Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel 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

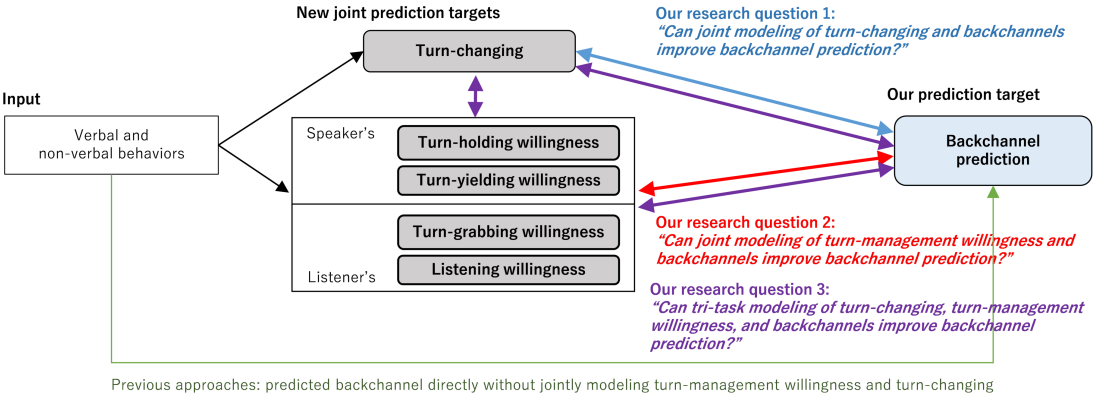

Figure 1: Overview of our research questions

ABSTRACT

The listener's backchannel has the important function of encouraging a current speaker to hold their turn and continue to speak, which enables smooth conversation. The listener monitors the speaker's turn-management (*a.k.a.* speaking and listening) willingness and his/her own willingness to display backchannel behavior. Many studies have focused on predicting the appropriate timing of the backchannel so that conversational agents can display backchannel behavior in response to a user who is speaking. To the best of our knowledge, none of them added the prediction of turn-changing and participants' turn-management willingness to the backchannel prediction model in dyad interactions. In this paper, we proposed a novel backchannel prediction model that can jointly predict turn-changing and turn-management willingness. We investigated the impact of modeling turn-changing and willingness to improve backchannel prediction. Our proposed model is based on trimodal inputs, that is, acoustic, linguistic, and visual cues from conversations. Our results suggest that adding turn-management willingness as a prediction task improves the performance of backchannel prediction within the multi-modal multi-task learning approach, while adding turn-changing prediction is not useful for improving the performance of backchannel prediction.

*The authors contributed equally to this research.

https://doi.org/10.1145/3472306.3478360

CCS CONCEPTS

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

KEYWORDS

backchannel, turn-management willingness, turn-changing, multitask learning, multimodal signal processing

ACM Reference Format:
Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling?. In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472306.3478360

1 INTRODUCTION

For smooth conversation, the listener's backchannel has the important function of encouraging the current speaker to hold their turn and continue to speak. To display the appropriate backchannel behavior (or start speaking instead), conversation participants must carefully monitor the willingness of other conversational partners to speak and listen (a.k.a. turn-management) and consider whether or not to speak or yield (i.e., displaying backchannel behavior) on the basis of their own willingness and other partners. Predicting a listener's backchannel can be beneficial for building conversational agents or robots as they are supposed to know when to display backchannel behavior at the appropriate time. In the field of human-agent/robot interaction, many studies have been dedicated to computational modeling of backchannels to add new dimensions to responses of agents and robots in interactions with humans. Many studies have focused on developing backchannel prediction models that can predict whether or not a backchannel occurs by using mainly acoustic and linguistic features, rarely the visual features of speakers [2, 11, 13, 18, 32, 33, 39, 40, 42, 47, 48, 50].

There is a need to explore novel multi-task prediction models for backchannels that incorporate multi-modal behaviors of speakers and listeners to enhance human-agent/robot interactions in many applications. Most previous work on backchannels has mainly investigated single-task learning approaches. Nevertheless, backchannels co-occur with other drivers of smooth conversations, such as turn-changing and turn-management willingness [27]. The relationship between the backchannel, turn-changing, and turn-management willingness has not been fully uncovered yet. We expect that multi-task learning approaches could capture the dependencies between these three conversation drivers and thus increase the performance of backchannel prediction. As a result, it could be possible to enrich human interactions with agents and robots.

In this paper, we investigate turn-changing and participants' turn-management willingness during dyadic interactions with the goal of jointly modeling turn-changing and willingness prediction to improve utterance backchannel prediction (see Fig. 1). In particular, we focus on 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) defined by [27] (see Section 3.4 for detailed definition) to improve backchannel prediction. In this study, we formulate three main research questions:

Q1) Can joint modeling of turn-changing and backchannels improve backchannel prediction?
Q2) Can joint modeling of turn-management willingness and backchannels improve backchannel prediction?
Q3) Can tri-task modeling of turn-changing, turn-management willingness, and backchannels improve backchannel prediction?

To address these questions, we first extract trimodal inputs (acoustic, linguistic, and visual features) to directly predict an utterance backchannel such as "yeah," "uh-huh," "hmm," and "right." We second incorporate neither or both turn-changing prediction and turn-management willingness prediction into backchannel prediction. This integrated modeling approach is motivated by the intuition that humans are likely to control the listener's backchannel on the basis of turn-management willingness and turn-changing (i.e., grab a turn). The multi-task learning framework can be useful for taking into account dependencies among human behavior (e.g. turn-changing) and internal state (e.g. turn-management willingness) and for improving the overall performance of our prediction model during dyad interactions [27]. We build a multi-prediction model for turn-changing and willingness in addition to backchannels, using a multi-modal multi-task learning paradigm, to increase prediction performance.

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

2 RELATED WORK

2.1 Backchannel and Turn-changing Prediction

There are several studies on predicting backchannels by using only the prosodic features of the preceding utterance of a speaker, such as pitch and power [11, 13, 39, 40, 47, 48, 50], as well as linguistic features [18, 42]. Moreover, some studies have used visual features [39]. Similarly, some 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, 13, 19, 28, 34, 36–38, 41, 44, 49], linguistic features [34, 37, 38, 41], and visual features, such as overall physical motion [3, 6, 8, 41] 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, 19, 21, 25, 28], head movement [19, 22, 23], mouth movement [24], and respiration [21, 26]. However, many turn-changing prediction studies have also mainly used features extracted from speaker behaviors rather than listener behaviors. Only a few studies have investigated limited features and modalities of listeners [21–26, 38]. Ishii et al. [27] used trimodal features, such as acoustic, language, and visual cues extracted from both speakers and listeners, and demonstrated that these trimodal features are useful for improving the performance of turn-changing prediction. This result supports our approach of using trimodal features from both speakers and listeners for predicting backchannels.

2.2 Multi-task Learning Approach

As an attempt to use explicit multi-task learning for backchannel and turn-changing prediction, Hara et al. [13] proposed a prediction model that could predict backchannels and fillers in addition to turn-changing by using a multi-task learning approach. The authors extracted only acoustic features from a speaker's acoustic signals to train a multi-task prediction model. In their experiments, the performance in predicting the listener's backchannel did not improve significantly. Our research utilizes not only speaker acoustic information but also speaker and listener multi-modal information

to explore the effects of simultaneously predicting turn-changing and willingness.

There has been an attempt to use explicit multi-task learning for turn-changing and turn-management willingness prediction. We investigated the impact of modeling willingness to help improve the task of turn-changing prediction [27]. We demonstrated that explicitly adding willingness as a prediction task improves the performance of turn-changing prediction. The results support our novel idea of improving the backchannel prediction performance through the joint prediction of backchannel, turn-management willingness, and backchannel.

3 CORPUS

3.1 MM-TMW Corpus

To implement a backchannel prediction model, we used data from the "MM-TMW Corpus" [27], which includes verbal and non-verbal behavioral information from human dialogues. 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, ages 20-50 (mean: 32.0, STD: 8.4). They were seated opposite each other. The conversations were structured and covered multiple topics, including taxes and social welfare balance. The lengths were unified to be around 10 minutes. The total time of all conversations was around 120 minutes.

3.2 Turn-changing

Professionals transcribed all of the Japanese utterances and identified the spoken utterance segments by 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, when an utterance was produced after a silent period of less than 200 ms, it was determined to be a continuation of the same utterance. We excluded backchannels without specific vocal content from the extracted IPUs. Next, we considered IPU pairs for the same person in temporally adjacent IPU pairs as turn-keeping and those for different people as turn-changing. The total number of pairs was 2208 for turn-keeping and 631 for turn-changing.

3.3 Backchannel

We focus on the utterance backchannel produced by the listener during turn-keeping, which is often known to immediately occur after the end of an utterance. We extracted listeners' backchannels during turn-keeping for use in this study. We extracted listeners' utterance backchannels between the end of an utterance and 1000 ms afterward for each turn-keeping occurrence. As a result, the total number of instances in which a backchannel occurred was 269, while the number of instances in which a backchannel did not occur was 362 for turn-keeping.

3.4 Turn-management Willingness

Turn-management willingness scores were collected with multiple external observers using, as reference, an annotation method for multiple external observers [18]. The ten annotators carefully watched each conversation 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 applied to avoid affecting the annotators' judgement on the willingness of the speakers and listeners to speak and listen. 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 the 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 the ten annotators score all videos to ensure good reliability. We calculated the rater agreement by 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 ten annotators as willingness scores.

4 PREDICTION MODELS

4.1 Motivation

To address **Q1** and **Q2**, we implemented backchannel prediction models that can jointly predict either turn-changing or turn-man agement willingness by using a two-task learning approach. We compared the performance of the backchannel prediction models with a single-task prediction model for backchannels. We also implemented three kinds of models for each of the three types of prediction models by using the multi-modal behaviors of either the speaker, listener, or both. To address **Q3**, we implemented models for predicting backchannels that jointly predict both turn-changing and turn-management willingness. We compared the performance of the backchannel prediction models using a three-task learning approach with a prediction model that can predict the backchannel and either turn-changing or turn-management willingness using the two-task learning approach.

4.2 Multi-modal Features

We used 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 backchannel or turn-changing prediction [3, 5, 6, 10, 13, 16, 17, 20, 27, 28, 30, 34–36, 38, 44]. This means that our models could predict backchannel,

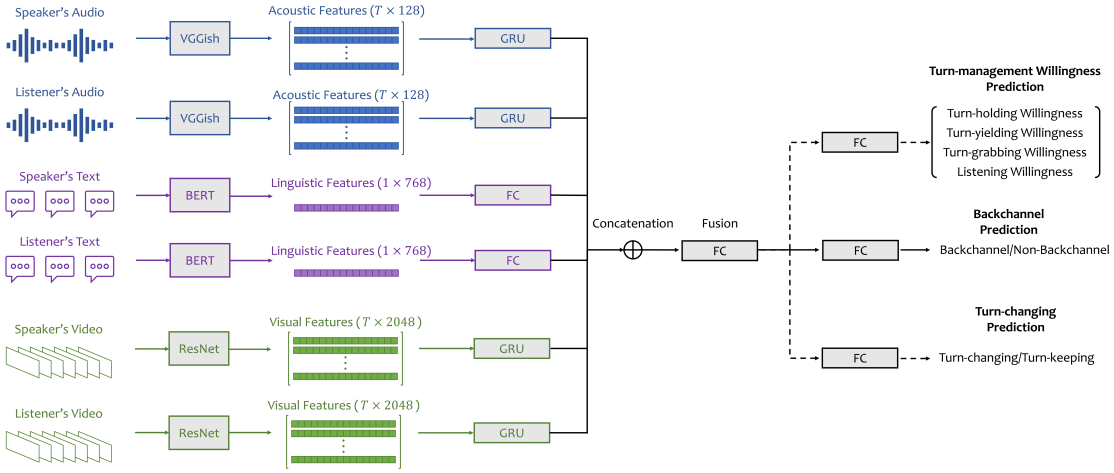

Figure 2: Architecture of multi-task prediction model for backchannels, turn-changing, and turn-management willingness with input features of acoustic, linguistic, and visual modalities from speaker and listener.

turn-management willingness, and turn-changing at the end of a speaker's utterance (IPU). Since the duration between the end of a speaker's utterance and the start of the next speaker's utterance is about 620 ms on average, our models could predict the three prediction targets about 620 ms before actual turn-keeping and turn-changing occurred.

In this study, we aimed to investigate the impact of turn-management willingness and turn-changing prediction on backchannel prediction rather than propose and implement a very complex multimodal fusion strategy to outperform existing backchannel prediction models. High-level abstracted features have recently been very informative for various prediction tasks. For example, in one of the most recent studies [46], a prediction model was proposed and implemented to estimate self-disclosure utterances on the basis of multi-modal features of acoustic, linguistic, and visual modalities while utterances take place. The study demonstrated that the latest high-level abstracted features, such as VGGish [15], BERT [7], and ResNet-50 [14], were more informative than interpretable features, such as MFCC [9], LIWC [29], and action unit [1], for estimating self-disclosure utterances in dyad interactions. Moreover, these features were very discriminative for predicting turn-changing and turn-management willingness in dyad interactions [27], which is similar to our study. To implement the 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 [27, 46].

Acoustic Modality. We used VGGish [15], which is a deep convolutional neural network, to extract features of the acoustic modality from the audio data. VGGish is a variant of the VGG model [45], trained on a large YouTube dataset to classify an ontology of 632 different audio event categories [12], 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 (that 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 time of a speaker's utterance in seconds.

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 transform each utterance into a 768-dimensional feature vector.

Visual Modality. For the visual modality, high-level representations were extracted by using ResNet-50 [14], which is a deep residual convolutional neural network for image classification. We used a ResNet-50 model that was trained on ILSVRC2012 [43], a large scale dataset that contains about 1.2 million training samples in 1000 categories, to provide good generalization and yield robust features. A feature set for a video sequence consisted of 2048-dimensional vectors obtained from the penultimate layer computed for each frame. As a result, the extracted features were in the shape of $T \times 2048$, where T is the number of frames during a speaker's utterance.

4.3 Implementation of Prediction Model

The backchannel was first predicted individually by using classification models (for backchannel/non-backchannel prediction). Then, turn-changing and turn-management willingness were jointly predicted in addition to backchannel prediction using classification models (for turn-changing/keeping prediction) and regression models (for predicting turn-management willingness scores). A multi-task model was then learned to jointly predict backchannels, turn-changing, and turn-management. This facilitated the understanding of the explicit impact of modeling turn-changing and turn-management willingness on backchannel prediction. Our architecture for a multi-modal multi-task model that can predict all three tasks for backchannels, turn-changing, and turn-management willingness is illustrated in Figure 2.

[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

Backchannel prediction. Backchannel prediction was considered a classification task. Either backchannel or non-backchannel was labeled depending on whether or not the listener performed an utterance backchannel behavior right after the end of the utterances collected in Section 3.3. 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 a loss function.

Turn-changing prediction. Turn-changing prediction was considered a classification task. Each turn was labeled as either turn-changing or turn-keeping, depending on whether or not the current listener became the next actual speaker, as described in Section 3.2. 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.

Turn-management willingness prediction. We formulated turn-management willingness prediction as a regression task and used the average willingness scores from the ten annotators as the ground truth described in Section 3.4. We learned a neural network-based model to address our regression task. 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 was 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 the four types of predicted willingness scores. It is worth mentioning that we selected mean squared error (MSE) as our loss function.

Multi-task prediction. Our proposed multi-task model jointly predicts backchannels, turn-changing, and turn-management willingness. The model is based on the main structure for predicting a single task, such as backchannel, turn-changing, or turn-management willingness prediction, explained above, with the difference being that there is an FC layer for each task after the fusion layer. The entire loss function is a weighted average of two MSEs and a CE with the same weights.

5 EXPERIMENTS

5.1 Experimental Methodology

To answer question **Q1**, we mainly implemented two types of backchannel prediction models: single-task learning prediction models and multi-task learning prediction models incorporating the turn-changing prediction task. We compared the performance of the multi-task learning models and single-task models to investigate whether incorporating turn-changing prediction into backchannel prediction models improves backchannel prediction. We also implemented three types of models for each model using the multi-modal behaviors of either the speaker, listener, or both to investigate the informativeness of features extracted from speakers and listeners.

To answer question **Q2**, we additionally implemented backchannel prediction models that jointly predict turn-management willingness. We compared the performance of the multi-task learning models and single-task models to demonstrate that incorporating turn-management willingness into backchannel prediction models improves backchannel prediction. In the same manner as question Q1, we also implemented three types of models using the multi-modal behaviors of either the speaker, listener, or both to investigate the usefulness of features from speakers and listeners.

To answer question **Q3**, we implemented backchannel prediction models that jointly predict turn-changing and turn-management willingness. We compared the performance of the multi-task learning models and other models mentioned above to demonstrate that incorporating turn-changing and turn-management willingness prediction into backchannel prediction models improves backchannel prediction. Moreover, we implemented three types of models using the multi-modal behaviors of either the speaker, listener, or both to investigate the usefulness of features from speakers and listeners.

All models were trained using the Adam [31] 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-fold cross-validation method) was used to evaluate model performance. With the testing, we evaluated how much backchannels, turn-changing, and turn-management willingness of new dyads could be predicted.

For classification tasks such as backchannel and turn-changin g prediction, we evaluated the performance using F1 scores. The predictions of pairs of classifiers were compared by means of a McNemar test at a 0.05 significance level. For a regression task such as turn-management willingness prediction, we report the concordance correlation coefficients (CCCs) between predicted and actual scores (i.e., annotated ground truth). A high CCC value indicates a high agreement between the values of the predicted scores and ground truth. This means that the prediction and ground truth values are similar to each other, and general trend changes for both signals are the same. We compared the predictions of pairs of regression models by means of two-sided Wilcoxon signed rank tests at a 0.05 significance level.

5.2 Results

The models were fed with combinations of different multi-modal features of speakers and listeners. The results of backchannel, turn-changing, and turn-management willingness prediction are summarized in Table 1. Model (0-B) is the base model of backchannel prediction. There was a random prediction model that randomly generated backchannel classes from learning data without using the feature values of speakers and listeners. The F1 score of model (0-B) was 0.557. Models (1) ~ (3) and (10) ~ (12) for the backchannel prediction task significantly outperformed model (0-B) (p-value < 0.001). This suggests that features extracted from speaker and listeners are informative for backchannel prediction. However, models (4) ~ (6) did not significantly outperform model (0-B) (p-value < 0.001).

Results of backchannel prediction using speaker/listener behaviors. As shown in Table 1, models (1), (2), and (3) used features of speakers, listeners, and both.

Comparing models (1) and (2), the F1 score of backchannel prediction for model (2), which was 0.752, was significantly higher than the 0.658 F1 score of model (1) (p-value < 0.001). This suggests that listener features are more useful for predicting listener backchannels than speaker ones.

Table 1: Results of backchannel, turn-changing, and turn-management willingness prediction. Each row represents results of model with different configuration of input features. Section 5 describes experiments in detail. F1 score is reported for backchannel and turn-changing prediction. CCC is reported for each model for turn-management willingness prediction. Spk and Lis note speaker and listener. B, T, and W note backchannel, turn-changing, and turn-management willingness prediction in column of multi-task learning. Significantly highest performance for each prediction task is shown in bold with start mark next to it.

Model #	Features		Multi -task Learn.	Backchannel Prediction (F1 score)	Turn-changing Prediction (F1 score)	Turn-management Willingness Prediction (CCC)			
						Spk		Lis	
	Spk	Lis				Turn-holding	Turn-yielding	Turn-grabbing	Listening
(0-B)			B	0.557±0.060	–	–	–	–	–
(0-T)			T	–	0.528±0.036	–	–	–	–
(0-W)			W	–	–	0.006±0.056	-0.017±0.054	-0.003±0.042	0.005±0.033
(1)	✓		B	0.658±0.062	–	–	–	–	–
(2)		✓	B	0.752±0.046	–	–	–	–	–
(3)	✓	✓	B	0.772±0.078	–	–	–	–	–
(4)	✓		B, T	0.344±0.065	0.592±0.045	–	–	–	–
(5)		✓	B, T	0.428±0.070	0.551±0.055	–	–	–	–
(6)	✓	✓	B, T	0.378±0.045	**0.617**±0.057*	–	–	–	–
(7)	✓		B, W	0.612±0.038	–	0.385±0.081	0.372±0.120	0.266±0.116	0.271±0.083
(8)		✓	B, W	0.742±0.043	–	0.194±0.102	0.212±0.084	0.342±0.121	0.261±0.103
(9)	✓	✓	B, W	**0.852**±0.081*	–	0.393±0.048	**0.410**±0.075*	0.447±0.120	0.372±0.082
(10)	✓		B, T, W	0.628±0.049	0.471±0.085	0.589±0.079	0.344±0.083	0.265±0.095	0.264±0.117
(11)		✓	B, T, W	0.749±0.070	0.368±0.090	0.561±0.048	0.224±0.069	0.361±0.122	0.259±0.095
(12)	✓	✓	B, T, W	**0.849**±0.077*	0.527±0.071	**0.642**±0.060*	**0.414** ±0.075*	**0.471**±0.081*	**0.406**±0.099*

Comparing model (3) with models (1) and (2), the F1 score of model (3) with all features, which was 0.772, was significantly higher than the performance of models with speaker features [model (1)] or listener features [model (2)] (p-value < 0.001). This suggests that a model using features extracted from both modalities of speakers and listeners will outperform a model using features from one person. We found an overall improvement in backchannel prediction by fusing multiple features of speakers and listeners.

Results of multi-task prediction of backchannels and turn-changing (related to Q1). We analyzed whether or not applying multi-task learning to backchannel and turn-changing prediction could improve backchannel prediction. Models (4), (5), and (6) contained a multi-task learning technique for turn-changing prediction in addition to backchannel single-task prediction models (1), (2), and (3). We compared the performance of models (4) and (1), models (5) and (2), and models (6) and (3) for backchannel prediction. Models (4), (5), and (6), which had F1 scores of 0.344, 0.428, and 0.378, respectively, had a significantly lower F1 score than models (1), (2), and (3). This suggests that multi-task learning incorporating turn-changing prediction into backchannel prediction models did not improve the performance of backchannel prediction.

Results of multi-task prediction of backchannels and turn-management willingness (related to Q2). We analyzed whether or not applying multi-task learning to backchannel and turn-management willingness prediction could improve backchannel prediction. Models (7), (8), and (9) used a multi-task learning technique for turn-management willingness prediction in addition to backchannel prediction models (1), (2), and (3) independently. We compared the performance between models (7), (8), and (9) and (1), (2), and (3) for backchannel prediction. Model (9), which had an F1 score of 0.852, had a significantly higher F1 score than model (3). However, models (7) and (8), which had F1 scores of 0.612 and 0.742, respectively, did not have higher F1 scores than models (1) and (2). This suggests that multi-task learning incorporating turn-management willingness prediction into backchannel prediction models can improve the performance of backchannel prediction when only using both the speaker's and listener's features.

Additionally, we compared the performance among models (7) ~ (9), with incorporated multi-task learning for backchannel prediction and turn-management willingness. Model (9) fed with all features performed best, and it achieved an F1 score of 0.852, being significantly higher than model (7) with speaker features or model (8) with listener features (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, taking into account the results of models (0) ~ (9).

Results of multi-task prediction of backchannels, turn-changing, and turn-management willingness (related to Q3). We analyzed whether or not applying multi-task learning to backchannel, turn-changing, and turn-management willingness prediction could improve backchannel prediction. We compared the performance of models (10) and (1), models (11) and (2), and models (12) and (3) for backchannel prediction independently. The results show that only model (12), which had a score of 0.849, had a significantly higher F1 score than model (3). Models (10) and (11) did not have a significantly higher F1 score than models (1) and (2). This suggests that applying multi-task learning to backchannel, turn-changing, and turn-management willingness prediction

could improve backchannel prediction when only using both the speaker's and listener's features.

Finally, we compared the performance of models (9) and (12), which used multi-task learning frameworks incorporating only turn-management willingness prediction and incorporating both turn-changing and turn-management willingness prediction. There was no significant difference in the performance of backchannel prediction between them. This suggests that there is no significant difference in the performance of backchannel prediction between incorporating only turn-management willingness prediction and incorporating both turn-changing and turn-management willingness prediction.

These results suggest that multi-modal fusion using speaker and listener behaviors and multi-task learning applied to turn-mana gement willingness prediction alone are useful for backchannel prediction.

6 DISCUSSION

6.1 Answer to Q1 Research Question

Backchannel prediction did not become more accurate when the two tasks of turn-changing and backchannels were predicted simultaneously using the multi-task learning framework. This demonstrates that adding turn-changing as a prediction target does not improve the performance of backchannel prediction. Backchannel and turn-changing/keeping are very closely related in dyad interactions, but jointly predicting them at the same time does not seem to improve the performance of backchannel prediction. This result is in line with the results of another previous study [13] that attempted to simultaneously predict turn-changing and backchannel using only the listener's voice acoustic information.

6.2 Answer to Q2 Research Question

Backchannel prediction became more accurate when the two tasks of turn-management willingness and backchannel were predicted simultaneously by using multi-task learning. This demonstrates that explicitly adding turn-management willingness as a prediction target improves the performance of backchannel prediction. This introduces new possibilities for more accurately predicting human behavior by predicting human psychological states at the same time in conversations. The multi-task learning approach allows a prediction model to learn the underlying relationship between turn-changing willingness and backchannels. This result was similar to the results of another previous study [27] that investigated the simultaneous prediction of turn-management willingness and turn-changing using the trimodal information of speakers and listeners. Therefore, multi-task learning supports a prediction model in learning the underlying relationship between turn-management willingness and backchannels.

6.3 Answer to Q3 Research Question

There was no significant difference in the performance of backchannel prediction when two-task learning for turn-management willingness and backchannel prediction and three-task learning for turn-changing, turn-management willingness, and backchannel prediction were applied. This demonstrates that adding turn-management willingness alone as a prediction target improves the performance of backchannel prediction.

From these results, it is thought that predicting multiple tasks of human behaviors at the same time has no effect on improving the prediction performance, but rather predicting human behavior and mental states at the same time can improve the prediction of human behavior. In other words, it was shown that it is important to predict internal states simultaneously in order to increase the performance of predicting human behaviors during dialogues. This introduces new possibilities for more accurately predicting human behaviors by predicting human psychological states simultaneously in conversations.

6.4 Future Work

Our goal was to study the impact of jointly predicting turn-changing and turn-management willingness on backchannel prediction. We used automatic high-level abstract features extracted from acoustic, linguistic, and visual modalities of speakers and listeners. We plan to extract interpretable features, such as prosody [10, 16, 17, 20, 37, 38, 41] and gaze behavior [3, 21, 25, 28, 30], and implement more advanced prediction models [37, 38, 41, 49] that were proposed for turn-changing prediction but could be applied to backchannel prediction to take into account temporal dependencies.

We treated only the utterance backchannel, which was a short utterance such as "yeah," "uh-huh," "hmm," and "right," as the prediction target. We plan to investigate the prediction of multi-modal backchannels such as nodding and facial expressions by using our proposed multi-modal and multi-task learning approach.

We also plan to incorporate our multi-modal multi-task prediction models into conversational agent systems that can display backchannel behaviors and increase the naturalness of agent behaviors. Since our backchannel prediction model uses multi-modal behavior information and turn-management willingness from both the speaker and the listener as input features, it is necessary to include the past multi-modal behaviors and willingness of a conversational agent as well when generating backchannels of the agent. Therefore, we aim to apply our prediction model to agents that can express human-like multi-modal behaviors and have turn-management willingness.

7 CONCLUSION

We built multi-modal and multi-task machine learning models for predicting a listener's backchannel as well as turn-changing and turn-management willingness on the basis of trimodal behaviors in conversations, that is, acoustic, linguistic, and visual cues. An evaluation of our prediction models showed that a backchannel is predicted most accurately when all of three of these modalities from the speaker and listener are processed. Furthermore, backchannel prediction becomes more accurate when turn-management willingness and backchannels are predicted jointly by using multi-task learning. However, joint turn-changing prediction is not useful for backchannel prediction. These results suggest that more accurate prediction models of human behaviors could be proposed and built by incorporating another prediction task 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] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *FG*. 59–66.
[2] P. Blache, Massina Abderrahmane, S. Rauzy, and R. Bertrand. 2020. An integrated model for predicting backchannel feedbacks. In *IVA*.
[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] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *INTERSPEECH*. 889–892.
[12] 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.
[13] 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.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.
[15] 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.
[16] Judith Holler and Kobin H. Kendrick. 2015. Unaddressed Participants' Gaze in Multi-person Interaction: Optimizing Recipiency. *Frontiers in Psychology* 6 (2015), 515–535.
[17] 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.
[18] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. *AAMAS* 2, 1265–1272.
[19] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. A Multimodal End-of-Turn Prediction Model: Learning from Parasocial Consensus Sampling. In *AAMAS*.
[20] 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.
[21] 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.
[22] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting Next Speaker Using Head Movement in Multi-party Meetings. In *ICASSP*. 2319–2323.
[23] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2017. Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings. In *HAI*. 181–187.
[24] 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.
[25] 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.
[26] 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.
[27] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2020. Can Prediction of Turn-Management Willingness Improve Turn-Changing Modeling?. In *IVA*.
[28] 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.
[29] Jeffrey Kahn, Renée Tobin, Audra Massey, and Jennifer Anderson. 2007. Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *J. psychology* 120 (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] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*. 13.
[32] N. Kitaoka, M. Takeuchi, Ryota Nishimura, and Seiichi NAKAGAWA. 2005. Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems. *Trans. Japanese Society for Artificial Intelligence* 20 (2005), 220–228.
[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] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2008. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In *IVA*. 176–190.
[40] Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using Neural Networks for Data-Driven Backchannel Prediction: A Survey on Input Features and Training Techniques. In *Human-Computer Interaction: Interaction Technologies*. 329–340.
[41] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. In *ICMI*. 186–190.
[42] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. *Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor*. 247–258.
[43] 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.
[44] David Schlangen. 2006. From Reaction to Prediction: Experiments with Computational Models of Turn-taking. In *INTERSPEECH*. 17–21.
[45] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.
[46] 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.
[47] Khiet P. Truong, Ronald Poppe, and Dirk Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information.. In *INTERSPEECH*. ISCA.
[48] N. Ward. 1996. Using prosodic clues to decide when to produce back-channel utterances. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 3. 1728–1731 vol.3.
[49] 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.
[50] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 32, 8 (2000), 1177–1207.