebastian Möller. 2020. Simulating turn-taking in conversations with varying interactivity. In *31th Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pages 208–215, Dresden. TUDpress.

Alexander Raake, Katrin Schoenenberg, Janto Skowronek, and Sebastian Egger. 2013. Predicting speech quality based on interactivity and delay. In *Proceedings of INTERSPEECH*, pages 1384–1388.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue and Discourse*, 2(1):83–111.

Ethan O Selfridge and Peter A Heeman. 2012. A temporal simulator for developing turn-taking methods for spoken dialogue systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 113–117. Association for Computational Linguistics.

Gabriel Skantze. 2017. Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230.