Improving a Robot's Turn-Taking Behavior in Dynamic Multiparty Interactions

Maike Paetzel-Prüsmann*
maike.paetzelprusmann@disney.com
Disney Research
Zurich, Switzerland

James Kennedy*
james.kennedy@disney.com
Disney Research
Glendale, California, USA

ABSTRACT

In this paper, we describe ongoing work to develop a robust and natural turn-taking behavior for a social agent to engage a dynamically changing group in a conversation. We specifically focus on discussing likely interaction scenarios for a social robot and how appropriate conversational behavior could unfold in each situation. Preliminary findings from annotations of more than 9,000 dialogue samples from a related domain are used to help judge the importance of different interaction scenarios. We conclude by outlining important general considerations for designing more robust dialogue systems as well as highlight next steps we are taking in developing our character's turn-taking behavior.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**; *Collaborative interaction*; • **Computing methodologies → Discourse, dialogue and pragmatics**; • **Computer systems organization →** *Robotic autonomy*.

KEYWORDS

turn-taking, multi-party interactions, human-robot dialogue

ACM Reference Format:
Maike Paetzel-Prüsmann and James Kennedy. 2023. Improving a Robot's Turn-Taking Behavior in Dynamic Multiparty Interactions. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580117

1 INTRODUCTION

For a conversation to be engaging, it is not only important to have something interesting to say - you also need to be able to express it in an appropriate and appealing tone and at the right time [Maat et al.(2010)]. Even as humans, we often struggle with these key aspects: When meeting a stranger, we may find it hard to choose a conversation topic, talk too fast, or unintentionally interrupt them. When more people partake in a conversation, these problems become increasingly challenging as they may have different topical interests, have opposing conversational goals, or compete for attention within the group [Traum(2004)]. Artificial characters often still lack even fundamental aspects of appropriate conversational behavior in one-on-one and multi-party settings. Hence, many designers of dialogue systems default to either rule-based solutions that restrict the naturalness of the interaction (see review in [Skantze(2021)]), or leverage a human wizard remote-controlling the robot's conversational dynamics (e.g., [DeVault et al.(2015)]).

In the ongoing work presented in this paper, we aim to provide the community with *data-driven guidelines for modeling turn-taking in multi-party settings.* Moreover, we are discussing our ongoing development of a *robust autonomous turn-taking behavior for a robotic character to engage a dynamically changing group of people in a conversation.* At each point in the interaction, the robot is considering its own conversational goals and their urgency, knowledge about norms in spoken conversations, and its beliefs about the other peoples' conversational goals and urgency. Based on these, the system should decide to wait for one of the human dialogue partners to start or continue speaking, to use a silence in the conversation to take the floor, to interrupt a human interlocutor if the urgency of its own content is deemed high enough, or to abandon its own speech if the urgency or content of the human is given priority.

To develop models supporting this autonomous behavior, we leverage data collected with another character in a similar interactive setting but an unrelated fictional world. Our aim is to transfer the models between these characters with as little additional training data from the new character and domain as possible. From the previously collected corpus, we extracted several scenarios that are important for a robot to consider in an interaction. While

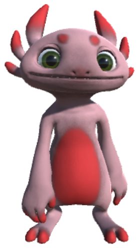


Figure 1: Our conversational agent in the current virtual embodiment. The character is designed to have human-like features, but be distinctly non-human, to provide flexibility in behavior design and expectations.

[*] Both authors contributed equally to this research.

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03...$15.00
https://doi.org/10.1145/3568294.3580117

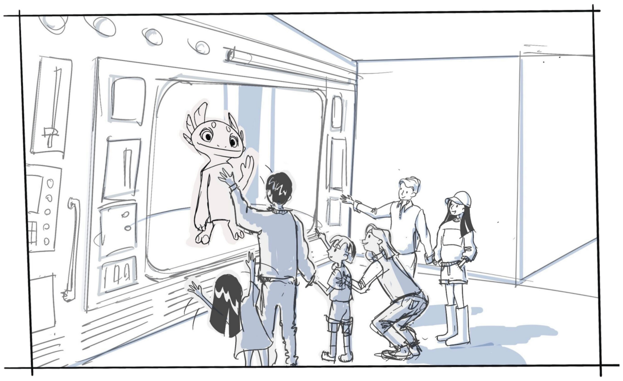


Figure 2: The setup of the multi-party interaction with our virtual character.

some of these are common in one-on-one dialogue (e.g., predicting end-of-turn signals), some are specific to multi-party interactions (e.g., identifying the addressee of a turn-yielding cue). Moreover, we found some scenarios like meta-conversations about the robot's behavior to be unique modes of interaction uncommon in human-human dialogue.

This paper first introduces our new artificial character (Fig. 1) which is the key AI agent in the multi-party interaction we are developing. We then discuss several conversational scenarios derived from the literature and a previous corpus collection that we consider important for a character to navigate when engaging in conversations. We support and prioritize our scenarios by annotating and analyzing more than 9,000 lines of human speech. Based on these preliminary findings, we summarize key considerations for the human-robot interaction community when developing dialogue systems for autonomous agents and highlight ongoing work in our effort to design more natural and robust models for our character.

2 THE ROBOT & CONVERSATIONAL SETUP

We developed a robotic character (Fig. 1) and use its virtual embodiment on a screen for this project. It is placed in a public space in which the immediate surroundings that it can refer to and interact with consists of virtual and physical objects alike. To make people more comfortable and increase privacy [Rueben(2018)], the dialogue model only gets access to the auditory input channel. Unlike in most of the related work, visual features like gaze and gestures are hence not available to our model [Mutlu et al.(2012), Skantze(2021), Żarkowski(2019)].

People engaging in a conversation with our character are either individuals or small groups of people, who may walk up and interact with the robot at any point. If people approach the robot in a group, they are usually acquainted with each other, which likely increases the level of meta-comments made about the robot and the setup. Events in the real world (people joining or leaving) as well as changes of the virtual environment can influence the urgency of conversational topics both for the human as well as the artificial interlocutor. Each interaction with our robot is designed to last about five minutes. However, parties can decide to leave the conversation at any point. Similarly, individuals or other groups may join the conversation, either as active participants or as bystanders, at any time. This provides a challenging conversational dynamic uncommon in the related work.

3 DESIRED ROBOT CAPABILITIES

By observing several multi-party interactions with previous characters we developed and by aligning it with related work in human-human and human-agent interaction [Bohus and Horvitz(2011), Sacks et al.(1978), Skantze(2021), Traum(2004)], we identified three areas pivotal to a character's natural turn-taking ability.

3.1 Taking the floor

The first set of capabilities evolves around judging when it is appropriate for a robot to take the floor given that the floor was previously held by a human interlocutor. This requires an understanding of (i) whether the same human is likely going to continue talking (identifying *turn-holding cues*) and (ii) whether the turn was handed to an interlocutor who is not the robot itself (*turn-yielding cues to a third party*). To understand the importance of identifying turn-holding cues, consider the following example:

PERSON: how about [*pause*]
PERSON: the red ball?

In this case, the voice activity detection of the dialogue system noted a pause after the first two words. In traditional dialogue systems, pause thresholds are often the only cue that is used to interpret whether the robot can take the turn [Skantze(2021)]. However, it is likely that the same person will continue the sentence and only interpreting the first part of the input would not lead to a satisfying response. Unless the robot's conversational goal is urgent or of high importance, an interruption at this point would not be considered appropriate dialogue behavior. Hence, the robot should learn to leave the floor to the person currently speaking.

Similarly, it would not be appropriate for the robot to take the floor if one interaction partner has handed the floor to another human interlocutor:

PERSON 1: how about [*pause*]
PERSON 1: I don't know help me out here [*pause*]
PERSON 2: the red ball?

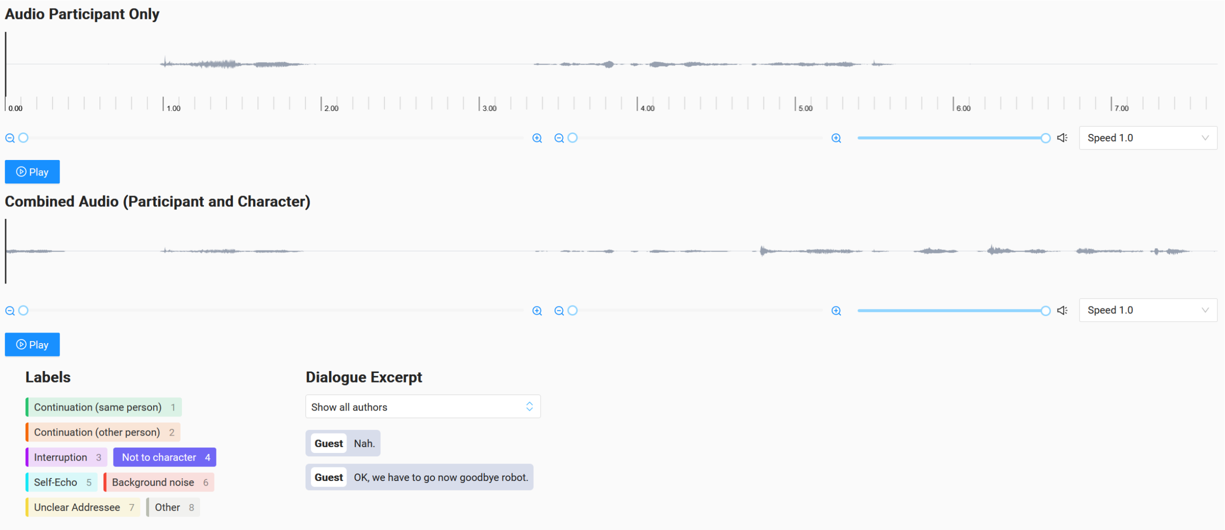
Note that depending on the context that preceded this dialogue, "help me out here" could also be an invitation for the robot to take the turn. However, if this is directed to another of the human interaction partners, the robot should not take the turn at this point.

3.2 Holding or handing over the turn after intentional interruptions

In our second set of scenarios, the robot is holding the turn and a human interaction partner is recognized to start speaking while the robot speech hasn't finished. In this case, it is important to detect whether the human comment is directed to the robot.

ROBOT: well, last time I checked, – [*interruption detected*]
PERSON: Sorry but we really gotta go

In this example, the speech is directed towards the robot and the information was urgent to share for the human interaction partner. In this case, it would be appropriate for the robot to abandon the speech act as soon as it understands that the human needs to stop



Guest Nah.

Guest OK, we have to go now goodbye robot.

Figure 3: The annotation setup in label-studio with the two audio samples on top, the eight labels as buttons to the left and the raw ASR transcript to the right.

the conversation and either share an important last conversational point, or say goodbye to the party [Chao and Thomaz(2012)]. In other situations, however, the robot may decide that the content is of lesser importance and decide to continue holding the turn regardless of the interruption.

We also observe several examples in which the interruptions are directed to another person participating in the conversation.

ROBOT: oh, I like the color red, too, but – [*interruption detected*]

PERSON: they know I said the word red!

In this case, the comment is neither marking that the person wants to take the floor in the conversation nor is it meant for the robot to interpret the speech as a topical reply. While we find remarks to be common in human-human dialogue as well, they are usually not concerned with the capabilities of one of the interaction partners. The dialogue system may still, depending on the robot's personality, decide to respond to the person's comment. However, it should then move on with the main topic of the conversation according to the robot's conversational goals.

There are times in which the addressee of a sentence cannot be determined with certainty. This is especially true if no visual channel can be utilized to understand the direction of the speaker's gaze. In this case, the dialogue model may, depending on the content, decide to ask for clarification or keep the floor and continue talking if the content of the comment is considered of low importance.

3.3 Keeping the floor after unintentional interruptions

We identified two types of scenarios in which a robot may record incoming speech acts that should not lead to any change in the robot's dialogue behavior. First, the robot may occasionally hear a *self-echo*, a playback of its own speech. This can be considered a technical artifact that is not observable by the human interaction partners and should hence be disregarded by the robot.

The second set involves background noises, which are often tricky to detect as they can lead to peculiar transcriptions from the Automatic Speech Recognition (ASR) system:

ROBOT: oh, I like the color red, too, but – [*interruption detected*]

PERSON: box

In this case, the robot may choose to ignore background noises the first times they occur, but may comment on the noisy environment if they become too frequent. In any case, it should not shift the robot's permanent conversational goals.

4 DATASET & ANNOTATIONS

To appropriately react to the aforementioned scenarios, there are several capabilities necessary for a dialogue system. It needs to (i) predict if the same human wants to continue speaking, (ii) detect turn-yielding cues, (iii) determine the addressee of a line, (iv) judge the urgency of conversational content, and (v) filter self-echo and background noise. To train models with these capabilities, we use a corpus previously collected with a robotic character situated in a similarly dynamic multi-party interaction setting. The character was, however, placed in a different domain with different conversational goals and a different personality than our new character. From that dataset, we selected 5,291 dialogue excerpts in which our ASR system either recorded the input speech to consist of two or more consecutive utterances divided by small pauses, or the incoming speech was detected while the robot was in a speaking state itself. The system did not differentiate between the different speakers. Each excerpt is composed of the last sentence uttered by the robot, followed by all of the lines recorded by human interaction partners. As dialogue excerpt are often composed of multiple lines, the final number of training examples is 9,342.

A professional annotator who received training for this task was asked to label all dialogue excerpts in label-studio[1]. The annotator

[1] https://labelstud.io/

had access to two audio excerpts, one that only included the microphone input and one combined with the robot's utterances (Fig. 3). They then saw the raw ASR transcripts of the human speech and were tasked with applying one of eight labels per line. The labels match the different scenarios discussed in the previous section. We added "Other" to give the annotator the option to label anything that we had not foreseen in our initial screening of the data.

5 PRELIMINARY RESULTS & GUIDELINES

In this paper, we will focus on analyzing how prevalent each of the conversational scenarios outlined in Sec. 3 is in the dataset we annotated and how this can help us and the research community to prioritize developing features for robotic turn-taking models.

An analysis of the distribution of labels (Fig. 4) shows that the most common annotation is "Not to character". Indeed, more than a third of all utterances included in our samples were not directed to the robot but to someone else in the interaction group. Since interpreting these comments as valid input for the conversational topic has a high chance to take the conversation off track, classifying input by addressee should be an important focus point for developing dialogue systems. An initial concern was that reliably determining the addressee of a sentence based on audio data alone is too challenging. To evaluate this, we did not give the annotator access to the visual scene and asked them to determine the addressee based on the audio data alone. We found that the annotator only marked 3.6% of the utterances to have an unclear addressee, which gives us confidence that developing privacy-preserving dialogue models with high accuracy is possible.

The second most common phenomenon we observed in our data were utterances being continued either by the same or another person. Continuations by the same person make up about a third of all samples, and in about 13.7% of cases another person is taking the turn and continues speaking.

While our data do not allow us to determine whether the pause would have been an appropriate position for the robot to jump in and interrupt regardless of whether one of the human members of the party were planning to continue speaking, it shows that a simple turn-taking model that is based on the length of the pause alone may lead to frequent interruptions of human speech [Skantze(2021)]. It is hence advisable to base the prediction of turn-holding and turn-yielding cues on a more diverse set of features.

About 10% of the samples presented to the annotator were marked as interruptions. Keeping the microphone channel open and allowing the robot to process speech while it is still talking is therefore a valuable addition to a dialogue system. However, as the number of samples that were not directed to the robot are much higher than the actual interruptions of the robot's speech directed to it, classifying the addressee is a prerequisite to further deciding how to handle interruptions more naturally.

Both self-echo and background noise are phenomena that are rather infrequent in our corpus. This is likely related to a well tuned audio setup created for our character. We consider tuning the audio setup to the specific environment or applying other means of preprocessing incoming speech to filter echos and background noise fundamental for reliable intent detection of the language understanding unit.

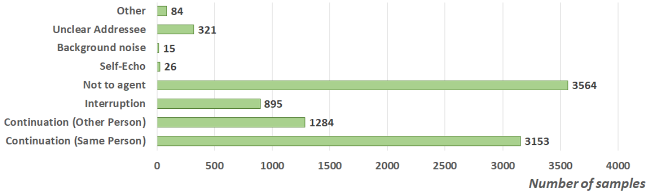


Figure 4: The number of samples by label

Interestingly, the annotator did label 84 samples as "Other". A manual review of these revealed a few conversational phenomena that we had not considered in our initial review of the data:

- **Backchannels:** In many of the examples labeled as "Other", the human gave a verbal backchannel like "Oh". Backchannels are likely even more common in our conversations, but are often omitted in the ASR transcriptions. As backchannels should not lead to any behavioral change in the robot, a separate classifier is only necessary if this is of specific research interest.
- **Coughing/Clearing throat:** While these can be considered background noise, the annotator decided to label them as "Other" as the source of the noise was one of the members actively engaged with the robot. Again, these do not have any implication for a robot's behavior, so we do not deem separately labeling them to be necessary.
- **Hallucinations:** This refers to instances in which an input transcribed by the ASR was impossible to relate to any audible phenomenon in the recording.

6 ONGOING & FUTURE WORK

We believe that improving the conversational turn-taking behavior for robots is a meaningful step for developing more natural and engaging interactions. Our work is currently focused on a classifier detecting turn-holding cues and investigating the utility of several features like lexical structure [Ekstedt and Skantze(2020)], prosody, volume, pitch and speed [Skantze(2021)]. At the same time, we are finalizing the new conversational setup and aim to collect and annotate an initial set of data in this new domain. For this data collection, we will transfer our models and test how well they are applicable to this character which has a different personality and conversational goals. We hope that our tests will show that the models we developed can be applied generically to improve a character's turn-taking behavior in multi-party settings. The next step will then involve the development of a classifier to incrementally decide whether a human speech act is directed towards the robot or to another member of the party as well as improving the filtering of artifacts transcribed by the ASR system. Moreover, we would like to test a physical instance of our character as well as a character with a different appearance to see whether our models are applicable across embodiments.

ACKNOWLEDGMENTS

We thank Kyna McIntosh & Michelle Ma for the character design, Samantha Ho for Fig. 2, and James Ryan for the data annotations.

REFERENCES

[Bohus and Horvitz(2011)] Dan Bohus and Eric Horvitz. 2011. Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions. In *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 98–109.

[Chao and Thomaz(2012)] Crystal Chao and Andrea L Thomaz. 2012. Timing in Multimodal Turn-Taking Interactions: Control and Analysis Using Timed Petri Nets. *Journal of Human-Robot Interaction* 1, 1 (2012), 1–16.

[DeVault et al.(2015)] David DeVault, Johnathan Mell, and Jonathan Gratch. 2015. Toward Natural Turn-Taking in a Virtual Human Negotiation Agent. In *2015 AAAI Spring Symposium Series*.

[Ekstedt and Skantze(2020)] Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2981–2990.

[Maat et al.(2010)] Mark ter Maat, Khiet P Truong, and Dirk Heylen. 2010. How Turn-Taking Strategies Influence Users' Impressions of an Agent. In *International Conference on Intelligent Virtual Agents*. Springer, 441–453.

[Mutlu et al.(2012)] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational Gaze Mechanisms for Humanlike Robots. *ACM Transactions on Interactive Intelligent Systems* 1, 2 (2012), 1–33.

[Rueben(2018)] Matthew Rueben. 2018. *Privacy-Sensitive Robotics*. Ph. D. Dissertation. Oregon State University.

[Sacks et al.(1978)] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A Simplest Systematics for the Organization of Turn Taking for Conversation. In *Studies in the Organization of Conversational Interaction*. Academic Press, 7–55.

[Skantze(2021)] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (2021), 101178.

[Traum(2004)] David Traum. 2004. Issues in Multiparty Dialogues. In *Workshop on Agent Communication Languages*. Springer, 201–211.

[Żarkowski(2019)] Mateusz Żarkowski. 2019. Multi-party Turn-Taking in Repeated Human–Robot Interactions: An Interdisciplinary Evaluation. *International Journal of Social Robotics* 11, 5 (2019), 693–707.