Learning Turn-Taking Behavior from Human Demonstrations for Social Human-Robot Interactions

Pourya Shahverdi[1], Alexander Tyshka[1], Madeline Trombly[1], and Wing-Yue Geoffrey Louie[1]

***Abstract*— Turn-taking is a fundamental behavior during human interactions and robots must be capable of turn-taking to interact with humans. Current state-of-the-art approaches in turn-taking focus on developing general models to predict the end of turn (EoT) across all contexts. This demands an all-inclusive verbal and non-verbal behavioral dataset from all possible contexts of interaction. Before robot deployment, gathering such a dataset may be infeasible and/or impractical. More importantly, a robot needs to predict the EoT and decide on the best time to take a turn (i.e. start speaking). In this research, we present a learning from demonstration (LfD) system for a robot to learn from demonstrations, after it has been deployed, to make decisions on the appropriate time for taking a turn within specific social interaction contexts. The system captures demonstrations of turn-taking during social interactions and uses these demonstrations to train a LSTM RNN based model to replicate the turn-taking behavior of the demonstrator. We evaluate the system for teaching the turn-taking behavior of an interviewer during a job interview context. Furthermore, we investigate the efficacy of verbal, prosodic, and gestural cues for deciding when to begin a turn.**

I. Introduction

Communicating with humans is a fundamental skill required of social robots. Humans communicate and interact through both verbal and nonverbal behaviors, and social robots should be equipped with human-like behaviors if we want them to interact naturally with humans [1]. Turn-taking is an especially essential nonverbal behavior used between humans because it enables fluid conversations [2]–[4]. This is because when a speaker takes the conversational floor at an inappropriate time it can disrupt the flow of a conversation [5], [6]. Namely, there are three ways to disrupt the flow of a conversation when inappropriately taking the conversational floor: 1) speaking before a turn has ended (interruption, overlap), 2) speaking too early after a turn has ended (short gap), or 3) speaking too late after a turn ended (long gap). These situations lead to the conversational partner perceiving the speaker's behavior as inappropriate due to the speaker not actively listening or not knowing when to take the conversational floor [4], [6], [7].

Context is important in turn-taking because it influences the appropriate time for a person to take the conversational floor [3], [4], [8]. Social interaction context is defined as any information that can be used to characterize a social interaction such as the social status, interaction goals, personalities, cultures, verbal and nonverbal behaviors, time, and environment of the interaction [9]. The influence of context can be exemplified by contrasting turn-taking behavior within a structured setting, such as an interview context, to turn-taking behavior in a less structured controversial debate. In an interview context, participants will seldom interrupt each other because doing so would be considered unprofessional. In contrast, it is often more common and acceptable in a controversial debate to interrupt the other speaker without waiting for them to finish their turn because participants are more prone to voicing their opinion or refuting the other's claims. In these examples, the social norms of the context affect participants' turn-taking behavior [3], [4], [10]. Hence, it is vital that social robots exhibit context-specific turn-taking behavior so they can effectively interact with humans according to the context of the social interaction [10]. However, it is infeasible to pre-program the turn-taking behavior of a social robot for all the potential social interaction scenarios it will face prior to it being deployed.

Computational models of turn-taking behavior are rapidly advancing from prior approaches that only considered single feature vectors over a brief window of time to identify a speaker's End of Turn (EoT) [6], [11]. These recent research works utilize state-of-the-art machine learning techniques such as Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) [12]–[14] and Transformer-based architectures [15] to model turn-taking behavior. These models are often trained on datasets gathered within specific social interaction contexts such as the MapTask corpus [11]–[13], [16], puzzle solving corpus [14], or a variety of human-robot interaction (HRI) scenarios [15], [17]. However, a general turn-taking model that could be applicable to all possible contexts has yet to be achieved, and models trained from data gathered in one context (e.g., MapTask) cannot be utilized in contexts it has not been trained on (e.g., HRI) [12]. It is also infeasible to predict all the potential social interaction contexts a social robot may face prior to deployment. Hence, there is currently an open opportunity to develop an approach to rapidly teach a robot context-specific turn-taking behavior after it has been deployed.

Learning from Demonstration (LfD) enables non-technical experts to teach a social robot new behaviors or tasks after it has been deployed [18], [19]. LfD has primarily been used to learn high-level social tasks, such as facilitating Bingo or robot-mediated therapy, and discrete robot actions such as a greeting [20]–[22]. LfD enables humans to demonstrate ideal behaviors in a distinctive social interaction context, and then teach the robot how to imitate these behaviors in a similar context. Furthermore, teaching a social robot

This work was supported by the National Science Foundation grant #1948224

[1]Intelligent Robotics Laboratory, Oakland University, Michigan, USA, 48309 (e-mail: louie@oakland.edu)

Digital Object Identifier (DOI): see top of this page.

using LfD enables it to implicitly capture and understand the context of the situation. Hence, LfD could be a potentially effective approach for teaching a social robot context-specific turn-taking behavior. This can be accomplished by training turn-taking models utilizing human demonstrations of the non-verbal behavior in a specific context. However, to date there has been limited research in learning nonverbal behaviors (e.g., turn-taking) from human demonstrations.

In this paper, we present a LfD system for a human to teach ideal turn-taking behavior within a specific social interaction context via demonstration and enable a robot to exhibit such turn-taking behavior within a similar context. Namely, we use LfD to capture a demonstrator's context-specific turn-taking behavior in response to the verbal, prosodic, and gestural cues of the individual they are interacting with during a specific social interaction context. External observation-based LfD is used to gather verbal, prosodic, and gestural data in a human-human interaction. This demonstration data is then used to train an LSTM RNN to model the demonstrator's turn-taking behavior. The model can then be applied to a robot to exhibit context-specific turn-taking behavior in a human-robot interaction based on the verbal, prosodic, and gestural cues of the human partner. We evaluate the performance of this LfD system in a dyadic interaction.

Overall, this paper has three primary research contributions. First, we present a LfD system which learns nonverbal behavior from human demonstration. Second, we extend prior turn-taking models by learning from demonstration context-specific turn-taking behavior from a limited number of demonstrations. This contrasts prior work that trains a single model of turn-taking behavior on a large dataset containing a variety of contexts and attempts to generalize to new contexts; but the results of these models have performed poorly in these new contexts [12], [17]. Third, we extend current research on turn-taking for chat bots and conversational agents to robots and investigate the effect of verbal, prosodic, and gestural cues on predicting the appropriate time for a robot to take the conversational floor. Current research for chat bots and conversational agents only utilize verbal and prosodic cues to predict a turn shift during human-computer interactions as opposed to selecting the most appropriate time to take a speaking turn based on context-specific social norms [8].

II. Related Works

Current state-of-the-art turn-taking approaches aim to learn a predictive and general model of turn-taking that can be applicable to all contexts by training a model with datasets containing a single or several contexts of conversational turn-taking [12], [13], [15], [17]. Namely, current predictive turn-taking models aim to predict a speaker's end of turn so that the model can be applied alongside other algorithms to determine the appropriate time for an agent (e.g., robot, chatbot, conversational agent) to take a speaking turn.

In [12], an LSTM RNN model was trained on the MapTask [16] dataset for predicting an end of turn and generalizing turn-taking predictions to new contexts. The dataset included 18 hours of spontaneous speech that was recorded from 128 dyadic conversations. A combination of the final Part of Speech (POS) tags and prosody features were used to train the LSTM RNN model. The model was evaluated on a subset of the MapTask dataset and a separate HRI dataset consisting of a robot actively listening (e.g., providing backchannels and follow-up questions) to users while they recounted their past travels. Results demonstrated that the LSTM RNN model outperformed the baseline silence-based and Inter-Pausal Units-based models in predicting turn-shifts. However, the model could not accurately predict turn-shifts in the novel HRI context without re-training in that context [12].

In [13], a larger RNN model with multiple LSTM layers was again trained on the MapTask dataset. Using only prosodic cues from the speaker, the trained model was more accurate in predicting the EoT than the model presented in [12]. The turn-taking model was also retrained on a Japanese speaking dataset and a telephone call dataset consisting of five different languages to investigate the effects of language on the model's performance. The model accurately predicted end of turns across four languages but did not accurately predict end of turns in Japanese. This suggests that prosodic features alone cannot predict EoTs in Japanese. The authors further elaborated that due to the variability in turn-taking, future work has the potential to improve performance by rapidly learning a novel style of interaction.

In [17], a study was conducted to investigate differences in the ability of LSTM RNN models to predict an EoT when trained on a dataset containing multiple contexts versus training on data collected in a specific context. A total of 105 human-robot interaction sessions were conducted with participants over a wide range of ages and backgrounds. Scenarios included a robot interviewing participants as candidates for a job (30 sessions), actively listening to participants (20 sessions), acting as a secretary during interactions with participants (19 sessions), acting as a single woman during a speed-dating scenario with participants (32 sessions), and guiding participants through a tour of a lab (4 sessions). The LSTM RNN models were trained using only prosodic features, only linguistic (verbal) features, or a fusion of the two features. The results of their study found that a model trained on all the aforementioned scenarios using verbal and prosodic features performed better than models trained in a single specific scenario when the context of interaction was closely related in structure (i.e., interview, speed-dating, secretary, active listening) to the scenarios found in the aggregated dataset but performed worse when the scenario's structure (i.e., job interview) was not close to those found in the aggregated dataset. The authors further elaborated that a generalized turn-taking model based on a large dataset is more suited for unstructured informal conversation, and structured task-dependent conversation would require training a model with data derived from the context to perform successfully.

TurnGPT [15] is an adaptation of Open AI's GPT-2 [23] and a transformer-based model for turn-taking. The TurnGPT model was trained with eight verbal datasets including: transcripts of dialogues between humans and automated assistants, human-human written dialogues, and scripts from

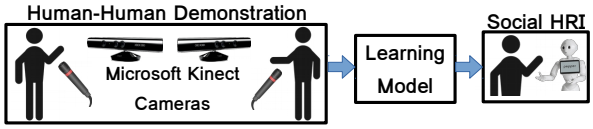


Fig. 1: Overview of the proposed LfD system for learning turn-taking behavior from human demonstrations

the MapTask and Switchboard corpus. Compared to POS model and text-based LSTM baselines, the TurnGPT model significantly outperformed both in prediction of EoTs. This advantage can be attributed to the fact that this model not only considers completion of a turn syntactically but also considers the pragmatic completion of a turn. However, the Turn-GPT model is designed for text-based data and lacks the important concept of time (i.e., when to take a turn given a high probability EoT word), which is an important ability for spoken dialogue.

Current approaches have focused on learning turn-taking models which generalize to many different contexts [12], [15], [17]. However, the results of these models have demonstrated that models trained on data from one context struggle to generalize to novel contexts [12], [17].

Lastly, current models determine an upcoming EoT using inputs including: verbal [15], prosody [13], and a combination of verbal and prosodic cues [12], [17]. To the best of our knowledge, the combination of verbal, prosodic, and gestural cues in detecting turn-shifts has yet to be explored. Studies have found that gesture could be a potentially useful cue for identifying whether a speaker is holding the conversational floor [8], [24]–[26].

In this work, the aforementioned gaps were addressed by investigating the impact of gesture, verbal, and prosodic cues on the performance of a turn-taking model and creating a LfD system which achieves the following: 1) learning context-specific task-oriented turn-taking from human demonstrations and 2) making discrete decisions of when to take a turn instead of only predicting turn-shifts.

III. Learning from Demonstration System

Our LfD system for a social robot to learn contextual turn-taking models is presented in Figure 1. The LfD system first gathers demonstrations of turn-taking during dyadic human-human social interactions within a specific context. The demonstrations are then used to train a LSTM RNN turn-taking model utilizing verbal, prosodic, and gestural cues to decide the appropriate time for a robot to take the next turn in the demonstrated context. This model can then be implemented on a social robot to identify the beginning of its turn to speak when it takes on the role of one of the individuals within the dyad for the demonstrated context. Namely, while predicting the EoT is the ultimate goal of current turn-taking models, our LfD system is able to learn a demonstrator's decision-making process on when to take a speaking turn during a conversation in a specific context.

A. Human-Human Demonstration Data Gathering

Our setup for gathering demonstration data during dyadic human-human social interactions within specific contexts is depicted in Figure 1. In our setup, the two individuals

TABLE I: Set of Interviewer Behavioral Questions

1	How are you today?
2	Tell me about yourself.
3	Are you good at working in teams?
4	Can you make decisions quickly?
5	What kind of skills do you think are important in research?
6	Are you quick in completing tasks?
7	Why do you find social robotics interesting?
8	Would you please tell me about your strengths?
9	Would you like to discuss some of your weaknesses?

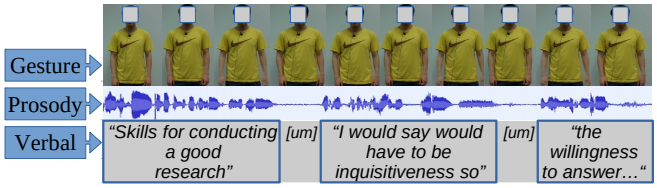


Fig. 2: A sample of data collected using our LfD system

are standing and facing each other during a natural social interaction. During the interaction two Microsoft Kinect depth cameras, one directed at each individual, are being used to record both participants' skeleton joint locations. Each individual is also wearing a lapel microphone which captures their audio during the interaction.

B. Job Interview LfD Scenario and Dataset

We utilize a job interview scenario as a representative example of a dyadic interaction where our LfD system could be applied. The procedures for collecting this dataset with our LfD system were reviewed and approved by the Institutional Review Board at Oakland university (#IRB-FY2022-103). Written informed consent was obtained from all participants prior to the data collection and participants could withdraw from the data collection at any time.

We had a researcher (32 years old and male) acting as an interviewer in order to simulate an end-user demonstrating a new turn-taking behavior to the robot within a specific context. The interviewer had a set of nine behavioral interview questions (Table I) which the interviewer could naturally vary in phrasing. The interviewer conducted interviews with undergraduate students using the set of behavioral questions as well as greeting and closing statements. Each student participated in three interview sessions consisting of the same questions phrased differently to increase the number of conversational turns taken and, consequently the size of the dataset for training a turn-taking model. The students were not provided any specific prompts on how to answer the interview questions other than to answer them naturally.

A total of five students were interviewed. The students were all English speakers with an age range of 21-24 years old (μ = 22). There were three male and two female participants. The average duration of the interviews was 1 minute and 52 seconds and the entire dataset included 28 minutes and 7 seconds of dyadic conversations. In total there were 150 turns taken within the dataset. A sample of data from our interview dataset is presented in Figure 2.

C. Feature Extraction

The features for our model can be divided into 3 categories: verbal, prosody, and gesture. All features are time-synchronized and sampled over 50 millisecond frames. Verbal

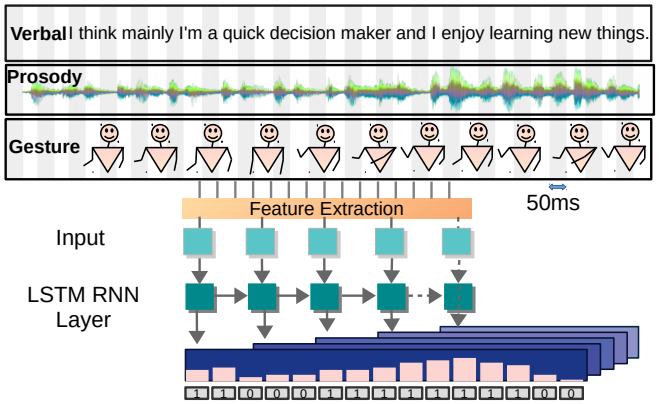


Fig. 3: Turn-taking model architecture

features are generated by applying an Automatic Speech Recognition (ASR) service, Google Speech-to-Text, to raw audio data and obtaining timestamped transcriptions as well as word-level confidences. These transcriptions often contained misrecognitions and we do not correct these so as to preserve the unlabeled as well as LfD nature of our data gathering. These transcriptions are then input to a TurnGPT model which has been pre-trained on the turn-taking datasets used in [15]. The output is a score for the probability of a word being an EoT. The scores on a word being an EoT are used as features for all frames after the speaker has uttered the word and before the end of the next word-level utterance. This choice simulates a causal, real-time system, where the features could not be computed until ASR outputs a word. The final verbal features for our model are the time-transformed ASR confidence and TurnGPT score.

For interviewee prosody features, we utilize a modified feature set found in [12] which included: voice activity, pitch, energy, and spectral flux. Voice activity was derived from the ASR system. The other features were computed using librosa [27] and z-normalized per speaker. Note that we use spectral flux rather than the spectral stability used by Skantze [12] because Ward [13] achieved better results using flux.

Finally, we derive interviewee gesture features via principle component analysis (PCA). The skeleton joint positions of the interviewee are first transformed to be relative to the torso frame. We sample the 3D euclidean position of both hands, both elbows, and the head and z-normalize these 15 features. We then perform PCA to reduce the dimensionality of the features to obtain five output features.

D. Turn-taking Model Architecture

Once the features have been extracted from the dataset, they are utilized to train a LSTM RNN based turn-taking model, Figure 3. Our model consists of one LSTM layer with 15 hidden neurons per LSTM cell. The model output is the probability that the interviewer should speak in each prediction frame for the next three seconds. Each frame comprises 50ms of data and we use 60 frames for a total of 3 seconds of prediction time. The LSTM output is transformed to the 60-dimensional output vector via a fully connected layer. Both our recurrent and fully connected layers use sigmoid activation functions, while the LSTM inputs use tanh.

E. Model Training

We employ k-fold cross validation to train the model and split our dataset of five speakers into four training samples and one test sample. This is due to the relatively small size of our dataset (28 mins) in comparison to typical turn-taking datasets (e.g., the popular Switchboard dataset which contains 240 hours of data [28]). We consider the participant's turn-taking behavior within the test sets as ground truth for the evaluation of the model's performance. In contrast to prior work that train models for both speakers on the dataset, we train our model from the perspective of a single speaker (i.e., the interviewer). The verbal features of both speakers are input to the TurnGPT model so its internal state can gain context from both sides of the conversation. However, the verbal features from the interviewer's current speaking turn is masked out from the output of TurnGPT because it would not be available while making a real-time decision on turn-taking for the interviewer. For prosody and gesture, we use only the features from the interviewee. These model design/training decisions ensure that the output of the model (i.e., turn-taking decision) is only determined by information available to the interviewer including history of the conversation and interviewee verbal as well as nonverbal behaviors. Specifically, our loss function targets the demonstrator's (interviewer) decisions on taking the turn for training the model while the loss functions for current state-of-the-art turn-taking models target the EoTs.

Given the small size of our dataset, we select hyperparameters that minimize the chance of overfitting. We train for 60 epochs using a batch size of 4 and a learning rate of 0.005. We use a 0.2 dropout on the LSTM inputs but do not apply dropout on recurrent connections. We also apply 0.001 L2 regularization on the LSTM and output layer weights to reduce overfitting. We use a loss function of mean squared error for training. To divide our data into training samples, we select windows of 10 seconds, using features from the first seven and labels from the last three. We designed our models in Tensorflow and trained our models on a 32-core AMD Ryzen CPU with 128 GBs of memory. The average model training time was 76 minutes.

F. Turn Decision Making

The output of the learned LSTM Model only provides probabilities on whether the interviewer will speak for each frame over the next 3 seconds. However, the choice in making a turn is discrete. Herein, we make discrete decisions with our model by first accumulating the predictions over a fixed time window of past predictions. Formally, this can be defined as:

$$pred_i = (\sum_{n=0}^{size} p_{t^i}^{t^{i-n}})/size \quad (1)$$

where $pred_i$ is the probability of whether to take a turn in the current frame i, *size* is the number of frames of past predictions to accumulate, and $p_{t^i}^{t^{i-n}}$ is the probability predicted at frame t^{i-n} of whether to take a turn at time frame t^i. In this case, a small window size of 10 frames (i.e., 0.5 seconds) will take into account more short-term predictions and be more responsive, whereas a larger window of 3s will take into account more long-term predictions, thus

having a filtering effect on any short-term variance. Applying a threshold to $pred_i$ will then allow us to make a discrete decision to take a turn or not.

IV. Experiments

We evaluate our system's ability to learn a model that replicates a demonstrator's turn-taking behavior in a dyadic social interaction utilizing an ablation test and F_1 scores.

A. Ablation Test

We evaluate how different features contribute to turn taking and identify the best performing model by conducting an ablation test while using Mean Absolute Error (MAE) as our measure of performance. We evaluated models using silence, verbal, prosodic, and gestural cues, as well as combinations of these cues. All models included silence (VAD) as a feature due to its critical importance for turn-taking [8].

B. F_1 Score Evaluation

While MAE provides an overall metric of the continuous prediction accuracy of the model, we note that evaluating a model only in the continuous domain fails to provide an idea of how the model will perform for real-world robot decision making on discrete turn or no turn decisions. To account for this, we utilize F_1 scores and precision-recall to evaluate the robot's discrete turn-taking decisions. The discretization approach previously explained in Sub-Section F of Section III requires tuning of the parameters including the accumulation window size and the probability threshold value. First, to find the optimal window size, we identified the best-performing model from the ablation test and plot the F_1 score versus threshold value for different window sizes. Maximizing the F_1 score gives us an idea of an optimal balance between False Positives (FP) and False Negatives (FN). For the purpose of this evaluation, a True Positive (TP) is counted as any prediction within two seconds of the ground truth turn-taking event. This two-second tolerance is according to studies on the distribution of human turn-taking latency [7]. Also, note that we do not consider probability thresholds below 0.2 due to the limitation that for very low thresholds, the model predicts always speaking, and given our rising-edge method of discretization, we would only predict a single turn for the whole sequence. For traditional classification tasks, lowering thresholds would increase FPs and lower precision (and consequentially F_1 score) but in this case lowering thresholds too far decreases FPs. Given that the metric does not provide meaningful information past this point, we omit evaluating probability thresholds below 0.2. Once we obtain the plot with the F_1 score versus threshold value for different window sizes we select the optimal accumulation window from it. This optimal accumulation window is then used to plot the F_1 score performance of the different models.

V. Results

The results of our ablation test and F_1 score evaluations are summarized in Table II and Figure 4.

A. Ablation Test

The results of our overall ablation test are described in Table II. We observe that the model with verbal and prosodic features performs best over all prediction lengths. Prosody scores second over all prediction lengths, in agreement with prior work [13]. Notably, gesture and the model with all features including gestures perform worst in this test. We attribute this to strong overfitting on the gesture features (training loss was 0.148 and evaluation loss was 0.217).

B. F_1 Score Evaluation

The plots for the F_1 score versus threshold value for different windows sizes for the verbal and prosody model are shown in Figure 4a. The optimal accumulation window was ten previous frames. The plot using this accumulation window to determine the performance of the different models based on changing probability thresholds is presented in Figure 4b. These results suggest overall performance remains mostly stable across a significant range of thresholds from 0.2 to 0.7. We note that aside from the worst-performing models which used gesture cues, the others appear marginally different. Given this result, we turn to a more granular approach of plotting precision and recall to determine the differences in FPs and FNs. We plot precision and recall using the same value of accumulation window size of ten in Figure 4c. From this figure, we can see that the verbal and prosodic model skews toward optimizing precision while simpler models such as silence sacrifice precision for recall.

VI. Discussion

In this study, we present a LfD based system capable of making decisions on the appropriate time to take a turn during a one-on-one social interaction context such as in an interview scenario. We argue that creating a model generalizable to numerous contexts can be difficult and may not be necessary in some cases. Instead, gathering small amounts of data using a LfD approach has proven valid for learning turn-taking behavior as demonstrated in our interview context. Our model is also the first to demonstrate the benefits of including transformer-based verbal features from TurnGPT in combination with nonverbal features for turn-taking. Including this data provides a sense of grammatical sentence completion and helps indicate whether the interview question was completely answered; these are features that simpler POS verbal features could not capture [15].

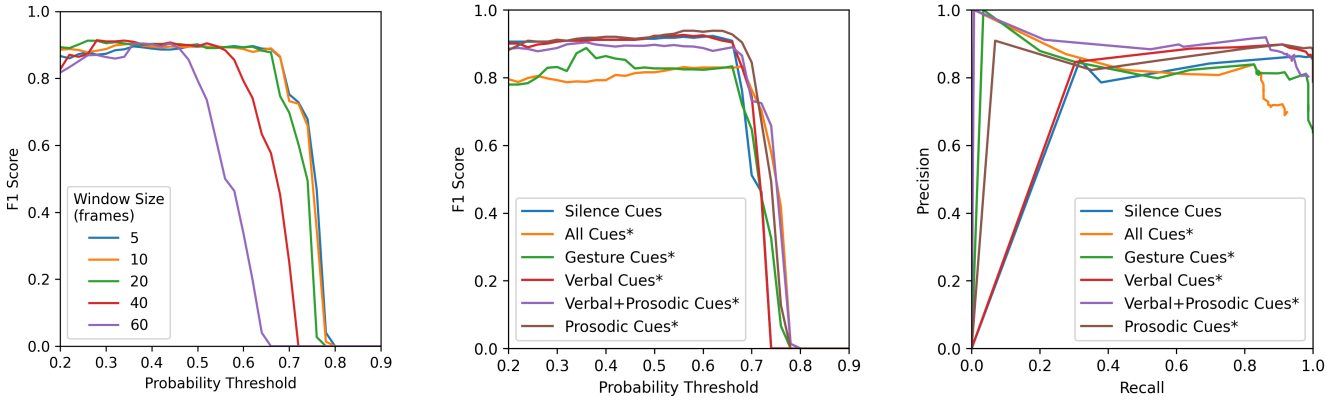
In terms of adapting a model to a specific context, our results show that we can use a limited dataset of interviewer turn-taking behavior to build a performant model for that context. Both gathering and labeling of data can be prohibitively difficult. Our approach limits the former and eliminates the latter, which is a significant finding for adapting robots to new social interaction contexts after they have been deployed. However, overfitting is a significant challenge with high-dimensional features like gesture. Our findings demonstrated that the lower-dimensional features provided by prosody and verbal features were better for generalization in this context.

One phenomenon that we also observed in our data was that FPs were strongly correlated with the interviewees' use

TABLE II: Mean absolute error ablation test results (mean ± std. dev. across participants)

Prediction Length	Silence	Verbal*	Prosody*	Gesture*	Verbal+Prosody*	Verbal+Prosody+Gesture*
250ms	0.270±0.031	0.272±0.024	0.263±0.027	0.327±0.045	**0.258±0.023**	0.307±0.045
500ms	0.274±0.028	0.277±0.021	0.268±0.025	0.334±0.042	**0.264±0.022**	0.315±0.043
1s	0.287±0.023	0.290±0.016	0.281±0.021	0.351±0.033	**0.278±0.021**	0.334±0.038
2s	0.319±0.018	0.321±0.013	0.312±0.017	0.381±0.016	**0.311±0.019**	0.367±0.024
3s	0.342±0.016	0.343±0.013	0.335±0.014	0.396±0.011	**0.334±0.018**	0.385±0.015

* Feature set also includes the silence feature.

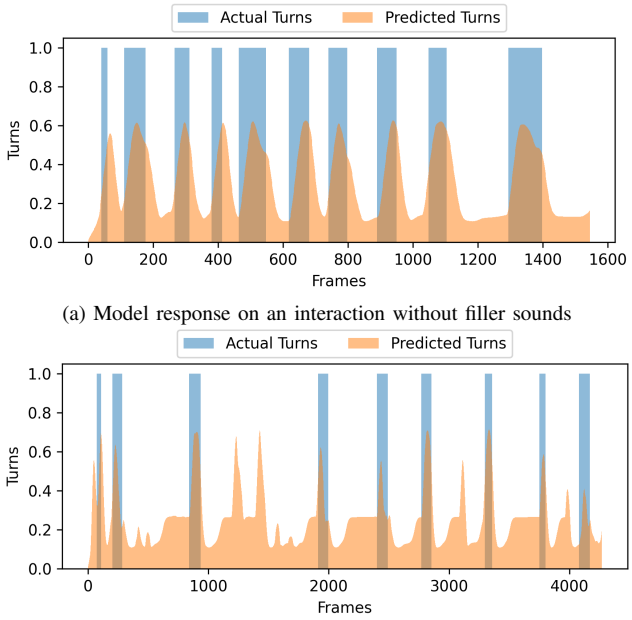


(a) Verbal+Prosodic model F_1 scores for different sizes of accumulation windows

(b) F_1 score over probability thresholds for different models

(c) Precision-recall curve for models trained on different turn-taking cues

Fig. 4: F_1 score and precision-recall



(a) Model response on an interaction without filler sounds

(b) Model response on an interaction that includes filler sounds

Fig. 5: Comparison of the predicted probability to take a turn against the ground-truth turns for two different test inputs

of filler sounds such as 'um'. This is illustrated in Figure 5. Figure 5a shows model performance with a speaker who did not use filler sounds and the model correctly identifies all TPs and produces no FNs or FPs. On the other hand, the speaker in 5b frequently utilized filler sounds. Our ASR and VAD detection systems did not detect such occurrences but instead labeled these instances as silence because recognition of disfluencies (e.g., filler sounds) remains an open research challenge and tools are still unavailable to reliably recognize them [29]. The model falsely predicts to take the turn when presented with these misrecognitions of long silences. Future models should aim to address such disfluencies because they are common in human speech and play an important role in signaling an incomplete turn in natural human social interactions [30]. Hence, it is necessary to address these challenges with disfluencies in future work and we hypothesize that representing these instances in our feature set would greatly improve the model's performance.

Another notable finding in the data is the seeming disparity between the MAE evaluation and precision-recall evaluation. The verbal and prosody model performs best on MAE but not necessarily on precision-recall. This suggests that the model is better at predicting near and long-term speaking activity than when predicting exact speaking onset. In other words, this model makes conservative judgments about when to speak whereas the other models are more eager. We hypothesize that this phenomenon is also correlated with the issue of filler sounds because such instances may penalize the model for being eager and prompt during training. Consequently, the model utilizes late turn-taking to better optimize the loss function. We again expect that the results would improve on recall-precision if filler sounds were better represented.

We believe that in future work it is important to further study the ideal trade-off between early turns and late turns from a human-robot interaction perspective. Objective metrics such as F_1 score weigh each equally but this does not account for subjective human evaluations on socially appropriate turn-taking policies or their specific expectations of robot turn-taking. Prior work further indicates that the context affects this trade-off [3], [4], [10]. For example, in a job interview one would be more wary of interruption than in casual conversation with a friend. Moreover, it may be possible

to recover from mistakes in turn taking. An example would be identifying an interruption has occurred and yielding the turn much like a human would.

Lastly, interviews are fairly structured, and interruption patterns, filler words, and backchanneling are likely to be different in other unstructured contexts [8], [10]. Our approach performed successfully in a structured setting and we plan to evaluate this approach for training a model in unstructured settings (e.g., a controversial discussion) as our future work. Also, interruptions during HRIs have been studied in several other contexts but as far as we are aware, interruptions during turn-taking has yet to be explored [31]. Such scenarios are challenging to study as systems that change the flow of conversation are closed-loop in nature and less friendly to offline training and evaluation. Often the solution for closed-loop learning in robotics involves simulation environments, but this approach is infeasible for this task given that we still lack a full model of turn-taking [7].

VII. Conclusion

In this paper, we present a LfD approach for a robot to learn appropriate context-specific turn-taking behavior from human demonstrations. In comparison to existing work in turn-taking that learns a model that predicts the EoT for a speaker and has an agent speak immediately after, our model specifically learns when a robot should speak in a given context. This accounts for differences in social norms as well as appropriate uses of short gaps, long gaps, and interrupts in turn-taking within different social contexts. Results from experiments on applying our LfD approach to a job interview context demonstrates that our system can learn a turn-taking model that replicates human-like turn-taking behavior in the given context. Furthermore, we evaluated the role of verbal, prosodic, and gestural turn-taking features for enabling a learned model to accurately make a decision on when a robot should take a speaking turn. Ablation analysis on these features suggest that the combination of verbal and prosodic features perform better in training a context-specific model, with limited demonstration data, to determine when a robot should take a speaking turn.

References

[1] D. O. Johnson, *et al.*, "Socially assistive robots: a comprehensive approach to extending independent living," *International Journal of Social Robotics*, vol. 6, no. 2, pp. 195–211, 2014.

[2] E. J. Baesler and J. K. Burgoon, "Measurement and reliability of nonverbal behavior," *Journal of Nonverbal Behavior*, vol. 11, no. 4, pp. 205–233, 1987.

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

[4] W. J. Levelt, *Speaking: From intention to articulation*. MIT press, 1993.

[5] J. M. Wiemann and M. L. Knapp, "Turn-taking in Conversations," *Journal of Communication*, vol. 25, no. 2, pp. 75–92, 1975.

[6] J. Holler, K. H. Kendrick, M. Casillas, and S. C. Levinson, *Turn-taking in human communicative interaction*. Frontiers Media SA, 2016.

[7] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, p. 731, 2015.

[8] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, p. 101178, 2020.

[9] G. D. Abowd, *et al.*, "Towards a better understanding of context and context-awareness," in *International symposium on handheld and ubiquitous computing*. Springer, 1999, pp. 304–307.

[10] T. P. Wilson, J. M. Wiemann, and D. H. Zimmerman, "Models of turn taking in conversational interaction," *Journal of Language and Social Psychology*, vol. 3, no. 3, pp. 159–183, 1984.

[11] M. Johansson, T. Hori, G. Skantze, A. Höthker, and J. Gustafson, "Making turn-taking decisions for an active listening robot for memory training," in *International Conference on Social Robotics*, 2016, pp. 940–949.

[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] N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, "Turn-taking predictions across languages and genres using an lstm recurrent neural network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 831–837.

[14] A. Maier, J. Hough, D. Schlangen, *et al.*, "Towards deep end-of-Turn prediction for situated spoken dialogue systems," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 1676–1680, 2017.

[15] E. Ekstedt and G. Skantze, "Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog," *arXiv preprint arXiv:2010.10874*, 2020.

[16] A. H. Anderson, *et al.*, "The hcrc map task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.

[17] D. Lala, K. Inoue, and T. Kawahara, "Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 78–86.

[18] C. G. Atkeson and S. Schaal, "Robot learning from demonstration," in *ICML*, vol. 97, 1997, pp. 12–20.

[19] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[20] M. Clark-Turner and M. Begum, "Deep reinforcement learning of abstract reasoning from demonstrations," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 160–168.

[21] W.-Y. G. Louie and G. Nejat, "A social robot learning to facilitate an assistive group-based activity from non-expert caregivers," *International Journal of Social Robotics*, pp. 1159–1176, 2020.

[22] K. Winkle, *et al.*, "In-situ learning from a domain expert for real world socially assistive robot deployment," *Proceedings of Robotics: Science and Systems. Corvalis, Oregon, USA.*, 2020.

[23] A. Radford, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[24] S. Duncan, "Some signals and rules for taking speaking turns in conversations." *Journal of personality and social psychology*, vol. 23, no. 2, p. 283, 1972.

[25] M. Zellers, D. House, and S. Alexanderson, "Prosody and hand gesture at turn boundaries in swedish," *8th Speech Prosody 2016, 31 May 2016 through 3 June 2016*, pp. 831–835, 2016.

[26] J. Holler, K. H. Kendrick, and S. C. Levinson, "Processing language in face-to-face conversation: Questions with gestures get faster responses," *Psychonomic bulletin & review*, vol. 25, no. 5, pp. 1900–1908, 2018.

[27] B. McFee, *et al.*, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[28] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. USA: IEEE Computer Society, 1992, p. 517–520.

[29] T. Kourkounakis, A. Hajavi, and A. Etemad, "Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2986–2999, 2021.

[30] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and Speech*, vol. 44, no. 2, pp. 123–147, 2001.

[31] P. Saulnier, E. Sharlin, and S. Greenberg, "Exploring minimal nonverbal interruption in hri," in *2011 RO-MAN*, 2011, pp. 79–86.