SIMULATING TURN-TAKING IN CONVERSATIONS WITH VARYING INTERACTIVITY

Thilo Michael[1], *Sebastian Möller*[1,2]

[1]*Quality and Usability Lab, Technische Universität Berlin*
[2] *German Research Center for Artificial Intelligence (DFKI), Berlin, Germany*
thilo.michael@tu-berlin.de

Abstract: The impact of transmission delay on a telephone conversation depends not only on the severity of the delay, but also on the interactivity of the conversation. To model the perceived quality it is thus important to first model the interactivity of a conversation. In this paper we propose the method of simulating a conversation with different interactivity levels by simulating human turn-taking behavior. We simulate two types of conversations with different conversational interactivity with and without turn-taking. We then perform a parametric conversation analysis on the resulting conversations to show that adding a turn-taking mechanic to a simulation creates differences in interactivity that can be seen in human conversation.

1 Introduction

The experienced quality of a conversation over a telephone network is mainly formed by the parameters of the network and thus the different kinds of degradations that influence the call. However, research showed that the impact on the perceived quality of some degradations varies based on the type of conversations and its content [1]. 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 [2]. Furthermore, the severity of audible degradations like packet-loss can also depend on the interactivity, information density and available context. For example, if a lost packet renders a word unintelligible that bears critical information and can't be derived from context, it will have a greater impact on the quality perception than an unintelligible word that is not important or can be derived from context. Additionally, repairing the dialogue by resubmitting information requires further dialogue which in-turn increases the length of a conversation and leads to more possibilities for misunderstandings.

Transmission planning models like the narrowband E-model already include the interactivity-based changes in quality perception by changing the impact of delayed transmission based on *interactivity levels* of conversations [3, 2]. Based on Parametric Conversation Analysis (P-CA), where the on-off characteristics of speech are used to derive conversational metrics, this Conversational Interactivity (CI) can be determined [4, 5]. However, the P-CA and the determination of the CI requires recorded conversation and thus expensive conversation tests.

To overcome this problem, we propose a conversation simulation based on models from the field of spoken dialogue systems. Such a simulation could consist of two dialogue systems that exchange information via speech signals and models human turn-taking. In this paper we present such a system and evaluate how a simulation with and without turn-taking is able to model the conversational metrics of the P-CA. For this, we simulate two conversation scenarios, namely the Short Conversation Test (SCT) with a low CI and the Random Number Verification test (RNV), both standardized by the ITU [6]. We then compare the CI of the simulated

conversations with and without turn-taking with the interactivity of recorded human-to-human conversations to assess to which extend the simulation is able to model the interaction of each scenario.

Section 2 briefly reviews the fundamentals of conversational quality and the E-model as well as previous work in simulation of conversations on turn-taking level. Section 3 describes the setup of the conversation simulation and Section 4 describes the turn-taking model in detail. The resulting simulations are discussed in Section 5 and Section 6 concludes the discussion and suggests topics for future work.

2 Related Work

Subjective evaluation of telephone quality [7] and especially the conversation quality [6] has been a research focus, with recent research proposing to separate the analysis of the conversation into three phases: the *listening* phase, the *speaking* phase and the *interaction* phase [8] and analyzing the conversational quality in these different phases over multiple dimensions [9, 10].

Because of the interactive nature of conversations, common degradations like packet-loss not only degrade the perceived listening quality but also influence the conversational quality by altering the flow of the conversation [5]. In contrast to the degradations that influence the signal and thus the information that gets transmitted, delay is not affecting the characteristics of the signal, but the *timing* of it. The delayed arrival of turn-taking cues results in increased double talk and mutual silence. However, this varies not only with the amount of delay, but also with the interactiveness of the conversation [11, 1, 2, 12]. As a method to evaluate conversational quality, conversation tests with different conversational interactivity (CI) have been standardized. Two examples are the Random Number Verification test (RNV) [13] with a high conversational interactivity and the Short Conversation Test (SCT) [6] with a lower conversational interactivity. The RNV test consists of a list of 24 numbers in 4 blocks that the participants have to compare by alternatingly reading one block. The SCT provides scenarios such as ordering a pizza or booking a flight, where various kinds of information have to be exchanged.

Parametric Conversation Analysis (P-CA) is a framework that assesses the structure of conversations by parameters that can be instrumentally extracted from recordings of conversations [5]. It is based on the on the "on-off characteristics" of conversational speech that splits up the conversation into four *states*. States *A* and *B* represent part of the conversation where either only speaker *A* or only Speaker *B* talks. State *M* ("mutual silence") represents situations where neither conversation partner is talking and *D* ("double talk") corresponds to the state where both are talking at the same time [4, 14]. From these states, metrics like speaker alternation rate, double talk rate and interruption rate can be derived as well as overlaps and gaps between speaker changes measured [1, 15]. The interactivity of conversations as measured by P-CA has been included into the *delay sensitivity* factor of the Narrowband E-model [3]. The Wideband E-model [16] and Fullband E-Model [17] however do not take conversational interactivity into account.

While simulating dialogues was previously not used for prediction of conversational quality, it has been long used as a way to generate dialogues for spoken dialogue systems [18]. These simulations may be used to train dialogue managers or to evaluate the human-computer-interaction automatically [19, 20]. Most recent approaches to user simulation use statistical approaches. However, those rely on large sets of training data when complex behavior has to be modeled [21]. To overcome those shortcomings, different methods have been described to progressively generate data that can be used to train statistical dialogue manager. In [22] Schatzmann et al. propose a probabilistic, agenda based method for training statistical dialogue manager. In this method, the user is modeled to have an agenda and a goal. The agenda is

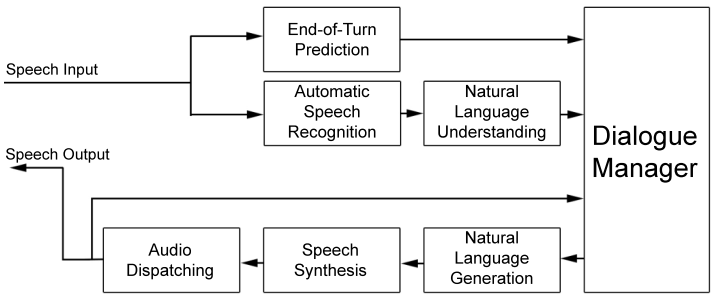


Figure 1 – An incremental spoken dialogue network containing parts for speech understanding and end-of-turn prediction on the top, the and dialogue managing unit on the right and the speech generation and audio dispatching on the bottom.

described as a stack-like data structure that contains the next relevant dialogue acts. The goal is split into constraints and requests that model the information that the user requires and provides. With this agenda-based structure, goal oriented, semi-structured conversations can be modeled.

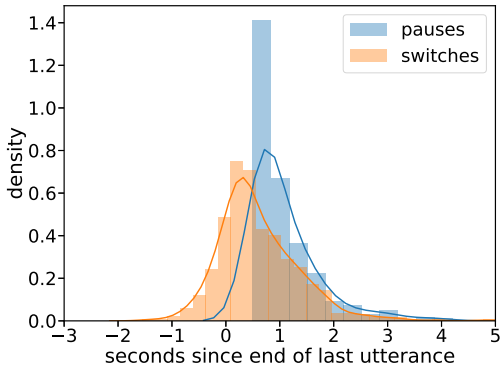
A simulated spoken dialogue, especially with a focus on turn-taking, needs to be incremental. This allows for a timely processing of small information units which is necessary for the precise timing of turn-taking. While parts of spoken dialogue systems like automatic speech recognition [23] and natural language processing [24] have been designed incrementally, the incremental processing in every part of a spoken dialogue system has become the focus of research as of recent [25, 26]. Most notably, in [27], Skantze and Schlangen described a general and abstract model of incremental processing in spoken dialogue systems. In this model, incremental modules consume, process and produce small information bits called incremental units [27]. Most of these concepts were implemented in the incremental processing toolkit InProTK [26], a framework that allows for the modeling and implementation of incremental spoken dialogue systems and Retico [28], a framework for modeling human-to-human conversations in a real-time environment.

While turn-taking behavior is a long studied phenomenon [29], recent work has investigated the human turn-taking behavior in conversations [15], end-of-turn prediction [30, 31, 32] and rules for modeling turn-taking behavior [14, 33, 34]. Simulation of human-to-human dialogue has been part of that effort of modeling turn-taking behavior. For example, in [33], Baumann describes a dialogue simulation with simple rules to enable turn-taking. These simulations however only operate on the signal level and the utterances exchanged are generated, speech-like sound [14] or randomly selected utterances [33, 34].

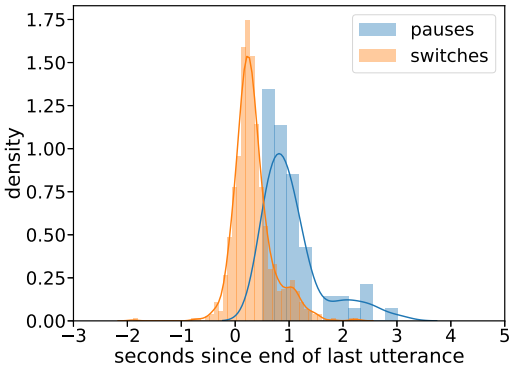
3 Simulation Setup

The simulation is based on a set of conversation test carried out with untrained participants. In this test, participants carried out conversations after the standardized SCT and RNV scenarios. For the simulation one scenario was selected from each conversation type and 20 SCT conversations and 20 RNV conversations were annotated with dialogue acts, transcripts and turn-taking information. 20 different conversations from each conversation type were used to evaluate the simulation.

For implementing the simulation, we made use of the Retico framework [28], which uses incremental processing for a timely transmission of hypotheses in the dialogue pipeline. The setup of one agent in the simulation is shown in Figure 1. The dialogue manager uses an agenda-based model to determine the next dialogue act. Natural language generation and speech synthesis is modeled with the snippets of speech from the annotated training data. However, also real synthesis can be used by the simulation. An audio dispatching module handles the dis-



(a) Turn-taking in SCT conversations



(b) Turn-taking in RNV conversations

Figure 2 – Timing of pauses and switches relative to the end of the last utterance. Negative values represent overlaps during turn-taking.

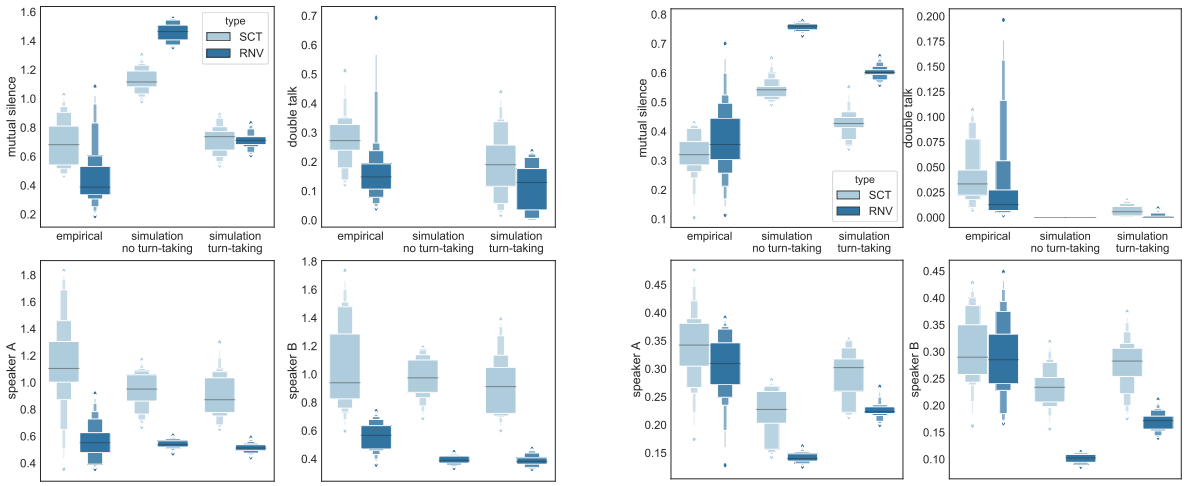
patching of the speech data to the interlocutor and reports back to the dialogue manager how much of the current speech act is completed. The speech output from the other agent in the simulation is routed through an automatic speech recognition and natural language understanding also trained on the speech data of the training conversations. Also, an end-of-turn module predicts the time until the end of the interlocutor's current utterance.

The simulation setup is able to use various different state-of-the-art systems for speech synthesis, speech recognition, natural language understanding and end-of-turn prediction. However, for the simulations evaluated in this paper, models based on meta-information from the training material is used instead, so that no errors are introduced in these processing steps. During the simulations the transmitted speech (one channel per agent), the transcript (as generate by the NLG) and the according dialogue acts are persisted onto disk. The simulation consists of two of the dialogue systems shown in Figure 1 that are connected to each other - each with their own agenda and set of utterances. Throughout the conversation, the dialogue manager receives the latest dialogue act hypotheses, predictions about the state and progress of the interlocutor's as well as it's own speech.

4 Turn-Taking Model

The turn-taking of the simulated agents is modeled by probability distributions that based on the work by Lunsford et al. [15]. For this, we measured the offsets of utterances in response to turn-keeping and turn-giving utterances of the interlocutor. These offsets in seconds are relative to the ending of the last utterance. A negative value denotes a speaker change with double talk and a positive value denotes a speaker change with mutual silence or alternatively that no speaker change occurred (turn-keeping).

This analysis results in the distributions for pauses and switches as shown in Figure 2 (a) for SCT conversations and Figure 2 (b) for RNV conversations. The turn-taking in each agent is then determined by one of the following four rules: (1) If the agent speaks and the interlocutor does not speak, the agent keeps talking until the end of the utterance. (2) If only the interlocutor is speaking, the agent randomly samples the *switches* distribution and uses the predictions of the end-of-turn module to determine when to speak. (3) If the agent is no longer talking but spoke last, it randomly samples the *pauses* distribution to determine when to continue speaking. (4) If both the agent and it's interlocutor are speaking and one of them is not at the beginning or the end of an utterance, it stops talking. Each agent decides based on the current dialogue act if it samples from the distributions of SCT or RNV conversations. Dialogue act corresponding to



(a) Sojourn times in seconds

(b) State probabilities

Figure 3 – Sojourn times in seconds (a) and probabilities (b) of the states mutual silence, double talk, speaker A and speaker B for RNV and SCT scenarios in empirical conversations, simulations without turn-taking and simulation with turn-taking.

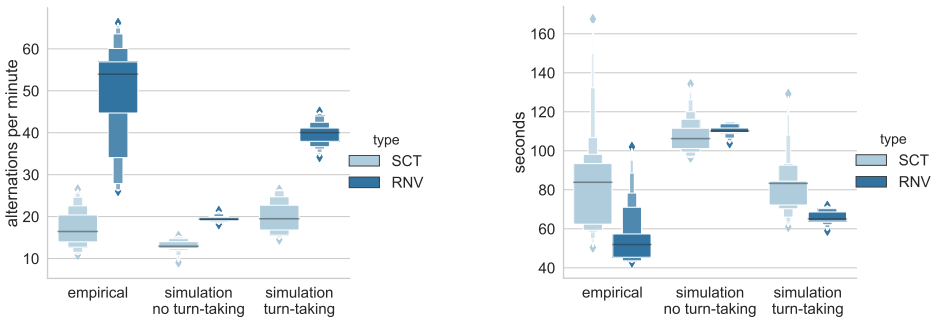
quick exchanges of information are sample from the RNV distribution and all other dialogue acts are sampled from the SCT distribution.

5 Results and Discussion

We simulated 100 RNV and 100 SCT conversations turn steps (with a pause of one second between turns) and with our turn-taking model. Figure 3 shows the sojourn times and the state probabilities of mutual silence, double talk, speaker A and speaker B. In general, the empirical conversations result in a larger variance than the simulations. The sojourn times for speaker A and B are very similar between all conversations, which is due to the fact that the simulations use the same speech segments. For simulations without turn-taking, the sojourn times of mutual silence is more than a second and for double talk it is zero, which is to be expected. The simulation with turn-taking matches the sojourn times for mutual silence in the SCT scenario. However, for the more interactive RNV scenario however, the sojourn times are higher than in the empirical data.

The state probabilities for mutual silence and double talk are too high and too low respectively when comparing the turn-taking simulations to the empirical data. This may be due to the pessimistic predictions of the end-of-turn module (predicting the interlocutor is still speaking while the utterance was already over) which leads to more gaps and less overlaps when turns are being taken. This would also explain the increase in mutual silence probability of the RNV scenario, considering that it consists of more turns than the SCT conversations.

Figure 4 (a) shows the speaker alternation rate (SAR) of the empirical and simulated conversations, that is how many turn are occurring per minute. The SAR of the simulation without turn-taking already differs between SCT and RNV. This is due to the fact that the utterances are shorter in the RNV scenario. However, the simulation with turn-taking shows that difference more pronounced. The conversation length in seconds for the simulation in turn-steps as shown in Figure 4 (b) is too high and does not differ between SCT and RNV. Adding turn-taking brings the conversation length of SCT scenarios to the same level as empirical conversations. For RNV conversation however it is still too hight, which again could be due to the additional mutual silence introduced in the turn-taking.



(a) Speaker alternation rate **(b)** Conversation length

Figure 4 – Speaker alternation rate (a) and conversation length (b) for RNV and SCT scenarios in empirical conversations, simulations without turn-taking and simulation with turn-taking.

6 Conclusion

In this paper we showed that conversations with different levels of interactivity can be simulated using user-simulation methods from the field of spoken dialogue systems. The simulations of the two scenarios in turn-steps already show differences in the P-CA due to the fact that RNV and SCT conversations have a different structure (i.e. number of turns and length of utterances). When adding turn-taking, a clear distinction in SAR can be seen between the simulated SCT and RNV conversations. Especially the state probabilities of the simulations are shifted towards mutual silence, which might be due to pessimistic end-of-turn prediction.

In future work we are planning to evaluate the simulation approach with different scenarios without the use of annotated data and to build a model that predicts the delay sensitivity from those simulations. We also plan to modify the turn-taking mechanism to account for timing problems with the end-of-turn predictions and to shifting behavior over the course of a conversation. Finally, we plan to insert delay into the simulations to analyze if and how this approach is able to model the according changes in turn-taking behavior.

Acknowledgements

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

References

[1] HAMMER, F., P. REICHL, and A. RAAKE: *The well-tempered conversation: interactivity, delay and perceptual VoIP quality*. In *IEEE International Conference on Communications*, vol. 1, pp. 244–249. Institute of Electrical and Electronics Engineers (IEEE), 2005. doi:10.1109/ICC.2005.1494355. URL `http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=1494355`.

[2] RAAKE, A., K. SCHOENENBERG, J. SKOWRONEK, and S. EGGER: *Predicting speech quality based on interactivity and delay*. In *Proceedings of INTERSPEECH*, pp. 1384–1388. 2013.

[3] ITU-T RECOMMANDATION G.107: *The E-model: a computational model for use in transmission planning*. International Telecommunication Union, Geneva, 2011. URL `http://handle.itu.int/11.1002/1000/12505`.

[4] LEE, H. and C. UN: *A study of on-off characteristics of conversational speech*. *IEEE Transactions on Communications*, 34(6), pp. 630–637, 1986.

[5] HAMMER, F.: *Quality Aspects of Packet-Based Interactive Speech Communication*. Forschungszentrum Telekommunikation Wien, 2006.

[6] ITU-T RECOMMENDATION P.805: *Subjective Evaluation of Conversational Quality*. International Telecommunication Union, Geneva, 2007.

[7] ITU-T RECOMMENDATION P.800: *Methods for subjective determination of transmission quality*. International Telecommunication Union, Geneva, Switzerland, 1996.

[8] GUÉGUIN, M., R. LE BOUQUIN-JEANNÈS, V. GAUTIER-TURBIN, G. FAUCON, and V. BARRIAC: *On the evaluation of the conversational speech quality in telecommunications*. *EURASIP Journal on Advances in Signal Processing*, 2008, p. 93, 2008.

[9] KÖSTER, F.: *Multidimensional Analysis of Conversational Telephone Speech*. Springer, 2017.

[10] MÖLLER, S., F. KÖSTER, and B. WEISS: *Modelling speech service quality: From conversational phases to communication quality and service quality*. In *Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*, pp. 1–3. IEEE, 2017.

[11] EGGER, S., R. SCHATZ, and S. SCHERER: *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*, pp. 1321–1324. ISCA, 2010.

[12] EGGER, S., R. SCHATZ, K. SCHOENENBERG, A. RAAKE, and G. KUBIN: *Same but different? — Using speech signal features for comparing conversational VoIP quality studies*. In *Communications (ICC), 2012 IEEE International Conference on*, pp. 1320–1324. IEEE, 2012.

[13] KITAWAKI, N. and K. ITOH: *Pure delay effects on speech quality in telecommunications*. *IEEE Journal on selected Areas in Communications*, 9(4), pp. 586–593, 1991.

[14] ITU-T RECOMMENDATION P.59: *Artificial Conversational Speech*. International Telecommunication Union, 1993.

[15] LUNSFORD, R., P. A. HEEMAN, and E. RENNIE: *Measuring Turn-Taking Offsets in Human-Human Dialogues*. In *Proceedings of INTERSPEECH*, pp. 2895–2899. 2016.

[16] ITU-T RECOMMANDATION G.107.1: *Wideband E-model*. International Telecommunication Union, Geneva, 2015.

[17] ITU-T RECOMMANDATION G.107.2: *Fullband E-model*. International Telecommunication Union, Geneva, 2019.

[18] ECKERT, W., E. LEVIN, and R. PIERACCINI: *User modeling for spoken dialogue system evaluation*. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pp. 80–87. IEEE, 1997.

[19] HILLMANN, S.: *Simulation-Based Usability Evaluation of Spoken and Multimodal Dialogue Systems*. Springer, 2017.

[20] PIETQUIN, O. and H. HASTIE: *A survey on metrics for the evaluation of user simulations*. *The knowledge engineering review*, 28(1), pp. 59–73, 2013. doi:doi:10.1017/S0269888912000343.

[21] SCHATZMANN, J., K. WEILHAMMER, M. STUTTLE, and S. YOUNG: *A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies*. *The knowledge engineering review*, 21(2), pp. 97–126, 2006.

[22] SCHATZMANN, J., B. THOMSON, K. WEILHAMMER, H. YE, and S. YOUNG: *Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System*. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 149–152. Association for Computational Linguistics, 2007.

[23] YOUNG, S. J., N. RUSSELL, and J. THORNTON: *Token passing: a simple conceptual model for connected speech recognition systems*. Cambridge University Engineering Department Cambridge, UK, 1989.

[24] KILGER, A. and W. FINKLER: *Incremental generation for real-time applications*. Tech. Rep., DFKI, Saarbrücken, Germany, 1995.

[25] AIST, G., J. ALLEN, E. CAMPANA, and C. G. GALLO: *Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods*. *Decalog 2007*, p. 149, 2007.

[26] BAUMANN, T. and D. SCHLANGEN: *The INPROTK 2012 release*. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pp. 29–32. Association for Computational Linguistic, 2012.

[27] SCHLANGEN, D. and G. SKANTZE: *A General, Abstract Model of Incremental Dialogue Processing*. *Dialogue and Discourse*, 2(1), pp. 83–111, 2011.

[28] MICHAEL, T. and M. SEBASTIAN: *Retico: An open-source framework for modeling real-time conversations in spoken dialogue systems*. In *30th Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pp. 238–245. TUDpress, Dresden, 2019.

[29] SACKS, H., E. SCHEGLOFF, and G. JEFFERSON: *A Simplest Systematics for the Organization of Turn-Taking for Conversation*. *Language*, 50(4), pp. 696–735, 1974. doi:10.2307/412243.

[30] LIU, C., C. ISHI, and H. ISHIGURO: *Turn-Taking Estimation Model Based on Joint Embedding of Lexical and Prosodic Contents*. In *Proc. Interspeech 2017*, pp. 1686–1690. 2017.

[31] MAIER, A., J. HOUGH, and D. SCHLANGEN: *Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems*. In *Proceedings of INTERSPEECH 2017*. 2017.

[32] SKANTZE, G.: *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*, pp. 220–230. 2017.

[33] BAUMANN, T.: *Simulating Spoken Dialogue With A Focus on Realistic Turn-Taking*. *13th ESSLLI Student Session*, pp. 17–25, 2008.

[34] SELFRIDGE, E. O. and P. A. HEEMAN: *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*, pp. 113–117. Association for Computational Linguistics, 2012.