Can Prediction of Turn-management Willingness Improve Turn-changing Modeling?

Ryo Ishii*
rishii@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Xutong Ren*
xutongr@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Michal Muszynski
mmuszyns@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Louis-Philippe Morency
morency@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

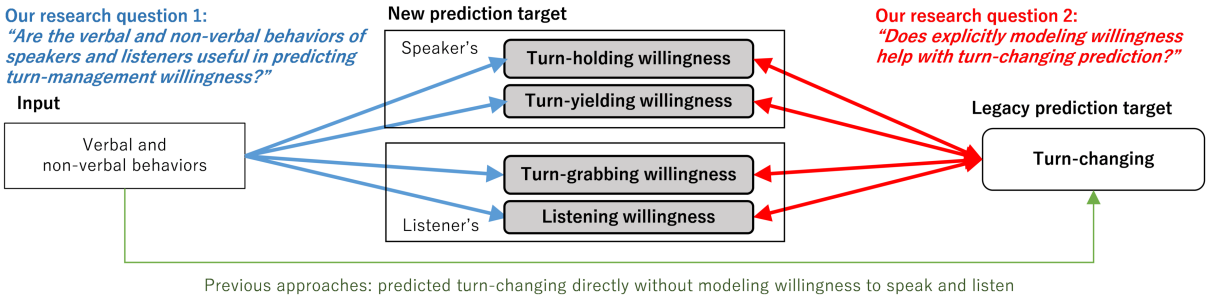

Figure 1: Overview of Our Research Questions

ABSTRACT

For smooth conversation, participants must carefully monitor the turn-management (*a.k.a.* speaking and listening) willingness of other conversational partners and adjust turn-changing behaviors accordingly. Many studies have focused on predicting the actual moments of speaker changes (*a.k.a.* turn-changing), but to the best of our knowledge, none of them explicitly modeled the turn-management willingness from both speakers and listeners in dyad interactions. We address the problem of building models for predicting this willingness of both. Our models are based on trimodal inputs, including acoustic, linguistic, and visual cues from conversations. We also study the impact of modeling willingness to help improve the task of turn-changing prediction. We introduce a dyadic conversation corpus with annotated scores of speaker/listener turn-management willingness. Our results show that using all of three modalities of speaker and listener is important for predicting turn-management willingness. Furthermore, explicitly adding willingness as a prediction task improves the performance of turn-changing prediction. Also, turn-management willingness prediction becomes more accurate with this multi-task learning approach.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative interaction**; **HCI theory, concepts and models**; **Collaborative and social computing theory, concepts and paradigms**.

KEYWORDS

turn-management willingness, turn-changing prediction, multi-task learning, multimodal signal processing

ACM Reference Format:
Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2020. Can Prediction of Turn-management Willingness Improve Turn-changing Modeling?. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423907

1 INTRODUCTION

Turn-changing is an important aspect of smooth conversation, where the roles of speaker and listener change during conversation. For smooth turn-changing, participants must carefully monitor the willingness of other conversational partners to speak and listen (a.k.a turn-management) and consider whether to speak or yield on the basis of their own willingness and that of other partners. Predicting turn-changing can be helpful to conversational agents

*Both authors contributed equally to this research.

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423907

or robots as they need to know when to speak and take turns at the appropriate time. The field of human-computer interaction has long been dedicated to computational modeling of turn-changing. Furthermore, many studies have focused on developing actual turn-changing (*i.e.*, next speaker or end-of-turn) models that can predict whether turn-keeping or turn-changing will happen using participants' verbal and non-verbal behaviors [3, 5, 6, 10, 12, 15, 16, 19–26, 30, 34–38, 43, 47, 50].

In this paper, we study turn-management willingness during dyadic interactions with the goal of incorporating the modeling of willingness into the computational model of turn-changing prediction (see Fig. 1). We study four types of willingness for speakers and listeners: turn-holding (a.k.a speaker's willingness to speak), turn-yielding (a.k.a speaker's willingness to listen), turn-grabbing (a.k.a listener's willingness to speak), and listening (a.k.a listener's willingness to listen). We focus on two new research questions:

Q1) Are the verbal and non-verbal behaviors of speakers and listeners useful in predicting turn-management willingness?

Q2) Does explicitly modeling willingness help with turn-changing prediction?

Firstly, we study the behavioral usefulness of features obtained from acoustic, linguistic, and visual modalities from both speakers and listeners. Predicting willingness directly could help conversational agents and robots with starting and ending utterances.

Secondly, we study prediction models for actual turn-changing. As a first step, we use trimodal inputs (acoustic, linguistic, and visual inputs) to directly predict turn-changing. As a second step, we integrate willingness prediction with turn-changing prediction. This integrated modeling approach is motivated by the intuition that humans are likely to control actual turn-changing on the basis of turn-management willingness. We build a multi-prediction model for turn-changing and willingness using a multi-task learning paradigm and evaluate the performance improvement.

2 RELATED WORK

2.1 Turn-changing Prediction Technology

Research on the mechanisms of yielding and taking conversation turns was initiated mainly in the field of sociolinguistics. Sacks *et al.* [46] proposed a turn-changing model, arguing that speaker switching occurs only at transition-related points (TRPs). Kendon [31] analyzed conversations and discovered that verbal and non-verbal behaviors contributed to smooth turn-keeping and turn-changing. Other studies have demonstrated verbal and non-verbal cues for a person to consider the presence or absence of turn-changing in two-person conversations [5, 35]. Several studies have recently examined that non-verbal cues of conversation partners are discriminative for turn-changing. It has been shown that eye-gaze behavior [15, 24, 26, 30], eye blinking [19], head movement [21, 22], respiration [25], and hand gestures [16] are related to turn-changing.

With such knowledge, many studies have developed models for predicting actual turn-changing, *i.e.*, whether turn-changing or turn-keeping will take place, on the basis of acoustic features [3, 6, 10, 12, 18, 26, 34, 36–38, 43, 47, 50], linguistic features [34, 37, 38, 43], and visual features, such as overall physical motion [3, 6, 8, 43] near the end of a speaker's utterances or during multiple utterances. Moreover, some research has focused on detailed non-verbal behaviors such as eye-gaze behavior [3, 6, 18, 20, 24, 26], head movement [18, 21, 22], mouth movement [23], and respiration [20, 25]. However, many turn-changing prediction studies use mainly features extracted from speakers. Several studies used limited features and modalities of listeners [20, 20–25, 38].

To the best of our knowledge, our paper is the first to study the prediction of turn-management willingness in dyad interaction and the first attempt to explicitly add the willingness prediction task to the turn-changing prediction model. Furthermore, there is no prior research that investigates all acoustic, linguistic, and visual modalities of speakers and listeners for turn-changing prediction. Our study is the first to construct a model for predicting willingness and turn-changing using trimodal information, including acoustic, linguistic, and visual cues of both speakers and listeners.

2.2 Human-Agent Interaction with Turn-changing Prediction

In the literature, researchers have mainly attempted to ensure smooth turn-changing, where the agent waits for its turn, which is not the rule in human-human conversations. For example, in [2, 47], algorithms were developed that predict turn-endings as soon as possible such that the system can behave immediate enough to simulate human-like behavior. In [42], how audio features are used to detect an end-of-turn as soon as possible was demonstrated; thus, an agent can start to speak as soon as possible. In human-agent interaction, an agent attempts to acquire a turn and start uttering at an appropriate time by using the prediction of a turn-changing prediction model. In [27, 28], a real-time turn-changing model was developed that was optimized to minimize the silence gap between the speech turn of a human and the system.

Also, using our estimation of turn-management willingness, agents may be able to facilitate users' speaking on the basis of the users' willingness. For example, although a listener may strongly want to take a turn, they may not actually do so (i.e. the speaker does not yield to him/her). At such times, the agent may be able to prompt the listener to start speaking using verbal and non-verbal behavior (the discrepancies between the turn-management willingness of speakers and listeners and actual turn-changing will be reported in Section 4.)

3 NEW MM-TMW CORPUS

3.1 Dialogue Collection

We collected a new corpus (named the "MM-TMW Corpus") that includes verbal and non-verbal behavioral information on human-human dialogue. It consists of 12 face-to-face conversations of people who had never met before (12 groups of 2 different people). The participants were 24 Japanese in their 20s to 50s (mean: 32.0, STD: 8.4). They were seated opposite each other. The conversations were structured to be about multiple topics, including taxes and social welfare balance. The lengths were unified to be around 10 minutes. The total time of all conversations was 120 minutes. The participants' voices were recorded by a headset microphone. The entire discussions were recorded by a camera. We also took upper body videos of each participant recorded at 30 Hz. A professional

transcribed all Japanese utterances, and another double-checked transcripts.

3.2 Annotation of Turn-management Willingness

As a first step, professional annotators identified the spoken utterance segments using the annotation scheme of the inter-pausal unit (IPU) [33]. Each start and end of an utterance was denoted as an IPU. When a silence interval of 200 ms or more occurred, the utterance was separated. Therefore, if an utterance was made after a silent period of less than 200 ms, it was determined to be a continuation of the same utterance. We excluded back-channels without specific vocal content from the extracted IPUs. Next, we considered IPU pairs by the same person in temporally adjacent IPU pairs as turn-keeping and those by different people as turn-changing. The total number of pairs was 2208 for turn-keeping and 631 for turn-changing.

We collected turn-management scores with multiple external observers using as reference an annotation method for multiple external observers [17]. The 10 annotators carefully watched each video from the beginning of one utterance (IPU) to the point just one frame (33 ms) before the beginning of the next utterance to annotate willingness scores. The annotators were not aware of who would become the next speaker because they could only watch the video until the point just before the start of the next speaker. This approach was taken to avoid affecting the annotators' judgement on the willingness of the speakers and listeners to speak and listen. For very short IPUs of less than one second, we set the start of the video to a moment earlier than the start time of the IPUs so that the annotators could view at least one second of video. In addition, the content of the current utterance and that of the past dialogue were considered to be important for judging turn-management willingness. Therefore, the annotators observed the utterances in order, starting with the first at the beginning of the dialogue. They could refer to contextual information on past dialogue to annotate the willingness score. The annotation order for the 12 dialogues was randomized for each annotator. For each video, they gave scores to four types of turn-management willingness of speakers and listeners.

- **Turn-holding willingness (a.k.a speaker's willingness to speak**: Does the speaker have the will to hold the turn (continue speaking)?
- **Turn-yielding willingness (a.k.a speaker's willingness to listen**: Does the speaker have the will to yield the turn (listen to listener speak)?
- **Turn-grabbing willingness (a.k.a listener's willingness to speak)**: Does the listener have the will to grab the turn (start speaking)?
- **Listening willingness (a.k.a listener's willingness to listen)**: Does the listener have the will to continue listening to the speaker speak?

The annotators scored each willingness index on a 5-point Likert scale, where 1 meant "He/she is not showing willingness," 5 meant "He/she is showing strong willingness," and 3 meant "uncertain." We had 10 annotators score all videos to ensure good reliability. We calculated the rater agreement using the Intraclass Correlation Coefficient (ICC). The ICC scores for all four categories were over **0.870**: $ICC(2,10) = \mathbf{0.904}$ for speaker's willingness to speak, $ICC(2,10) = \mathbf{0.877}$ for speaker's willingness to listen, $ICC(2,10) = \mathbf{0.878}$ for listener's willingness to speak, and $ICC(2,10) = \mathbf{0.875}$ for listener's willingness to listen. This suggests that the data was very reliable. We used the average values of the 10 annotators as willingness scores.

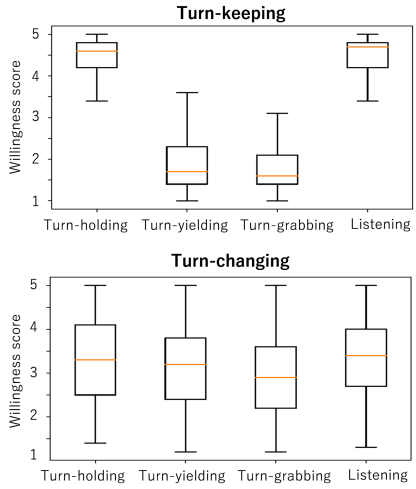

Figure 2: Box plots of turn-management willingness scores in turn-keeping (top) and turn-changing (bottom).

4 ANALYSIS OF WILLINGNESS IN TURN-KEEPING/CHANGING

In this section, we analyze the relationship between willingness scores and the actual turn-changing or turn-keeping as an empirical study. Figure 2 shows box plots of each score in our corpus, separated between turn-keeping and turn-changing to investigate the overall relationship between them. When turn-keeping happened, the average scores of the speaker's turn-holding willingness and listener's willingness to listen were more than 4.5, which was very high. In contrast, those for the speaker's turn-yielding and listener's turn-grabbing willingness were less than 2.0, which was very low. This means that a current speaker who continues to speak always has a high turn-holding willingness and a listener who continues to listen is highly willing to listen in turn-keeping. This suggests that a person's turn-management willingness and actual next speaking behavior are always consistent in turn-keeping.

When turn-changing happened, all average willingness scores were from 3.0 to 3.5, with larger standard deviations. This suggests that the current listener who becomes the next speaker may not always have a high turn-grabbing willingness and the current speaker who becomes the listener next may not always have high a turn-yielding willingness. We explored the discrepancies between the turn-management willingness of speakers and listeners and actual turn-changing, where turn-changing happens even though scores of turn-yielding and turn-grabbing willingness are not high. In detail, we calculated the occurrence probability of the discrepancies where the scores of turn-holding willingness were higher than those of turn-yielding willingness or those of willingness

to listen were higher than those of turn-grabbing willingness in turn-changing. As a result, the discrepancies were **44.8**% in turn-changing. This means that the willingness scores sometimes had discrepancies with actual turn-changing. The accuracy could be further improved by performing multi-task learning on willingness and turn-changing since they have a strong relationship [44]. Therefore, simultaneously predicting turn-management willingness could improve turn-changing prediction.

The result raises the possibility that willingness prediction could be beneficial for realizing an agent with smooth turn-management according to the discrepancies between willingness and actual turn-changing. For example, the agent may be able to prompt the listener to take a turn and start speaking using verbal and non-verbal behavior.

5 TURN-MANAGEMENT WILLINGNESS AND TURN-CHANGING PREDICTION MODELS

5.1 Motivation

To address **Q1**, we implemented three kinds of models for predicting turn-management willingness using the multimodal behaviors of either speaker or listener or both of them. To address **Q2**, we also implemented models for predicting turn-changing that jointly predict turn-management willingness on the basis of single turn-management prediction models.

5.2 Multimodal Features

We used the feature values of behaviors extracted during IPUs (i.e., the time between the start and end of an IPU) as input for the prediction models the same as other research on turn-changing prediction [3, 5, 6, 10, 15, 16, 19, 26, 30, 34–36, 38, 47]. This means that our models could predict willingness and turn-changing at the end of a speaker's utterance (IPU). Since the duration between the end of ones speaker's utterance and the start of the next speaker's utterance is about 620 ms on average, our models could predict willingness and turn-changing about 620 ms before actual turn-keeping and turn-changing happens.

Our goal is not necessarily to implement the most complex multimodal fusion but we aim to study willingness and its impact on turn-changing precision. Recently, high-level abstracted features have been very useful for many various prediction tasks. For example, in one of the most recent pieces of research [49], a model was implemented that estimates self-disclosure utterances using multimodal features of acoustic, linguistic, and visual modalities while utterances take place. It demonstrated that the latest high-level abstracted features, such as those of VGGish [14], BERT [7], and ResNet-50 [13], are more useful than interpretable features, such as those of MFCC [9], LIWC [29], and action unit [1], for estimating self-disclosure utterances in dyad interactions. To implement willingness prediction models, we used automatically extracted high-level features from the recorded data of the acoustic, linguistic and visual modalities on the basis of an existing study [49].

Acoustic Modality. We used VGGish [14], which is a deep convolutional neural network, to extract features of the acoustic modality from audio data. VGGish is a variant of the VGG model [48], trained on a large YouTube dataset to classify an ontology of 632 different audio event categories [11], involving human sounds, animal sounds, natural sounds, etc. The audio files were converted into stabilized log-mel spectrograms and fed into the VGG model to perform audio classification. The output 128-dimensional embeddings were post-processed by applying a PCA transformation (which performs both PCA and whitening). Therefore, each audio sample was encoded as a feature with a shape of $T \times 128$, where T is the number of frames. During natural conversations, listeners are not always absolutely silent; there are short backchannel responses or echoes of what speakers have said. Therefore, the VGGish features could be extracted from listeners' acoustic signals in addition to speakers' acoustic signals.

Linguistic Modality. We applied a data-driven method (BERT) [7] to extract linguistic representations. BERT is a multi-layer bidirectional Transformer network that encodes a linguistic sequence into a fixed-length representation. We used a pre-trained BERT model on Japanese Wikipedia[1] to transfer each utterance into a 768-dimensional feature. The BERT feature could be extracted from listeners' speech in addition to speakers' speech similarly to acoustic features since listeners often have short backchannel responses.

Visual Modality. For visual information, high-level representations were extracted using ResNet-50 [13], which is a deep residual convolutional neural network for image classification. We used a ResNet-50 model that was trained on ILSVRC2012 [45], a large scale dataset that contains about 1.2 million training samples in 1000 categories, to provide good generalization and yield robust features. The feature vector for a video sequence consisted of a 2048-dimensional vector obtained from the penultimate layer for each frame. As a result, the extracted feature was in the shape of $T \times 2048$.

5.3 Prediction Models

Turn-management willingness and turn-changing were first predicted individually using regression models (for predicting turn-management willingness scores) and classification models (for turn-changing/keeping prediction). A multi-task model was then learned to jointly predict willingness and turn-changing/keeping. This will help to understand the impact of modeling willingness explicitly. Our architecture for the multi-task model is illustrated in Figure 3.

Turn-management willingness prediction. We formulated the turn-management willingness prediction as a regression problem and average willingness scores from the 10 annotators as the ground truth. We used the neural networks to learn our regression problem. The unimodal features were first fed into individual processing modules to be further processed as 64-dimensional embeddings. For acoustic and visual modalities, the processing module was a one hidden layer gated recurrent unit (GRU) [4]. A fully connected (FC) layer as used for the linguistic modality. The embeddings were then concatenated together and forwarded into a FC layer with an output size of 192 for fusion. A final linear layer followed, outputting four predicted willingness scores. We used mean squared error (MSE) as our loss function.

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB

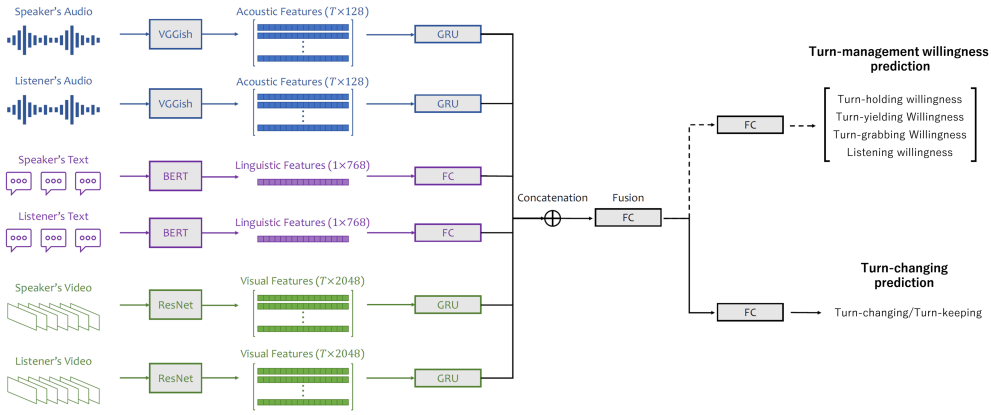

Figure 3: Architecture of Multi-task Model with Input Features of Acoustic, Linguistic, and Visual Modalities from Speaker and Listener.

Turn-changing prediction. Turn-changing prediction was considered a classification problem. Each turn was labeled as either turn-changing or turn-keeping, depending on whether the current listener became the next actual speaker. The classification model followed the same structure as the regression one, except that it output a two-dimensional vector for prediction. Cross entropy (CE) was used as the loss function.

Multi-task prediction. To embed willingness knowledge into turn prediction, our proposed multi-task model jointly predicts willingness scores and turn-changing/keeping. The model follows the main structure discussed above, with the difference being that, after the fusion layer, it has an FC layer for each task. The entire loss function is a weighted average of MSE and CE with weights of 1 and 2.

6 EXPERIMENTS

6.1 Experimental Methodology

To answer question **Q1**, we implemented the three kinds of models of turn-management willingness prediction using the multimodal behaviors of either the speaker or listener or both. We investigated and compared the performance of the models to demonstrate that turn-management willingness can be predicted using multimodal behaviors of speakers and listeners. To answer question **Q2**, we also implemented the models of turn-changing prediction that jointly predict turn-management willingness and turn-changing. We compared the performance of the multi-task learning models and single-task models to demonstrate that incorporating willingness into turn-changing prediction models improve turn-changing prediction.

All models were trained using the Adam [32] optimizer with a learning rate of 0.0001 for 50 epochs. The batch size was 64. Furthermore, we added dropout layers with a rate of 0.1 for the FC layers. Leave-one-dyad-out testing (12-hold cross-validation method) was used to evaluate model performance. With the testing, we evaluated how much willingness and turn-changing of new dyads can be predicted.

For the willingness prediction task, we report the concordance correlation coefficients (CCCs) between predicted and actual scores (i.e., annotated ground truth). A high CCC value indicates high agreement between the values of the predicted scores and ground truth. This means that prediction and ground truth values are similar to each other, and general trend changes for both signals are the same [40]. We compared the predictions of pairs of regression models by means of two-sided Wilcoxon signed rank tests at a 0.05 significance level [51]. For the classification task, we evaluated the performance using F1 scores weighted by the label proportion since the numbers of turn-changing and turn-keeping labels were imbalanced in our dataset. The predictions of pairs of classifiers were made by means of a McNemar test at a 0.05 significance level [39].

6.2 Results

Models were built using combinations of different input features. The results of willingness and turn prediction are shown in Table 1. Model (1) is the base model of prediction. It was a random prediction model that randomly generates scores and classes from learning data without using the feature values of speakers and listeners. The CCCs of the willingness prediction for model (1) were -0.011 for turn-holding, -0.013 for turn-yielding, -0.025 for turn-grabbing, and 0.007 for listening. The F1 score of turn-changing prediction was 0.528. All models (2) ~ (7) for turn-management willingness and turn-changing prediction tasks significantly outperformed model (1) (p-value < 0.001). This suggests that feature values from speaker and listeners are useful for prediction.

Results of turn-management willingness prediction using speaker/listener behaviors (related to Q1). As shown in Table 1, models (2), (3), and (4) used feature values of speaker, listener, and both independently.

Comparing models (2) and (3), the CCCs of turn-holding and turn-yielding prediction for model (2), 0.433 and 0.379, were significantly higher than those of model (3), 0.310 and 0.292 (p-value < 0.001). In contrast, the CCC of turn-grabbing prediction for model (3), 0.403, was significantly higher than that of model (2), 0.272 (p-value < 0.001). These suggest that speaker/listener feature values are more useful for predicting speaker/listener turn-management willingness than listener/speaker willingness.

Comparing model (4) with (2) and (3), model (4) with all features performed best, 0.502 for turn-holding, 0.464 for turn-yielding, 0.521

Table 1: Results of turn-management willingness and turn-changing prediction. Each row represents results of model with different configuration of input features. Section 6 describes experiments in detail. CCC is reported for each model for turn-management willingness prediction. F1 score is reported for turn-changing prediction. Results of runnning two-sided Wilcoxon signed rank among models (2) ~ (4) and among (5) ~ (7) are shown. Results for three pairs of two conditions under (2) vs (5), (3) vs (6), and (4) vs (7) are shown. * stands for p-value < 0.05, while ** stands for p-value $\ll 0.001$.

Model #	Features		Multi -task learning	Willingness Prediction (CCC)				Turn-changing Prediction (F1 score)
				Speaker		Listener		
	Speaker	Listener		Turn-holding	Turn-yielding	Turn-grabbing	Listening	
(1)				-0.011	-0.013	-0.025	0.007	0.528
(2)	×			0.443 $^{(3)**}$	0.379 $^{(3)**}$	0.272	0.327	0.759 $^{(3)**}$
(3)		×		0.310	0.292	0.403 $^{(2)**}$	0.373	0.711
(4)	×	×		**0.502** $^{(2)**,(3)**}$	**0.464** $^{(2)**,(3)**}$	**0.521** $^{(2)**,(3)**}$	**0.492** $^{(2)**,(3)**}$	**0.771** $^{(2)**,(3)**}$
(5)	×		×	0.433	0.381 $^{(2)**}$	0.272	0.321	0.760
(6)		×	×	0.320 $^{(3)**}$	0.303 $^{(3)**}$	0.422 $^{(5)**}$	0.400 $^{(3)**}$	0.730 $^{(3)**}$
(7)	×	×	×	0.534 $^{(5)**,(6)**}$	0.497 $^{(5)**,(6)**}$	0.517 $^{(5)**,(6)**}$	0.503 $^{(5)*,(6)*}$	**0.797** $^{(4)**,(5)**,(6)**}$

for turn-grabbing, and 0.492 for listening, being significantly than models with speaker's feature values (2) or listener's feature values (3) (p-value < 0.001). This suggests that a model using feature values from both speakers and listeners outperforms a model using them from one person. We found an overall improvement in turn-management willingness prediction by fusing multiple features of speaker and listener.

Results of turn-changing prediction using speaker/listener behaviors. We implemented and evaluated the performance of turn-changing prediction models (2), (3), and (4) similarly to the turn-management prediction models to assess the effect of multi-task learning on turn-changing prediction. We report the performance of the models to confirm whether our extracted speaker and listener features were useful for turn-changing prediction.

Comparing models (2) and (3), the F1 score of turn-changing prediction for model (2), 0.759, was significantly higher than that of model (3), 0.711 (p-value < 0.001). This suggests that the speaker features are more useful for predicting turn-changing than those of listeners.

Comparing model (4) with (2) and (3), model (4) with all features performed best, 0.771, significantly better than models with speaker features (2) or listener features (3) (p-value < 0.001). This suggests that a model using features from both speaker and listener outperforms using features from one person. We found an overall improvement in turn-changing prediction by fusing multiple speaker and listener features. These results are in line with previous research that similarly used both speaker and listener behaviors for turn-changing prediction [20, 23–25, 30].

The performance of our turn-changing prediction models was high [i.e., 0.771 for model (4)] even though the prediction task is known to be difficult and our dataset is relatively small. As an alternative, features from a pre-training model such as VGGish, BERT, and ResNet-50 could be used to mitigate our relative small dataset. Turn-changing prediction models (2) ~ (4) can serve as a baseline for evaluating the effect of using multi-task learning.

Results of multi-task prediction of turn-management willingness and turn-changing (related to Q2). We first analyzed whether applying multi-task learning to turn-management willingness and turn-changing prediction can improve turn-changing prediction. Models (5), (6), and (7) used multitask-learning in addition to models (2), (3), and (4), independently. We compared the performance between models (2) and (5), (3) and (6), and (4) and (7) for turn-changing prediction. Model (6) had a significantly higher F1 score, 0.730, than model (3), 0.711. Model (7) also had a significantly higher F1 score, 0.797, than model (4), 0.771 (p-value < 0.001). This suggests that multi-task learning incorporating turn-management willingness prediction into turn-changing prediction models improves the performance of turn-changing prediction.

We compared the performance among models (5) ~ (7), which used multi-task learning for turn-changing prediction. Model (7) with all features performed best, 0.797, being significantly better than models with speaker feature values (5) or listener feature values (6) (p-value < 0.001). This suggests that multimodal fusion using speaker and listener behaviors and multi-task learning incorporating turn-management willingness prediction were most useful for turn-changing prediction in our experiments.

We also analyzed whether multi-task learning is useful for predicting turn-management willingness. We compared the performance between models (2) and (5), (3) and (6), and (4) and (7). Model (6) only had significantly higher CCCs, 0.320 for turn-holding, 0.303 for turn-yielding, and 0.400 for listening, than model (3), 0.310 for turn-holding, 0.292 for turn-yielding, and 0.373 for listening (p-value < 0.001). This suggests that multi-task learning improved the performance of turn-management willingness prediction only when using the listener features.

We compared the performance of models (5) ~ (7), which use multi-task learning incorporating turn-management willingness prediction. Model (7) with all features performed best, 0.534 for turn-holding, 0.497 for turn-yielding, 0.517 for turn-yielding, and 0.503 for listening, being significantly higher than models with speaker feature values (5) or listener ones (6) (p-value < 0.001). These results suggest that multi-modal fusion using speaker and listener behaviors and multi-task learning applied to turn-management willingness prediction and turn-changing prediction are also useful for turn-changing prediction.

7 DISCUSSION

7.1 Relationship between Turn-management Willingness and Actual Turn-changing

In Section 4, we observed discrepancies between the willingness score and actual next speaking in turn-changing. We hypothesized that estimating willingness may be a helpful prediction target for avoiding such discrepancies. This is in contrast to prior works that ignored willingness information. For conversational agents or robots to start or stop speaking at the right time, we do believe that predicting human turn-management willingness is important, rather than simply predicting the next speaker (actual turn-changing). In this study, we tried to predict the willingness of two people simultaneously during dyad interaction. When applied to human-agent interaction (HAI) scenario, our approach will need to the adapted to predict only one user's willingness using the trimodal feature values, either the speaker or listener role. We see this as a great future direction.

Modeling turn-management willingness may help to detect discrepancies between the willingness toward turn-changing and actual turn-changing. A conversational system can then recognize users having a high willingness to speak (speaker's turn-holding or listener's turn-grabbing willingness) even though they cannot speak. It could even help to mediate meetings by possibly interrupting the current speaker if a person does not notice that the conversation partner has a low willingness to listen. Many studies are conducted to facilitate human interactions with agents and robots. For example, robots have been proposed that prompt the user who has the least dominance in conversation [41]. With such facilitation, the appropriate time when an agent can prompt a user to speak could be recognized with our prediction results on turn-management willingness and turn-changing.

7.2 Answer to Q1 Research Question

Our results show that the features of both speaker and listener are useful for predicting turn-management willingness. Individual turn-management willingness can be predicted better using features from individuals than from others. Individual willingness is well reflected in an individual's behavior. Moreover, the models using features of both speaker and listener performed better than those using only speaker or listener features. This suggests that the multimodal approach with trimodal features of speaker and listener is most useful in predicting the turn-management willingness of both persons. In the other words, the turn-management willingness of a speaker and listener can influence the verbal and non-verbal behaviors of both. This suggests that predicting the internal state of an individual, such as willingness, using features from not only the individual but also conversational partners could be greatly useful in dyad interaction.

7.3 Answer to Q2 Research Question

Turn-changing prediction becomes most accurate when turn-man agement willingness and turn-changing are predicted simultaneously using multi-task learning. This demonstrates that explicitly adding willingness as a prediction target improves the performance of turn-changing prediction. This introduces new possibilities for more accurately predicting human behavior by predicting human psychological states at the same time in conversations. Moreover, models that jointly learn two tasks also improve the performance of turn-management willingness compared with models that perform just one task. Multi-task learning leads a model to learn the underlying relationship between willingness scores and turn-changing. This results in both improved turn-changing and turn-management willingness prediction. These results also suggest that a multi-task prediction approach that predicts the internal state of people, such as their willingness and actual behaviors, could be greatly useful in dyad interaction. Applying such an approach to tasks other than turn-changing prediction will be part of our further investigation.

7.4 Future Work

Our goal is to study turn-management willingness and its impact on turn-changing precision. We used automatically high-level abstracted features extracted from acoustic, linguistic, and visual modalities. We plan to use other interpretable features, such as prosody [10, 15, 16, 19, 37, 38, 43] and gaze behavior [3, 20, 24, 26, 30], and implement more complex prediction models [37, 38, 43, 50] that take into account temporal dependencies.

Hara et al. [12] proposed a prediction model that can predict backchannels and filers in addition to turn-changing using multi-task learning. To analyze and model the relationship between turn-management willingness, backchannels and filers would be interesting future work.

We also plan to incorporate prediction models into conversational agent systems that can leverage the smooth turn-changing and facilitate the start of speaking for those who cannot speak despite having a high turn-holding or turn-grabbing willingness.

8 CONCLUSION

We found that many turn-changes happen even when the speaker has a high turn-holding willingness to continue speaking and the listener has a low turn-grabbing willingness to continue listening. This means that there are discrepancies between willingness and actual speaking behavior (i.e., turn-changing). Conversational agents would perform smooth turn-changing and facilitate the user in speaking with prediction results of turn-management willingness and actual turn-changing. We built models for predicting the turn-management willingness of speakers and listeners as well as turn-changing with trimodal behaviors, acoustic, linguistic, and visual cues, in conversations. An evaluation of our models showed that turn-management willingness and turn-changing are predicted most precisely when all of the modalities from speaker and listener are used. Furthermore, turn-changing prediction becomes more accurate when turn-management willingness and turn-changing are predicted jointly using a multi-task learning. Turn-management willingness prediction also becomes more accurate with it. These results suggest that more accurate prediction models of human behaviors could be built by incorporating other predictions related to human psychological states.

ACKNOWLEDGMENTS

Ryo Ishii was supported by the NTT Corpration. Xutong Ren and Louis-Philippe Morency were partially supported by the National

Science Foundation (#1722822, #1734868). Michal Muszynski was supported by the Swiss National Science Foundation (#P2GEP2 _184518).

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications*. Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[2] Michaela Atterer, Timo Baumann, and David Schlangen. 2008. Towards Incremental End-of-Utterance Detection in Dialogue Systems. In *COLING*. 11–14.

[3] Lei Chen and Mary P. Harper. 2009. Multimodal floor Control Shift Detection. In *ICMI*. 15–22.

[4] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*. 1724–1734.

[5] Stephen C.Levinson. 2016. Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in cognitive sciences* 20 (2016), 6–14.

[6] Iwan de Kok and Dirk Heylen. 2009. Multimodal End-of-turn Prediction in Multi-party Meetings. In *ICMI*. 91–98.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[8] Alfred Dielmann, Giulia Garau, and Herve Bourlard. 2010. Floor Holder Detection and End of Speaker Turn Prediction in Meetings. In *INTERSPEECH*. 2306–2309.

[9] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor. In *ACM MM*. 835–838.

[10] Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. 2002. Is the Speaker Done Yet? Faster and More Accurate End-of-utterance Detection using Prosody in Human-computer Dialog. In *INTERSPEECH*, Vol. 3. 2061–2064.

[11] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-labeled Dataset for Audio Events. In *ICASSP*. 776–780.

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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *ICASSP*. 131–135.

[15] Judith Holler and Kobin H. Kendrick. 2015. Unaddressed Participants' Gaze in Multi-person Interaction: Optimizing Recipiency. *Frontiers in Psychology* 6 (2015), 515–535.

[16] Judith Holler, Kobin H. Kendrick, and Stephen C. Levinson. 2018. Processing language in face-to-face conversation: Questons with gestures get faster responses. *Psychonomic Bulletin Review* 6 (2018), 25.

[17] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. *AAMAS* 2, 1265–1272.

[18] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. A Multimodal End-of-Turn Prediction Model: Learning from Parasocial Consensus Sampling. In *AAMAS*.

[19] Paul Hömke, Judith Holler, and Stephen C. Levinson. 2017. Eye Blinking as Addressee Feedback in Face-To-Face Conversation. *Research on Language and Social Interaction* 50 (2017), 54–70.

[20] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Multimodal Fusion using Respiration and Gaze for Predicting Next Speaker in Multi-Party Meetings. In *ICMI*. 99–106.

[21] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting Next Speaker Using Head Movement in Multi-party Meetings. In *ICASSP*. 2319–2323.

[22] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2017. Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings. In *HAI*. 181–187.

[23] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2019. Prediction of Who Will Be Next Speaker and When Using Mouth-Opening Pattern in Multi-Party Conversation. *Multimodal Technologies and Interaction* 3, 4 (2019), 70.

[24] Ryo Ishii, Kauhiro Otsuka, Shiro Kumano, and Junji Yamamoto. 2016. Predicting of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM TiiS* 6, 1 (2016), 4.

[25] Ryo Ishii, Kauhiro Otsuka, Shiro Kumano, and Junji Yamamoto. 2016. Using Respiration to Predict Who Will Speak Next and When in Multiparty Meetings. *ACM TiiS* 6, 2 (2016), 20.

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

[27] Gudny Ragna Jonsdottir and Kristinn R. Thórisson. 2009. Teaching Computers to Conduct Spoken Interviews: Breaking the Realtime Barrier with Learning. In *IVA*. 446–459.

[28] Gudny Ragna Jonsdottir, Kristinn R. Thorisson, and Eric Nivel. 2008. Learning Smooth, Human-Like Turntaking in Realtime Dialogue. In *IVA*. 162–175.

[29] Jeffrey Kahn, Renée Tobin, Audra Massey, and Jennifer Anderson. 2007. Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *J. psychology* 120 (02 2007), 263–86.

[30] Tatsuya Kawahara, Takuma Iwatate, and Katsuya Takanashii. 2012. Prediction of Turn-taking by Combining Prosodic and Eye-gaze Information in Poster Conversations. In *INTERSPEECH*. 726–729.

[31] Adam Kendon. 1967. Some Functions of Gaze Direction in Social Interaction. *ActaPsychologica* 26 (1967), 22–63.

[32] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*. 13.

[33] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs. In *Language and Speech*, Vol. 41. 295–321.

[34] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2018. Evaluation of Real-Time Deep Learning Turn-Taking Models for Multiple Dialogue Scenarios. In *ICMI*. 78–86.

[35] Imme Lammertink, Marisa Casillas, Titia Benders, Brechtje Post, and Paula Fikkert. 2015. Dutch and English Toddlers' Use of Linguistic Cues in Predicting Upcoming Turn Transitions. *Frontiers in Psychology* (2015), 6.

[36] Kornel Laskowski, Jens Edlund, and Mattias Heldner. 2011. A single-port non-parametric model of turn-taking in multi-party conversation. In *ICASSP*. 5600–5603.

[37] Ryo Masumura, Mana Ihori, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Takanobu Oba, and Ryuichiro Higashinaka. 2019. Improving Speech-Based End-of-Turn Detection Via Cross-Modal Representation Learning with Punctuated Text Data. *ASRU* (2019), 1062–1069.

[38] Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Ryuichiro Higashinaka, and Yushi Aono. 2018. Neural Dialogue Context Online End-of-Turn Detection. In *SIGdial*. 224–228.

[39] Quinn McNemar. 1947. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* 12, 2 (1947), 153–157.

[40] Michal Muszynski, Leimin Tian, Catherine Lai, Johanna Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. 2019. Recognizing induced emotions of movie audiences from multimodal information. *Trans. Affective Computing* (2019).

[41] Yukiko I. Nakano, Takashi Yoshino, Misato Yatsushiro, and Yutaka Takase. 2015. Generating Robot Gaze on the Basis of Participation Roles and Dominance Estimation in Multiparty Interaction. *ACM TiiS* 5, 4 (2015), 23.

[42] Antoine Raux and Maxine Eskenazi. 2008. Optimizing Endpointing Thresholds Using Dialogue Features in a Spoken Dialogue System. In *SIGdial*. 1–10.

[43] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. In *ICMI*. 186–190.

[44] Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. *CoRR* abs/1706.05098 (2017).

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 115, 3 (2015), 211–252.

[46] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organisation of turn taking for conversation. *Language* 50 (1974), 696–735.

[47] David Schlangen. 2006. From Reaction to Prediction: Experiments with Computational Models of Turn-taking. In *INTERSPEECH*. 17–21.

[48] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[49] Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal Analysis and Estimation of Intimate Self-Disclosure. In *ICMI*. 59–68.

[50] Nigel Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. 2018. Turn-Taking Predictions across Languages and Genres Using an LSTM Recurrent Neural Network. In *SLT*. 831–837.

[51] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83.