Smooth Turn-taking by a Robot Using an Online Continuous Model to Generate Turn-taking Cues

Divesh Lala
Graduate School of Informatics
Kyoto University
Kyoto, Japan
lala@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Graduate School of Informatics
Kyoto University
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Graduate School of Informatics
Kyoto University
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

Turn-taking in human-robot interaction is a crucial part of spoken dialogue systems, but current models do not allow for human-like turn-taking speed seen in natural conversation. In this work we propose combining two independent prediction models. A continuous model predicts the upcoming end of the turn in order to generate gaze aversion and fillers as turn-taking cues. This prediction is done while the user is speaking, so turn-taking can be done with little silence between turns, or even overlap. Once a speech recognition result has been received at a later time, a second model uses the lexical information to decide if or when the turn should actually be taken. We constructed the continuous model using the speaker's prosodic features as inputs and evaluated its online performance. We then conducted a subjective experiment in which we implemented our model in an android robot and asked participants to compare it to one without turn-taking cues, which produces a response when a speech recognition result is received. We found that using both gaze aversion and a filler was preferred when the continuous model correctly predicted the upcoming end of turn, while using only gaze aversion was better if the prediction was wrong.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Supervised learning by classification**; *Discourse, dialogue and pragmatics.*

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353727

KEYWORDS

turn-taking, multimodal interaction, human-robot interaction, machine learning, online model

ACM Reference Format:
Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth Turn-taking by a Robot Using an Online Continuous Model to Generate Turn-taking Cues. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3340555.3353727

1 INTRODUCTION

A long-term goal for android research is a robot which can hold a conversation with humans while maintaining human-likeness. Although there have been improvements for spoken dialogue systems in terms of natural language processing [29], there are other requirements for androids since their realistic appearance influences the expectations of the user [1, 10, 30]. In spoken dialogue systems such as smart speakers, conversational phenomena such as backchannels, fillers, and gaze behavior are redundant or even non-existent, but these are desired for androids since the goal is to match real human behaviors in conversation.

In this work we address human-like turn-taking, which is the switching in conversation from one speaker to another. It is known that across many languages and cultures there is little silence between turns in human conversation [13, 27]. Humans often respond as soon as the other turn has ended, even overlapping with the end of the previous turn. Coordination of turn-taking occurs naturally and even in cases of interruptions the conversation can continue smoothly.

On the other hand, turn-taking in human-robot conversations is more structured. Users often have to wait for the robot to recognize and then generate a response to their utterance. Designers of conversational systems aim to prevent the system from interrupting the user and disrupting the flow of the conversation. Although this method is safer, turn-taking is slower and not as human-like. Previous research has claimed users do not need fast response times from a robot [24]. However, we argue that increased user expectation due to the realism of androids means human-like turn-taking speed needs to be considered for these types of robots.

The major limitation of online turn-taking is that the system cannot respond as fast as a human. An automatic speech recognition (ASR) system can only generate an utterance result after a pause has been detected by a voice activity detection model, so there will always be a period of silence after the user's turn. On average there is about 100-200ms of silence in between turns in human conversation [11, 27] and in many cases turn overlap, but even fast ASR systems will not generate a result within this time.

Our research goal is to implement a turn-taking system for an android which exhibits human-like speed and behaviors. In order for human-like turn-taking speed, the system should take the turn before an ASR result has been received. Our approach is to generate multimodal cues which indicate the turn is going to being taken. We predict the upcoming end of the turn while the user is speaking, so that these behaviors are timed close to the end of the user's speech. We also evaluate if this approach is acceptable for users.

The next section details two main types of turn-taking models used in previous work and why these have limitations. We then present our combined architecture in Section 3. A continuous model to predict turn-taking cues is described in Section 4, and evaluated in Section 5. We then describe our subjective experiment in Section 6, the results in Section 7, before discussions of the outcomes of our work. In this paper Japanese is used as the target language.

2 TURN-TAKING MODELS

End-of-turn prediction is the problem of determining if the speaker has finished their turn. We assume a fixed silence threshold for turn-taking is always sub-optimal, as has been shown in other works [11, 15]. There are two main machine learning approaches.

The first approach uses the inter-pausal unit (IPU) or another lexical unit from an ASR system as a basis for prediction, as shown in Figure 1. Lexical information can be used to determine the end of the turn, such as if the IPU is a question. Prosodic and filter-bank features of the IPU can also be used as inputs to the model. Such models have been well studied in the literature [7, 11, 14–17].

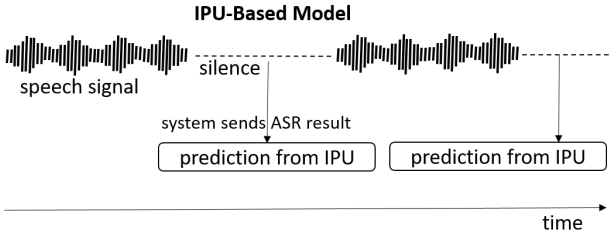

Figure 1: Diagram of IPU-based turn-taking model.

The second approach is a continuous model, shown in Figure 2. In this model prediction is done continuously (e.g. every 100ms), using only features which can be extracted in real-time. This includes prosodic information such as F0 and power, and also non-verbal signals such as eye gaze. Implementations of this model have become popular due to deep learning techniques [5, 22, 25].

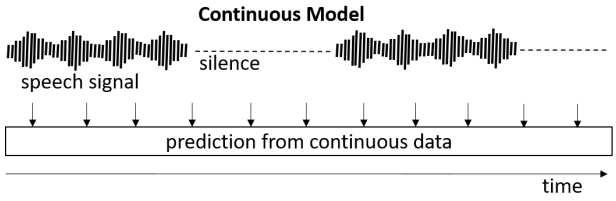

Figure 2: Diagram of continuous turn-taking model.

There are strengths and weaknesses to each approach for an online system. The IPU-based model is arguably more robust, since lexical content is a good indicator of the end of a turn. However, there is a limit to the speed at which we receive this IPU, since it cannot be recognized before a pause is detected. This detection means the IPU will not be received until some time after the user has finished their turn. A solely IPU-based approach cannot produce the very short or overlapping turn-taking times found in real conversation.

The continuous approach mitigates this problem as turn-taking prediction can be done at any time. However, this may be less accurate since we do not have lexical information and must predict using audio and body data streams, which may be unreliable. Continuous models have been implemented which predict the next speaker [5, 22, 25], but these make predictions after a defined pause time which is longer than human-like turn-taking speed. Incremental ASR can provide IPUs very quickly [25], but other work has shown there would still be a high number of false cut-ins because of relatively long pauses during the turn [11, 15].

3 HYBRID TURN-TAKING MODEL

We propose a hybrid turn-taking model which combines the speed of a continuous model with the more reliable IPU-based approach. The concept is shown in Figure 3.

We predict the end-of-turn while the user is speaking using a continuous model. The actual end-of-turn is a discrete time point, so it predicts if the turn will end within a certain period of time (e.g. 500ms). When the system does this, it should output some behavior which cannot be a response to the user's speech, since the ASR result is still unknown.

We define this behavior as a *turn-taking cue* which signals that a turn may be taken, but is not a response to the utterance. We focus on two types of turn-taking cues. Gaze

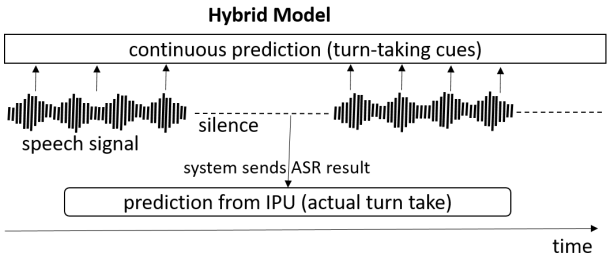

Figure 3: Diagram of proposed hybrid turn-taking model.

aversion is known to indicate that a speaker is beginning their turn [9, 20], where both parties engage in mutual gaze at the end of the first speaker's turn and then the second speaker averts their gaze during the turn switch. The second cue is a verbal filler. Fillers, or filled pauses, are non-lexical and used to regulate turn-taking [23]. In a previous analysis of our corpus it was found that 16.4% of turns used fillers as the first utterance [12].

Previous works analyzed fillers as a way to avoid unnecessary silences [18, 21, 24], but this is different from using fillers as a way to grab the turn quickly. Gaze aversion was also found to be successful in turn-taking management [3, 8], although participants read from a piece of paper so gaze information was a strong end-of-turn indicator. Another work similar to this paper implemented a method which used filler, gaze, smiling and breathing as turn-taking cues [26]. However, this scenario was a task with users looking at a table of objects, so again gaze behavior was a strong indicator of the end of a turn. It also used pauses as a basis for turn-taking behaviors and was heavily hand-crafted. We distinguish our work from others by targeting free conversation.

When an ASR result is received, the IPU-based model can more reliably detect if the speaker's turn has ended. In this work we assume we have a reliable IPU-based model to predict the actual end of the turn. Our goal is to show that even with an accurate IPU-based model and fast ASR, the addition of a continuous model to take the turn quickly will improve the overall system.

Our hybrid model makes two end-of-turn predictions - a continuous prediction to generate turn-taking cues and an IPU-based prediction for the actual response. Continuous prediction is done during speech, while more robust IPU-based prediction is done upon receiving an ASR result. Our approach predicts the end of the turn during speech rather than at pauses, so fast and overlapping turn-taking becomes possible.

4 CONTINUOUS MODEL IMPLEMENTATION

Our hybrid model requires both an IPU-based and continuous model. Previous work has already shown robust online IPU-based models can be implemented [11, 15–17], so rather than retraining and evaluating a new model, we assume that we are using an existing one. This section describes the implementation of the continuous model only, which is novel for this work.

Data collection

To train our model, we use 64 sessions of one-to-one interactions with a subject and an android robot, ERICA [4]. The corpus contains several scenarios, such as job interviews, speed dating and attentive listening [11]. Subjects were told of the scenario prior to interaction. ERICA was operated by a hidden remote operator - one of five trained voice actresses used during the sessions. The voice of the operator was synchronized with ERICA's mouth movements so that the conversation could be conducted naturally and without fixed responses. Each session lasted from 5-20 minutes with both genders and a diverse range of ages. Data recorded for each session was audio data through both a fixed microphone and microphone arrays. Body data of the subject was captured using a Kinect sensor. IPUs were fully transcribed, including annotations for fillers and backchannels.

Turns were also manually annotated as transition relevance places (TRPs). TRPs are points in the conversation where a turn could have switched, even if it did not occur in the actual conversation [23]. For example, a speaker may ask a question but get no response so asks it again. The end of the first question is a TRP, since a turn could have changed. We train the data using TRPs rather than the turns themselves, but use the terminology "turn" for simplicity.

Model details

We constructed two separate models - an operator model which was trained only on the operators of ERICA, and a subject model trained only on the subjects. Our objective is to predict the upcoming end of a turn (TRP) while the user is speaking. Technically, this is the time point when the speaker stops speaking, but defining it as such means the labels will become severely unbalanced. To reduce this imbalance, we label a positive classification as a time point where the end of the turn will occur within the next 500ms. We do not make any predictions within the first second of a turn or outside of IPUs. Based on our previous analysis of this corpus, we found other humans prefer a response time window of 200-500ms after user silence [12]. However, we also consider the processing time required to do the prediction and generate the robot speech and behaviors. The real system will generate a behavior slightly later than the actual ground truth, and this timing will actually be closer to the ideal timing window.

When constructing the model, we ensured that we could extract all the features in real-time. For audio features, we used F0 measured in hertz and power measured in decibels,

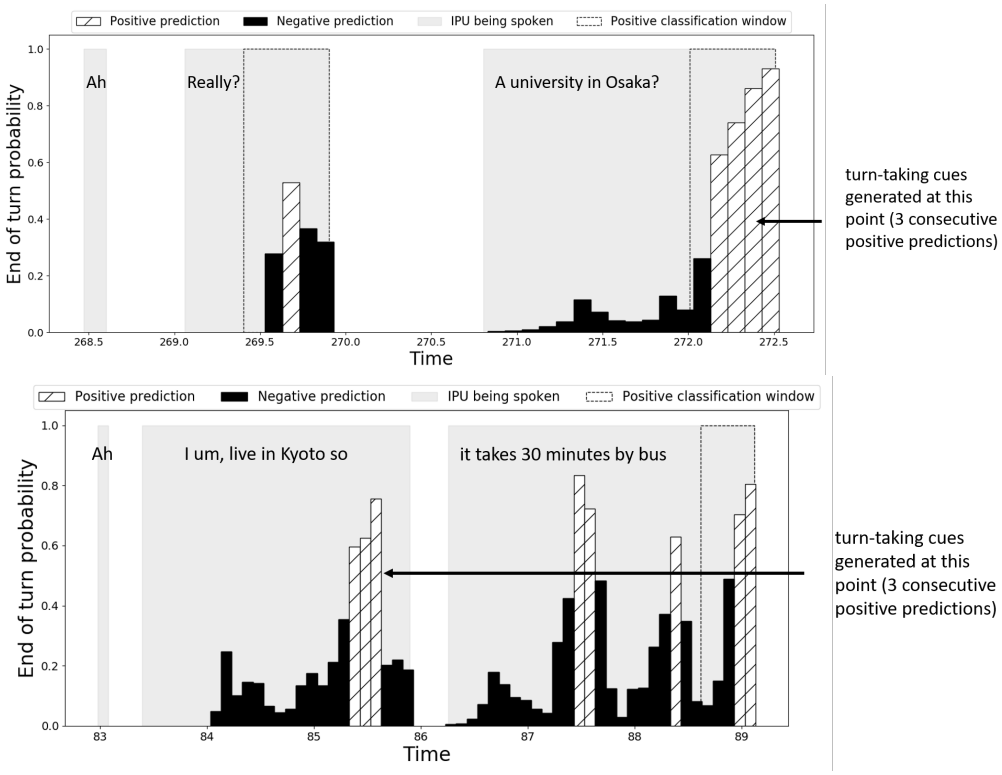

Figure 4: Visualizations of the continuous model prediction for turns classified as correct (top) and wrong (bottom), with $cpp = 3$. The dashed window indicates the target prediction window, 500ms before the end of a turn/TRP. Text corresponds to the actual IPU utterance. Note that in the correct turn there are two TRPs and the model correctly predicts the second one.

extracted with an online pitch tracker [6]. Each sample contained one second of user data with 100 frames of audio data per sample. Samples were taken every 100ms of speech. We also removed outlier F0 values. For each F0 value in the sample we checked if its value was 80-120% of the previous F0 value. If not, we set its value to zero and considered them unvoiced. We only kept sequences of F0 values which were at least 50ms in length and considered these to be voiced. This pre-processing can be done in real-time.

For every frame, we extracted 17 basic audio features which we assumed would be helpful for discriminating the end of a turn from mid-turn speaking and can be easily replicated in spoken dialog systems. These are:

- Raw F0 and power
- Δ and $\Delta\Delta$ of F0 and power
- maximum, minimum, range, slope and standard deviation of voiced F0 and power of previous 100ms
- ratio of voiced to unvoiced pitch in previous 100ms

For the subject model, we also extracted gaze features, using a Kinect sensor to estimate head direction at a lower sampling rate (approximately every 30ms). Gaze features were only used for the subject model, since the operator is in a remote location. The frame-based features for gaze were the differences of the 3-dimensional head positions and orientations. We performed late fusion of audio and gaze features using concatenation of hidden states to account for the difference in sampling rates. All features were z-normalized over the particular sample, so we do not have any speaker-dependent information used to train the model.

Since we have sequential data, we opted to use an LSTM network with 128 nodes and 3 layers. The batch size was 32 and dropout was used at a rate of 20%. We split the data by session, using a 60:20:20 ratio for training, validation and test sets. In total we used 64 sessions of data. The operator model contained 105,719 samples with 13,125 (12.41%) labeled as positive. The subject model contained 116,401 samples with 10,699 (9.19%) labeled as positive.

5 MODEL EVALUATION

In this work we assume the IPU-based model is perfect, but for reference we provide an estimate of its performance. We analyzed several Japanese turn-taking models and found the model by Masumura et al. [17] trained on a call center corpus to be best performing (F1 score = 0.821). The model trained

Table 1: Frame-based classification results for continuous models.

Model	Precision	Recall	F1
Proposed operator	0.409	0.491	0.446
Proposed subject	0.323	0.151	0.206
Proposed subject + gaze	0.327	0.169	0.223

on our corpus has lower performance (F1 score = 0.592) [11] but reports a metric where the average response time was around 1200ms for a false cut-in rate of 5%.

We evaluated our continuous model for every 100ms of speech data in the test set. Results are shown in Table 1 for operators only and subjects only.

The continuous models are less accurate than the IPU-based models, justifying our need for a hybrid model. The operator-only model is better than the subject-only model, likely because there were only five female operators used for the samples and the quality of the audio was higher as they were in a soundproof room. Gaze features did not improve the subject model by much, possibly due to the inaccuracy of the Kinect sensor.

This type of evaluation gives no indication about real system performance. Our test samples consist of many point-wise evaluations of the same turn so we need a more appropriate way to test turn-based performance.

We performed a turn-based analysis of our test set by identifying the time point during the turn where the system predicted the upcoming end of the turn. Since we can order our test samples chronologically, we know the probability of the end of a turn over its duration. Examples of visualizations of turns are in Figure 4.

We classify a turn depending on the time a positive classification is *first* predicted. If it is predicted within 500ms of the end of a TRP (Figure 4 inside the positive classification window), we classify it as *correct*, else it is *wrong*. If a correct prediction occurs later than a wrong prediction it is still classified as wrong. If no positive prediction is made during the turn, it is classified as *missed*. Many predictions are made during the turn so classification is not based on receiving one positive prediction, but *consecutive* positive predictions, which we represent as *cpp*. We classified every turn in our test set, with results shown in Table 2.

Reducing *cpp* decreases missed classifications but increases wrong classifications. Differences between the operator and subject model are because of the subject model's weaker metrics in Table 1. From these results we opted for $cpp = 3$ for the operator model and $cpp = 2$ for the subject model for the best balance between precision and recall. The model

Table 2: Turn classification for continuous models. *cpp* is the number of consecutive positive predictions needed to be classified as an end of turn.

Operator model (total turns = 452)

		cpp value 1	2	3	4
Classification	Correct	29.4%	35.8%	32.5%	21.2%
	Wrong	63.5%	49.1%	35.6%	25.0%
	Missed	7.1%	15.0%	31.9%	53.8%
		100%	100%	100%	100%

Subject model (total turns = 449)

		cpp value 1	2	3	4
Classification	Correct	20.9%	18.3%	10.9%	6.5%
	Wrong	49.2%	31.2%	18.3%	9.8%
	Missed	29.8%	50.5%	70.8%	83.7%
		100%	100%	100%	100%

will generate turn-taking cues at the time point when the final positive prediction has been received.

The objective evaluation of our model does not indicate how correct it is from a user perspective. One motivation behind turn-taking cues is to allow the conversation to continue smoothly even if we wrongly predict a turn end. In the next section we describe a user experiment to evaluate this.

6 EXPERIMENT METHODOLOGY

The goal of our experiment is to test if our hybrid model can improve an IPU-based model which only produces fast responses. Participants watch videos of ERICA to compare turn taking generated under different types of models. Although a live interaction experiment is possible, there are many unrelated factors we have to control for and we would also need a robust conversational spoken dialogue system which at this time is difficult to implement. In our experiment we can simulate the different conditions while controlling all other factors, including the content of the dialogue.

Turn-taking cues

First we designed turn-taking cues for ERICA to use in the hybrid model. ERICA gazes to the left or right side of the user's head, with a 50% chance per side. Looking upwards is also a gaze aversion behavior, but this is normally used when considering an answer to a question [3], and our model will not know the content of the user's utterance. We assume the user's head is situated in front of ERICA at eye level. To ensure that the gaze would be noticeable, we defined

that ERICA shift her focus to a random point between 30-40 degrees horizontally and up to 5 degrees downwards from her original gaze at the user. ERICA averts her gaze for up to 2500ms unless a response is needed, in which case she starts to gaze back at the user 500ms after she starts speaking.

Choosing an appropriate filler is also crucial [2, 19]. Although ERICA can generate many realistic-sounding fillers, the continuous model does not know the content of the user's utterance. Therefore, a filler as an emotional response would be unnatural for a simple question. We analyzed the lexical form of turn-taking fillers (fillers used at the beginning of turns) in our corpus according to categorizations in previous work [12, 19]. The two most common were notice fillers (*a* in Japanese) and proper fillers (*etto* in Japanese) comprising about 50% and 20% of all turn-taking fillers, respectively. We restricted filler usage to these two forms and chose fillers which were neutral in tone. The eventual choice of filler is based on a random distribution - a 67% chance of a notice filler and a 33% chance of a proper filler.

Simulation of turn-taking models

We extracted corpus samples from our test set consisting of one turn, followed by a response. The initial turn is spoken by either the subject or operator. The response was said by ERICA using synthesized text-to-speech, and the content of the response was what was actually spoken in the corpus. The continuous model used for evaluating a turn depended on the speaker of the initial turn (subject or operator).

We included both correct and wrong (turn-taking cues are generated more than 500ms before the end of a turn) samples in our experiment, from both the operator and subject models described in Section 5. We selected samples where the initial turn was short, but the context could still be understood, to prevent subject fatigue. We chose samples without cross-talk such as backchannels. In total, we selected a total of 51 samples (24 correct and 27 wrong) for the experiment.

In our experiment we simulate the decision made with the continuous model using the optimal *cpp* values defined in Section 5. We use the test set as samples, so the time point of the turn-taking cues is known precisely. For each sample we generated three simulations corresponding to different turn-taking models:

- **IPU (baseline)**: The continuous model is not used and ERICA only responds after the initial turn has ended. After tests using our current system, an ASR result will arrive approximately 700ms after the user has finished speaking if we use an end-to-end model [28]. This also assumes 100ms for IPU-based model processing. Therefore, ERICA also generates her response 700ms after the end of the initial turn.
- **Gaze**: If the decision is made to generate a turn-taking cue, ERICA will perform gaze aversion. 700ms after the end of the initial turn, ERICA will generate a response.
- **Gaze + Filler (G+F)**: If the decision is made to generate a turn-taking cue, ERICA will perform gaze aversion and also produce a filler. To limit the number of consecutive fillers, if a positive decision is found, ERICA will only say a filler if five seconds or more have passed before a previous filler has been said. She will always do gaze aversion for a positive decision, regardless of whether a filler has been said or not. 700ms after the end of the initial turn, ERICA will generate a response. If a filler is being said then ERICA will wait until the filler has been completed before speaking.

For simplification, we make some assumptions about the simulation. We ignore continuous model processing and response generation time and also assume the conversation continues as in the corpus. There is no guarantee that this will occur in a real situation, particularly for wrong samples. We also assume that our IPU-based model can always detect the end of the turn and therefore take it immediately, since we need to determine if we can improve a perfect IPU model.

Experimental setup

We have samples from one of two scenarios (correct and wrong) and one of three conditions (**IPU**, **Gaze** and **G+F**). We evaluate each combination of these, resulting in six different combinations which represent a scenario and evaluation pair.

Participants watched pairs of videos representing one of the six different combinations. The videos showed ERICA facing the camera at her eye level, representing the viewpoint of the other speaker. Figure 5 shows a video screen shot.

Figure 5: Screen shot of sample video used in the experiment.

The voice of the other speaker is played and ERICA responds according to the turn-taking model used as the condition. Each video in a pair had the same content but used different turn-taking models. Participants evaluated each pair by answering three questions:

Table 3: Fisher's exact test for distributions of correct and wrong samples.

Condition and measure	p-value
IPU vs. Gaze - Timing	0.902
IPU vs. G+F - Timing	**0.003**
Gaze vs. G+F - Timing	**0.000**
IPU vs. Gaze - Interest	1.000
IPU vs. G+F - Interest	0.689
Gaze vs. G+F - Interest	0.808
IPU vs. Gaze - Human-like	0.612
IPU vs. G+F - Human-like	0.119
Gaze vs. G+F - Human-like	**0.005**

- In which video was the timing of ERICA's responses more appropriate?
- In which video did ERICA seem more interested in the conversation?
- In which video was ERICA more human-like?

We recruited 29 participants (19 male, all students) for this experiment. Participants were shown 30 video pairs (5 videos for each of the 6 scenario and evaluation combinations). Each pair of videos was generated in a random order and no video was evaluated more than once. Left and right positions of the videos on the screen were randomized.

We forced participants to make a choice between two conditions. We could have included a neutral option but we wanted them to consider the differences in conditions, even if they were small. We could also have used Likert scale measures, but we wanted to reduce the workload of the participants, who would be watching many videos.

7 EXPERIMENT RESULTS

We conducted a binomial test to assess each turn-taking model for both correct and wrong samples. Results are shown in Figure 6.

For the timing of responses, both the **Gaze** and **G+F** models outperformed the **IPU** model for correct samples. The **G+F** model is much more preferred than the **Gaze** model for correct samples, but the reverse is true for wrong samples. The **G+F** model was the best for showing ERICA's interest, outperforming the other two models over both scenarios. In terms of the most human-like responses, the **G+F** model also outperformed the others. For wrong samples, there was no difference between the **Gaze** and **G+F** models.

We also performed Fisher's exact tests to test whether there were differences between correct and wrong samples for each comparison. Results are shown in Table 3.

We found the only significant differences were in the timing and human-like measures, where the **G+F** condition differed from **IPU** and **Gaze**. In all other comparisons, the distributions of answers were the same.

We were interested in videos evaluated as high or low by most participants, particularly for wrong samples. The context of the conversation seemed important. In several wrong samples, the filler came after the speaker asked a question or introduced a topic then continued talking to elaborate on it. ERICA used a filler to show interest in the utterance and the speaker continued talking, which made the interruption seem quite natural. We also did not find evidence in wrong samples that the timing was better if the filler was produced closer to the actual end of turn.

The filler form could also have had some effect. One of the lowest rated samples was a job interview interaction, where the speaker asked ERICA to give a self-promotion. The model produced a notice filler (*ah*) as an output, but this is unnatural for ERICA's role as an interviewee. A proper filler or even no filler would have been more suitable.

8 ANALYSIS AND DISCUSSION

Overall, our proposed model was promising. We found that when the continuous model is correct, the system is improved if turn-taking cues are used. Building on previous work [3, 26] we showed that gaze aversion improves perceptions of the robot and also found that using both filler and gaze aversion is effective. We can justify the use of fast turn-taking using fillers. Even with incorrect predictions, the hybrid model made ERICA seem more interested and was more human-like than only producing a response as fast as possible. We have now implemented the hybrid model in ERICA and aim to evaluate it in a live setting.

Fisher tests showed that correctly identifying the end of the turn made no difference to perceived interest in the conversation. For the other measures, the **G+F** model is perceived as worse when it gets the continuous prediction wrong. In other words, using a filler is riskier than gaze aversion, but more effective if the timing is correct. We also observed that wrong samples which were perceived as having good timing were ones where ERICA said fillers in the middle of the turn and between the speaker's clauses. This effect made ERICA's speech seem natural even though we didn't specifically train the model for this.

We have shown that our model performs well for isolated turns, but do not know if this would be seen as useful in a multi-turn full conversation. There are still a large number of turns in which the model does not predict a TRP. Fillers should be used sparingly, so the weak recall of the continuous model may not be so problematic. However, gaze aversion occurs more frequently so we have to improve our model if we prefer to use this cue.

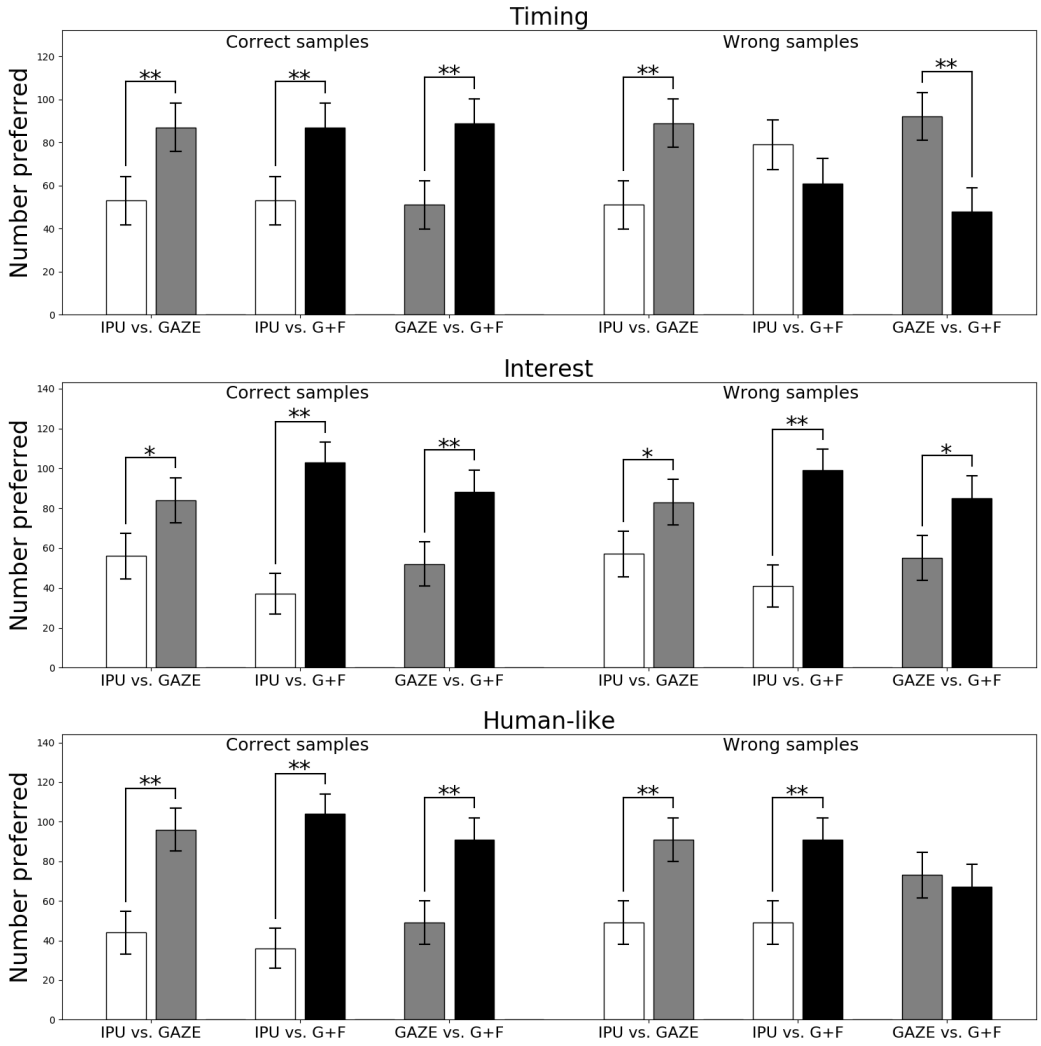

Figure 6: Results of subjective experiment. Significant p-values are indicated by * (<0.05) and ** (<0.01).

We made several assumptions in our experiment, as explained in Section 6. Firstly we assume the speaker's behavior would not change if they observed the turn-taking cues. In reality, the speaker may stop when they hear a filler in anticipation of ERICA taking the turn. Secondly we assume that our IPU-based model could always predict the actual end of turn and so ERICA would respond quickly. In reality, if the IPU-based model produces a false negative, then the response would be either delayed or not said. Another condition can be tested - the continuous model correctly predicts the end of the turn, but the IPU-based model misses the prediction. This results in a delay between the turn-taking cues and the response itself. Analysis of this condition needs to be confirmed in a similar experiment.

9 CONCLUSION

In this paper we describe a hybrid model of turn-taking, where a continuous model is used to generate turn-taking cues for an android robot. Once an automatic speech recognition result has been received, a separate IPU-based model predicts the end of the turn based on the utterance itself. A subjective experiment with this hybrid model showed that users preferred it over a perfect IPU-only model, with better perception of timing, interest in the conversation, and human-likeness. The next step in our work is to implement this in a live system and conduct a proper user evaluation.

ACKNOWLEDGMENTS

This work was supported by JST ERATO Grant Number JPMJER1401, Japan.

REFERENCES

[1] Abdulaziz Abubshait and Eva Wiese. 2017. You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in Psychology* 8 (2017), 1393.

[2] Sebastian Andersson, Kallirroi Georgila, David Traum, Matthew Aylett, and Robert AJ Clark. 2010. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Speech Prosody 2010 - Fifth International Conference*.

[3] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, USA, 25–32.

[4] Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. 2016. Erica: The ERATO intelligent conversational android. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 22–29.

[5] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers. In *Proceedings of INTERSPEECH 2018*. 991–995.

[6] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50, 6 (2008), 531–543.

[7] Yuichi Ishimoto, Takehiro Teraoka, and Mika Enomoto. 2017. End-of-Utterance Prediction by Prosodic Features and Phrase-Dependency Structure in Spontaneous Japanese Speech. In *Proceedings of INTERSPEECH 2017*. 1681–1685.

[8] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 12.

[9] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (1967), 22–63.

[10] Minae Kwon, Malte F Jung, and Ross A Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 463–464.

[11] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2018. Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In *Proceedings of the 2018 International Conference on Multimodal Interaction*. ACM, 78–86.

[12] Divesh Lala, Shizuka Nakamura, and Tatsuya Kawahara. 2019. Analysis of effect and timing of fillers in natural turn-taking. In *Proceedings of INTERSPEECH 2019*. (to appear).

[13] Stephen C Levinson. 2016. Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences* 20, 1 (2016), 6–14.

[14] Chaoran Liu, Carlos Ishi, and Hiroshi Ishiguro. 2017. Turn-Taking Estimation Model Based on Joint Embedding of Lexical and Prosodic Contents. In *Proceedings of INTERSPEECH 2017*. 1686–1690.

[15] Angelika Maier, Julian Hough, and David Schlangen. 2017. Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems. In *Proceedings of INTERSPEECH 2017*. 1676–1680.

[16] Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka. 2017. Online End-of-Turn Detection from Speech based on Stacked Time-Asynchronous Sequential Networks. In *Proceedings of INTERSPEECH 2017*. 1661–1665.

[17] Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Ryuichiro Higashinaka, and Yushi Aono. 2018. Neural Dialogue Context Online End-of-Turn Detection. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 224–228.

[18] Naoki Mukawa, Hiroki Sasaki, and Atsushi Kimura. 2014. How do verbal/bodily fillers ease embarrassing situations during silences in conversations?. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 30–35.

[19] Ryosuke Nakanishi, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot. In *IWSDS 2018*.

[20] David G Novick, Brian Hansen, and Karen Ward. 1996. Coordinating turn-taking with gaze. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, Vol. 3. IEEE, 1888–1891.

[21] Naoki Ohshima, Keita Kimijima, Junji Yamato, and Naoki Mukawa. 2015. A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 325–330.

[22] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale rnns. In *Proceedings of the 2018 International Conference on Multimodal Interaction*. ACM, 186–190.

[23] 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*. Elsevier, 7–55.

[24] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2009. How quickly should a communication robot respond? Delaying strategies and habituation effects. *International Journal of Social Robotics* 1, 2 (2009), 141–155.

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

[26] Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 67–74.

[27] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592.

[28] Sei Ueno, Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. 2018. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In *Proceedings of IEEE-ICASSP*. IEEE, 5804–5808.

[29] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13, 3 (2018), 55–75.

[30] Jakub Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F Glas, Christoph Bartneck, and Hiroshi Ishiguro. 2016. Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics* 7, 1 (2016).