Simulation of Conversational Behavior during Impaired Audio Transmissions

Thilo Michael

Technische Universität Berlin

thilo.michael@tu-berlin.de

1. Motivation

Measuring and predicting the quality of speech has been a research topic for a long time [1]. Subjective measurements like listening or conversation tests are used to quantify the perceived quality from the view of the users [2], while instrumental methods like the E-Model try to predict the results of those tests based on signals or parameters [3].

Recent research about influencing factors shifted towards transmission delay and packet-loss, as they are most prevalent in packet-base network switching. It has been shown that the effects of delay cannot be modeled on parameters of the transmission alone and are influence by a multitude of factors (e.g. the interactivity of the conversation) [4, 5, 6, 7]. Also with packet-loss it has been shown that the perceived quality varies depending on how much information was transmitted during the degraded part of the speech[8]. Current instrumental models either do not take the type of conversations or the pragmatics into account or try to define metrics to describe the influence of delay on the conversational behavior. In the research area of computer linguistics, a simulation of conversations (especially between humans and dialogue systems) is often used to train dialogue managers (e.g [9]). This is often done on a pragmatic level, where the dialogue system and the simulated interlocutor exchange dialogue acts. Recent research is also focusing on the prediction of end-of-turns to facilitate realistic turn-taking and making the dialogue systems more robust[10, 11, 12]. However, as far as we know this research was never used for the prediction of conversational quality.

The goal of this research is to transfer the research on simulation of human-computer conversation to the prediction of conversational quality. This is done in two concrete steps: First, a human-human voice-only conversation simulation that reacts to delayed transmissions and misunderstandings due to packet-loss is being created. Utilizing this simulation framework, a conversational quality estimator is being built that uses the information gained from multiple simulated conversations to give a mean opinion score estimation. Based on this goal, five research questions were formulated:

1. **Simulation of conversations**: How can the simulation methods of studying user behavior with spoken dialogue systems be transferred to telephone conversations between two humans?
2. **Simulation of turn-taking**: How can turn-taking in a telephone conversation be modeled and simulated on a pragmatic level?
3. **Turn-taking during delayed transmission**: How can the turn-taking models be applied to conversations under the influence of transmission delay?
4. **Intelligibility**: How can speech intelligibility of a specific utterance be predicted based on parameters of the transmission system?
5. **Evaluation**: How well can conversation parameters and the overall conversational quality be measured with this approach?

2. Related Work

Subjective evaluation of telephone quality [1] and especially the conversation quality [2] has been a research focus. Recent work has been analyzing the conversational quality in its different phases over multiple dimensions [13]. 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 [8]. 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[14].

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) [15] with a high conversational interactivity and the Short Conversation Test (SCT) [2] 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.

While turn-taking behavior is a long studied phenomenon [16], recent work has investigated the human turn-taking behavior in telephone conversations [17], end-of-turn prediction [10, 11, 12] and rules for modeling turn-taking behavior [18, 19, 20]. Simulation of human-to-human dialogue has been part of that effort of modeling turn-taking behavior. For example, in [19], Baumann describes a dialogue simulation with simple rules to enable turn-taking. These simulations only operate on the signal level and the transmitted audio is either artificial voice [21] or randomly selected utterances [19, 20].

3. Incremental Simulation Framework

For a human-to-human conversation simulation two virtual agents are needed that converse with each other. To model misunderstandings due to packet-loss as well as the timing-influences of transmission delay, the design of the agents needs to fulfill two main features:

1. in order to model misunderstandings, the agents need to exchange meaningful dialogue acts in a goal-oriented conversation scenario and need to identify how well each semantically important concept was understood.
2. In order to model the influences of transmission delay, the agents need to dynamically and timely negotiate the taking of turns in the conversation.

Both of these features require an incremental processing of the information (i.e. speech signal, transcription and dialogue acts) inside the simulation.

The framework for incremental processing created for this simulation is based on the conceptual model described in [22]

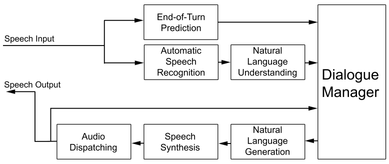

Figure 1: *An incremental spoken dialogue network containing parts for speech understanding and end-of-turn prediction on the top, the dialogue managing unit on the right and the speech generation and audio dispatching on the bottom.*

and is published in [23]. While the structure of the incremental network of the simulation as seen in Figure 1 is similar to other incremental spoken dialogue systems (e.g. [19]), it adds the capability to run multiple networks in succession and save the resulting dialogues as audio- and transcription-files. Combining two instances of the network shown in Figure 1 provides an environment where the two dialogue managers can interact with each other based on the dialogue acts and the turn-taking signals provided by the respective modules.

4. Incremental Dialogue Manager

The simulation is built to model short conversation tests that try to mimic ordinary telephone conversation (e.g. ordering a pizza or booking a flight) and random number verification tests where interlocutors try to exchange a list of numbers in the shortest amount of time possible [2]. For this we used 20 recorded conversations of these types as the baseline data. Because these conversational tests are structured and goal oriented, we decided to implement an agenda-based model that is based on [24].

Dialogue manager and user simulations that have previously been described in literature usually produce a dialogue act and concepts according to a given state of the dialogue and a dialogue act produced by the interlocutor [24, 25, 26], which results in a dialogue that progresses in *turn steps*.

We modified the agenda-based dialogue management model described in [24] to be able to pop dialogue acts from the stack independent of whether the interlocutor has taken over the turn. This is done by updating the stack with the most recent item of the agenda regardless of whether the interlocutor acknowledged the current dialogue acts of the agent. This way the agents may keep the turn by repeating a question or shift the topic of the conversation after they have transmitted information.

5. Turn-Taking Model

The turn-taking model in our simulation is controlling whether the incremental dialogue manager should produce an output for the natural language generation module, the speech synthesis module and the audio dispatching module (see Figure 1). For the decision of when to take or give a turn, it relies on two parameters: the input of the end-of-turn prediction module that tries to predict when the turn of the interlocutor ends (i.e. when it would be a good time to take over the turn) and the input of the audio dispatching module that provides information whether and how long the agent itself is currently speaking.

The turn-taking decision of the agents are based on the work by Lunsford et al. [17]. 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).

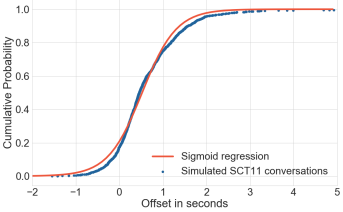

Figure 2: *A comparison of the sigmoid-function the turn-taking mechanism is modeled after (red) and the actual turn-taking points in the simulated SCT conversations (blue).*

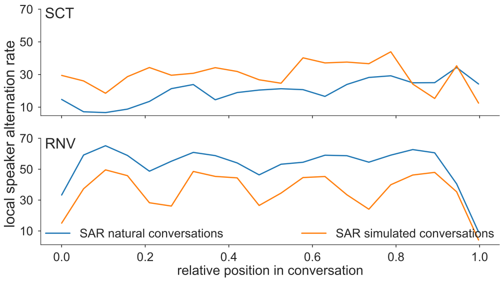

Figure 3: *The local speaker alternation rate (SAR) over the course of a SCT (top) and RNV test (bottom) averaged over all natural and simulated conversations.*

We used the conversations of the training data to determine the cumulative probability of when a turn is being given to the interlocutor and when a turn is kept (these matched very closely to the findings in [17]). To create a model that determines when an agent should start to speak, we approximated the cumulative probability with a sigmoid function and inverted it, so that the function maps the probability input to the offsets in seconds (see Figure 2). Figure 3 shows the speaker alternation rate (speaker alternations per minute) for the empirical data and simulations over 20 short conversation tests and random number verification tests. It can be seen that the simulation adapts to the different alternation-styles of the two conversation scenarios.

6. Roadmap

The next step for the simulation is to verify the turn-taking model by simulating conversations with different amounts of transmission delay added and comparing speaker alternation rate and conversation states (i.e. mutual silence, double talk and speaker states) to recorded data of short conversation tests and random number verification tests. With this, the first conversational quality model can be made that use parameters of the simulated conversations to predict the subjective quality and the parameters of the conversation.

Following that, a speech intelligibility model has to be created that decides whether an agent needs to request the transmitted information once more. With that, a conversational quality estimator factoring in delay and packet-loss can be created.

7. Acknowledgements

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

8. References

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

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

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

[4] S. Egger, 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*. ISCA, 2010, pp. 1321–1324.

[5] F. Hammer, P. Reichl, and A. Raake, "The well-tempered conversation: interactivity, delay and perceptual VoIP quality," in *IEEE International Conference on Communications*, vol. 1. Institute of Electrical and Electronics Engineers (IEEE), 2005, pp. 244–249. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1494355

[6] A. Raake, K. Schoenenberg, J. Skowronek, and S. Egger, "Predicting speech quality based on interactivity and delay," in *Proceedings of INTERSPEECH*, 2013, pp. 1384–1388.

[7] S. Egger, 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*. IEEE, 2012, pp. 1320–1324.

[8] A. Raake, "Short-and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1957–1968, 2006.

[9] W. Eckert, E. Levin, and R. Pieraccini, "User modeling for spoken dialogue system evaluation," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 80–87.

[10] C. Liu, C. Ishi, and H. Ishiguro, "Turn-Taking Estimation Model Based on Joint Embedding of Lexical and Prosodic Contents," in *Proc. Interspeech 2017*, 2017, pp. 1686–1690.

[11] A. Maier, J. Hough, and D. Schlangen, "Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems," in *Proceedings of INTERSPEECH 2017*, 2017.

[12] G. Skantze, "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*, 2017, pp. 220–230.

[13] F. Köster, *Multidimensional Analysis of Conversational Telephone Speech*. Springer, 2017.

[14] F. Hammer, *Quality Aspects of Packet-Based Interactive Speech Communication*. Forschungszentrum Telekommunikation Wien, 2006.

[15] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE Journal on selected Areas in Communications*, vol. 9, no. 4, pp. 586–593, 1991.

[16] H. Sacks, E. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[17] R. Lunsford, P. A. Heeman, and E. Rennie, "Measuring Turn-Taking Offsets in Human-Human Dialogues," in *Proceedings of INTERSPEECH*, 2016, pp. 2895–2899.

[18] ITU-T Recommendation P.59, *Artificial Conversational Speech*. International Telecommunication Union, 1993.

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

[20] E. O. Selfridge 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*. Association for Computational Linguistics, 2012, pp. 113–117.

[21] ITU-T Recommendation P.50, *Artificial Voices*. International Telecommunication Union, 1999.

[22] D. Schlangen and G. Skantze, "A General, Abstract Model of Incremental Dialogue Processing," *Dialogue and Discourse*, vol. 2, no. 1, pp. 83–111, 2011.

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

[24] J. Schatzmann, 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*. Association for Computational Linguistics, 2007, pp. 149–152.

[25] J. Schatzmann and S. Young, "The hidden agenda user simulation model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 4, no. 17, pp. 733–747, 2009.

[26] L. El Asri, J. He, and K. Suleman, "A sequence-to-sequence model for user simulation in spoken dialogue systems," in *Proceedings of Interspeech*, 2016, pp. 1151–1155.