Multimodal Turn Analysis and Prediction for Multi-party Conversations

Meng-Chen Lee
Department of Computer Science
University of Houston
Houston, Texas, USA
mlee45@uh.edu

Mai Trinh
Department of Computer Science
University of Houston
Houston, Texas, USA
mptrinh1918@gmail.com

Zhigang Deng
Department of Computer Science
University of Houston
Houston, Texas, USA
zdeng4@central.uh.edu

ABSTRACT

This paper presents a computational study to analyze and predict turns (i.e., turn-taking and turn-keeping) in multiparty conversations. Specifically, we use a high-fidelity hybrid data acquisition system to capture a large-scale set of multi-modal natural conversational behaviors of interlocutors in three-party conversations, including gazes, head movements, body movements, speech, etc. Based on the inter-pausal units (IPUs) extracted from the in-house acquired dataset, we propose a transformer-based computational model to predict the turns based on the interlocutor states (speaking/back-channeling/silence) and the gaze targets. Our model can robustly achieve more than 80% accuracy, and the generalizability of our model was extensively validated through cross-group experiments. Also, we introduce a novel computational metric called "relative engagement level" (REL) of IPUs, and further validate its statistical significance between turn-keeping IPUs and turn-taking IPUs, and between different conversational groups. Our experimental results also found that the patterns of the interlocutor states can be used as a more effective cue than their gaze behaviors for predicting turns in multiparty conversations.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in HCI**.

KEYWORDS

Multi-party conversations, conversational gesture understanding, Multimodal interaction, Machine learning, Human-human communication, Empirical studies

ACM Reference Format:
Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal Turn Analysis and Prediction for Multi-party Conversations. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3577190.3614139

1 INTRODUCTION

Multiparty conversation is a prevalent form of human communication in society, where participants naturally employ their voice, gaze, and gestures to facilitate idea exchange and manage conversational flow. In recent years, understanding and modeling the interactions and associations among multimodal communication channels - facial expressions, gaze, hand gestures, and linguistic content - have garnered increasing research interest in various fields, including communication, human-computer interaction, graphics, robotics, and multimodal interaction communities [10, 19, 33].

Turn-taking and turn-keeping in multiparty conversations are essential for effective communication among participants. By alternating speaking roles and responding appropriately, participants ensure equal opportunities to express their ideas and contribute to discussion. Ineffective turn management can result in interruptions, overlaps, and confusion, ultimately hindering conversation progress and reducing its overall quality.

A primary challenge in modeling and predicting turns in multiparty conversations is that turn management relies not only on linguistic content but also on non-verbal cues, such as gazes, head movements, and other gestures. For example, previous research has demonstrated a strong correlation between gazes and turn-taking in multiparty conversations [7, 8, 26], with gaze directions often signaling turn-yielding intentions or serving as a strategy to alleviate cognitive demands during response formulation [26].

Previous studies on turn prediction in multiparty conversations have relied on video recordings and manual annotations [13, 16, 17, 25], which inherently limits the accuracy of gesture signal data, particularly in 2D recordings. Additionally, while previous research has considered gaze transition patterns, none has explored the timing of gaze transitions or interlocutor states (i.e., speaking/back-channeling/silence) in multiparty conversations for turn prediction.

To address the above limitations of existing methods, we developed an in-house setup to capture large-scale, multi-modal conversational behaviors of interlocutors in multiparty conversations using a combination of optical motion capture, eye trackers, and high-definition microphones. Our dataset, including simultaneously recorded 3D head movements, 3D gazes, 3D body movements, 3D hand movements, and speech, provides more accurate conversational gesture signals for turn prediction and analysis, compared to conventional 2D video-based data. In this work, without loss of generality, we specifically focus on three-party conversations.

Leveraging this dataset, we propose a transformer-based turn prediction model. We adopt the inter-pausal unit (IPU) concept from previous studies [27], representing utterances preceding longer than 200 millisecond pauses, and extract all IPUs before turn-taking

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614139

and turn-keeping occurrences in our dataset. Based on the extracted interlocutor states (i.e., speaking, back-channeling, or silence) and gaze targets of each frame in one IPU, we develop and train a transformer-based model to predict turns for each IPU. Through extensive experiments, including cross-group validations, we demonstrate that our model can robustly and measurably outperform a Hidden Markov Models (HMM)-based baseline and previous related work. Additionally, our ablation study reveals that the interlocutor states with timing can be a more effective cue than gaze targets with timing for the purpose of turn prediction in multi-party conversations.

We also introduce a novel computational measure called the "Relative Engagement Level" (REL) to characterize each participant's participation in an IPU. For example, we assign decreasing involvement weights for speaking, back-channeling, and silence (i.e., neither speaking nor back-channeling) when computing the REL. Statistical analysis on our large-scale dataset indicates that the REL differences between turn-taking IPUs and turn-speaking IPUs for all interlocutors are statistically significant. Furthermore, the statistical significance of REL differences persists across various conversational groups. We contend that the REL measure introduced in our study provides a novel quantitative lens to examine turns in multi-party conversations.

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of previous works closely related to our study. Subsequently, we detail our data acquisition process in Section 3 and outline the extraction of interlocutor states and gaze features from our dataset in Section 4. In Section 5, we describe the REL measure and present our analysis results. We then discuss the specifics of our transformer-based turn prediction model and an HMM-based baseline model in Section 6, followed by the presentation of turn prediction experimental results in Section 7. Finally, we offer discussion points and future work remarks in Section 8.

2 RELATED WORK

2.1 Gaze Behavior for Turn Regulation

Gaze behavior and turn-taking are critical aspects of human conversations and have been extensively studied in the field of communication research. Early studies [7, 8, 26], for example, have empirically demonstrated that gaze signals are employed to regulate and monitor turn-taking. More recent research [12] has reported that speakers tend to direct their gaze toward the next speaker more frequently during turn-taking than turn-keeping, suggesting that gaze signals play a vital role in facilitating smooth turn transitions and maintaining communication flow in dynamic social contexts. Some studies have even shown that gaze behavior is more important than speech information for estimating turn-taking [5, 18, 23]. However, it has been argued that gaze should be considered as a part of the inherently multi-modal interaction rather than a singular occurrence [11].

In the context of multiparty conversations, researchers have found that unaddressed participants use their gaze behaviors to effectively manage and monitor the conversation [2, 24, 34], including preventing simultaneous speech and resolving disputes when two speakers vie for a turn [29, 35]. During multiparty conversations, even those not directly involved can comprehend each participant's intent and contribute to keeping the conversation on track. However, individuals who are not fully focused on the conversation and act as bystanders may disrupt the conversation flow by interrupting speakers [29].

2.2 Turn Prediction

According to the study by Aldeneh et al. [1], the decision to take turns in a dialogue must involve some level of anticipation or prediction. This prediction mechanism can explain the short average pause or gap that typically occurs between turns taken by different speakers in a conversation. Several studies have investigated the use of verbal and non-verbal cues to predict turn-taking in multi-party conversations. For example, researchers have utilized speech features to develop prediction models. Aldeneh et al. [1] perceive turn-taking in conversations as a problem of arranging a sequence of actions and proposed a forecasting model that employs Long Short-Term Memory (LSTM) algorithms. The model also takes into account speech characteristics such as loudness, intensity, and zero-crossing rate. Researchers also trained the model on the Switchboard dataset [9] and achieved an F1 score of 0.65 for predicting turn-taking in a conversation.

Meanwhile, researchers have explored the use of gaze information in the realm of automated turn-taking detection in multi-party meetings. For example, Kawahara et al. [25] proposed a model that uses participants' gaze information, including the person being gazed upon and the presence or absence of mutual gaze, along with prosody to detect turn-taking in a three-person poster conversation.

Ishii et al. [17] demonstrated that a detailed analysis of gaze transition patterns, such as changes in gaze targets and mutual gazes, is more effective in predicting turn-taking than relying on a single line of gaze. Their study used 12 gaze transition patterns and achieved an F1 score of 0.76 for the prediction of turn changes. Ishii et al. [13] also incorporated human gaze and respiratory behavior to predict turn changes and the next speaker in multi-party conversations. They conducted experiments on a custom dataset with four participants and used Sequential Minimal Optimization, a variation of SVM models, to train the models. Their results showed that the model based on the late fusion of eye gaze and respiratory behavior yielded an F1 score of 0.75 for the prediction of turn changes. Additionally, they also proposed a model that utilizes the head movements of both the speaker and listeners towards the end of an utterance [14], achieving an accuracy of 75% for turn-taking prediction using SVM models. Furthermore, Ishii et al. [16] investigated the role of mouth-opening transition patterns in turn-taking and trained SVM models to predict turn changes, reporting an F1 score of 0.80.

However, the above studies did not consider the timing of gaze transitions in the utterance unit. Hessels [11] emphasized the importance of considering the time-dependent nature of gaze during social interactions. Specifically, Hessels suggested linking gaze patterns to the evolving features of other interactional sub-states over time, rather than relying solely on aggregated measures of gaze that overlook such temporal dynamics.

3 DATA ACQUISITION

To acquire a large-scale, multi-modal, three-party conversational behavior dataset for this study, we specifically built an in-house, hybrid data acquisition system through the combination of optical motion capture, eye trackers, and HD microphones, as described below. Figure 1 illustrate the schematic view of the data acquisition setup. Note that the three interlocutors according to their locations are labeled as P1, P2, and P3, respectively, as shown in Figure 1. Also, Figure 2 shows a snapshot of our data acquisition process.

Motion Capture. A eight-cameras VICON optical motion capture system was set up in a controlled lab environment to capture the motions of interlocutors. We asked each interlocutor to wear a motion capture (mocap) suit attached with optical markers to record their body motions, including head, hands, torso, and lower body movements (see Figure 2). In addition, from the recorded 3D mocap data, we calculated the angles of all the joints for each interlocutor using inverse kinematics.

Gaze and Audio. In our study, we equipped each participant with an Ergoneers Dikablis Glass 3 eye tracker, chosen for its exceptional accuracy, wireless capabilities, and comprehensive software for detailed analysis. To assess the impact of wearing the eye tracker on participants’ natural conversational behavior, including gaze, we gathered their feedback. The consensus among participants was that the eye tracker had only minimal influence on their behavior. It should be noted that the eye tracker used in our research closely resembled conventional eyeglasses in terms of size and weight, ensuring its unobtrusive nature during the experiments. In addition, we provided each participant with a wireless microphone hooked to their neckline to record audio. We employed the D-LAB software to capture eye gaze and audio simultaneously.

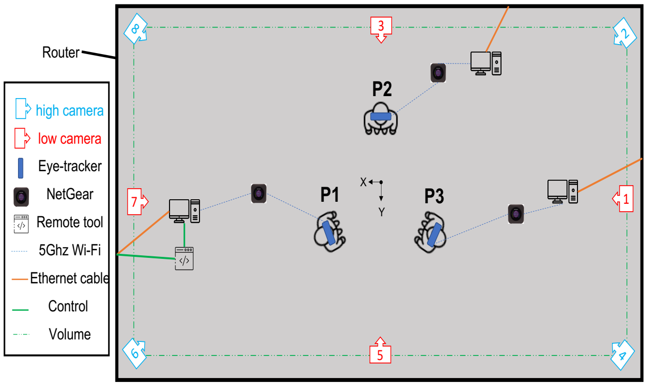

Figure 1: Schematic illustration of our hybrid data acquisition setup

Capture Design. We recruited 21 volunteer participants from a university campus for our data capture experiments. They were randomly divided into seven groups (each group had three participants). The 21 participants did not know each other prior to our data acquisition experiment. Among them, 12 are male and 9 are female, with ages ranging from 20 to 30. The majority of them (70%) have computer science majors. All of them are native English speakers.

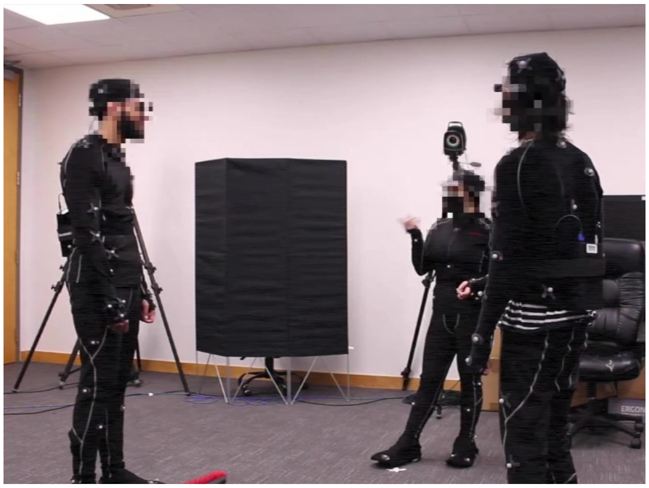

Figure 2: A snapshot of our three-party conversational motion data acquisition experiment.

For each group, we recorded 3-5 sessions of its three-party conversations. The lengths of the recorded data for the seven groups are: 46’18" (46 minutes 18 seconds), 44’47", 46’17", 42’49", 46’31", 44’27", and 48’19", respectively. The total amount of recorded data is 319 minutes 28 seconds.

During the data acquisition process, three interlocutors in a conversation were instructed to stay at fixed locations and form an equilateral triangle with a distance of approximately 1 meter from each other. Before recording each session, the participants were asked to stay at their original locations as much as possible. The spatial arrangement of interlocutors in our study is similar to previous studies that investigated three-party conversations [3, 6, 20–22]. Furthermore, our setup is comparable to the setup in [31], which ensured that all five participants formed a regular pentagon. The participants were allowed to discuss any topic of interest during the data acquisition process. A clapperboard was used to synchronize the start time of all different data. To this end, all acquired data (i.e. 3D mocap, eye tracking data, and speech) were temporally aligned and later down-sampled to 60 frames per second.

4 DATA POST-PROCESSING AND FEATURES EXTRACTION

In this section, we describe how we extract IPUs from the acquired dataset (Section 4.1), and extract features to build turn prediction models, including interlocutor states (Section 4.2) and gaze targets (Section 4.3).

4.1 IPU Extraction

Following previous IPU-based studies [13, 15, 18, 28, 30], we also extract Inter-Pausal Units (IPUs) [27] from our dataset as the *utterance unit of interest* in our study. Specifically, in the data we first identified any pauses lasting longer than 200 milliseconds, and then assigned the utterance preceding each such pause as an IPU. Figure 3 illustrate examples of the IPU extraction process.

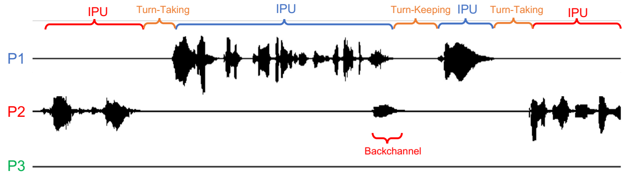

Figure 3: Example illustration of IPU extraction and turn events

Based on the extracted IPUs, we further categorized pauses into two types: turn-taking pauses, which occur between two IPUs by distinct speakers, and turn-keeping pauses, which occur for the same speaker. To the end, we extracted a total of 7,080 IPUs from the seven groups of three-party conversations. Among them, 5,975 IPUs occur before turn-taking pauses, and 1,105 IPUs occur before turn-keeping pauses. Thus, in this work we call them *turn-taking IPUs* and *turn-keeping IPUs*, respectively. The shortest IPU has 32 frames (i.e., 0.53 seconds), and the longest IPU has 1755 frames (i.e., 29.25 seconds), and the average length of the IPUs is 201 frames (i.e., 3.35 seconds). After the IPU extraction, we padded all the IPUs to the maximum length (1,755 frames) for model training later. Note that the IPUs do not temporally overlap since we discarded any overlapping speech segments. Table 1 provides a detailed breakdown of the extracted IPUs by subjects, while Table 2 offers a detailed analysis of the IPUs by groups.

Table 1: IPU distribution by subjects. The information in each cell contains the total number of the specific type of IPUs by the specific subject, with the average number of frames ± the standard deviation in the parentheses.

Subject	Turn-Taking	Turn-Keeping	Total
P1	1972 (183 ± 143)	355 (199 ± 165)	2327 (186 ± 146)
P2	2092 (198 ± 164)	513 (252 ± 206)	2605 (198 ± 164)
P3	1911 (201 ± 182)	237 (266 ± 217)	2148 (201 ± 179)

4.2 Interlocutor States

We extracted an interlocutor state feature vector from each IPU as follows. First, the audio was transcribed into texts using Google speech-to-text technology, followed by manual check and corrections (if needed). Then, to identify the back-channeling utterances, we employed the criteria defined in [25], which include "Yeah," "Um," "Cool," "Oh," "Okay," "Right," and "Uh", and flagged the corresponding places in the transcriptions. To this end, for each frame in one IPU, we created a 1 × 3 ternary-variables vector, called the *Interlocutor State Vector* (ISV), to represent the states of the three interlocutors in a conversation: the value of its i-th component (i is from 1 to 3) indicates the i-th interlocutor is speaking (=2), back-channeling (=1), or silence (neither speaking nor back-channeling) (=0).

Table 2: IPU distribution by groups. The information in each cell contains the total number of the specific type of IPUs by the specific subject, with the average number of frames ± the standard deviation in the parentheses.

Group	Turn-Taking	Turn-Keeping	Total
1	699 (171 ± 138)	232 (240 ± 183)	931 (188 ± 154)
2	617 (168 ± 131)	247 (191 ± 163)	864 (175 ± 141)
3	841 (190 ± 145)	112 (245 ± 195)	953 (196 ± 153)
4	1064 (208 ± 173)	137 (210 ± 179)	400 (208 ± 174)
5	655 (209 ± 192)	84 (265 ± 238)	739 (215 ± 198)
6	939 (190 ± 160)	158 (305 ± 248)	1097 (207 ± 180)
7	1160 (206 ± 178)	135 (249 ± 184)	1295 (211 ± 179)

4.3 Gaze Targets

Inspired by previous studies [21], we also computed the Direction-of-Focus (DFoc) of an interlocutor by combining the torso-head direction and the gaze direction. Specifically, assuming we use R_{hips}, R_S and R_H to denote the 3D rotations of the hip, spine, and head, respectively, and use G_h, O_S, O_H, and O_E to denote the global location of the hip and the offsets of the spine, head, and eyes, respectively, we can use the following equation to compute DFoc. Note that we eliminated outliers and interpolated gaps using shape-preserving piece-wise cubic interpolations for all V_E.

$$DFoc = G_h R_{hips} O_S R_S O_H R_H O_E V_E. \tag{1}$$

Instead of directly using the above computed DFoc, we further computed a high-level feature called *Focus-of-Attention* (FoA), which indicates which of the other interlocutors an interlocutor is looking at. To compute the FoA, a head-sphere is constructed to cover an interlocutor's head (refer to Figure 4), and then we check whether the DFoc of the interlocutor intersects with the head-spheres of other interlocutors. To this end, for each frame in one IPU, we created a 1 × 3 ternary-variables vector, called the *Gaze Target Vector* (GTV), to represent the FoAs of all the three interlocutors: the value of its i-th component is the index of the interlocutor who receives the attention of the i-th interlocutor at this specific frame.

5 RELATIVE ENGAGEMENT LEVEL

In this study, we introduce a novel computational measure, called *Relatively Engagement Level* (REL), to quantitatively analyze turn-taking and turn-keeping in multiparty conversations. Specifically, we study how the REL of each interlocutor in IPUs is associated

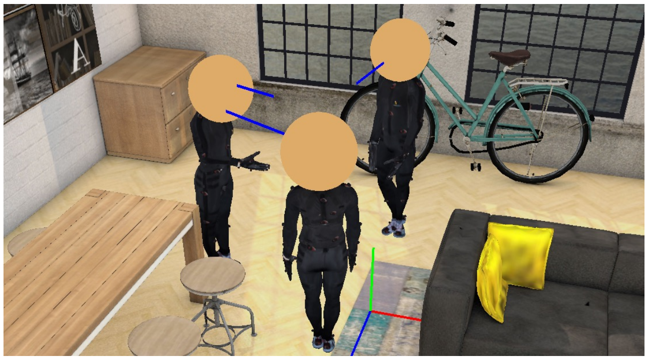

Figure 4: Visualization for the FoA computation process. The computed DFoc is visualized as a blue line.

Visualization for the FoA computation process. The computed DFoc is visualized as a blue line.

with turn-taking and turn-keeping. In the following, we describe how we compute the REL for an IPU.

Assume an IPU U^i has a total of n frames, and the interlocutor state vector of the j-th frame is v^j (a 1 × 3 ternary-variables vector, refer to Section 4.2), we can calculate the REL of the m-th interlocutor in U^i, $REL(U^i_m)$, using the following equation.

$$REL(U^i_m) = \frac{\sum_{k=1}^{n} v^k_m}{\sum_{k=1}^{n}\sum_{m=1}^{3} v^k_m}, (m = 1...3) \qquad (2)$$

where v^k_m denote the m-th component of v^k, and it has three possible values: 2 for speaking, 1 for back-channeling, and 0 for silence (neither speaking nor back-channeling). Conceptually, $REL(U^i_m)$ denotes the relative involvement extent of the m-th interlocutor during this specific IPU. Here we implicitly assume speaking represents a high level of conversational engagement, back-channeling is a medium level of conversational engagement, and silence is a low level of conversational engagement.

As mentioned in our data post-processing (Section 4), we categorized IPUs into turn-taking (TT) IPUs and turn-keeping (TK) IPUs. For TT IPUs (or TK IPUs), we further categorized them by subjects (P1, P2, and P3). To this end, we have six sub-categories of IPUs: TT IPUs for P_i (i = 1 to 3), TK IPUs for P_i (i = 1 to 3). For each sub-category, we compute the mean REL value and standard deviations for all the IPUs in this sub-category. Finally, we visualize these mean REL values and standard deviations in Figure 5.

As illustrated in Figure 5, for all the interlocutors, the average REL values for TK are significantly higher than those for TT. This implies that the speaker who holds the floor tends to have a higher REL value before turn-keeping than before turn-taking. This trend is observed consistently across all the interlocutors. Also, this suggests that turn-taking is more likely to occur when listeners are more actively involved with the conversation, such as interrupting the speech or actively providing feedback through back-channeling, which would decrease the REL value of the current speaker. Similar patterns can also be observed in Figure 6, which shows that listening interlocutors are more likely to speak before turn-taking.

We also performed an analysis of variance (ANOVA) to statistically compare different sub-categories of IPUs in terms of the REL values. The results of single-variable ANOVA on the REL values of IPUs are presented in Table 3. As shown in Table 3, for all the interlocutors, the REL differences between turn-taking and turn-keeping are statistically significant (that is, the p-values are well below 0.05). Further, we also used single-variable ANOVA to analyze the REL differences between different conversation groups. Recall that we acquired data for seven distinct conversation groups. As shown in Table 3, the REL differences between different conversational groups are not statistically significant (actually, the p-values are well above 0.05). This provides statistical evidence to support the robustness and generalizability of our findings between different groups of interlocutors. We believe that the REL measure can provide a new quantitative useful tool to look into turn management in multi-party conversations.

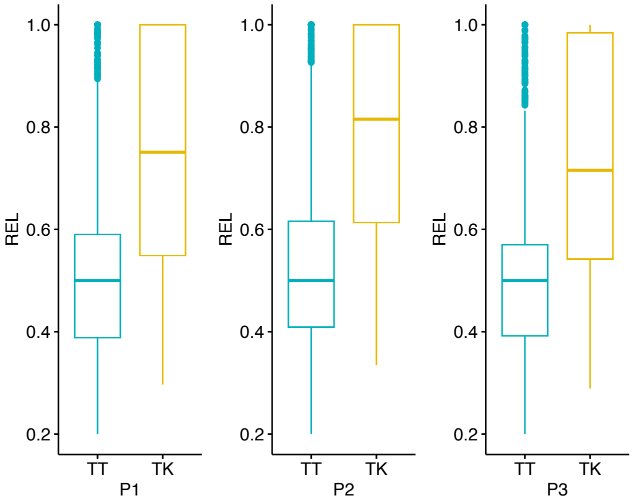

Figure 5: Plotting of the mean REL values and standard deviations of the six sub-categories of IPUs in our dataset. TT denotes turn-taking, and TK denotes turn-keeping.

Plotting of the mean REL values and standard deviations of the six sub-categories of IPUs in our dataset. TT denotes turn-taking, and TK denotes turn-keeping.

Table 3: Results of single-variable ANOVA on the REL values of IPUs between different conversation groups and between turn taking/keeping.

Comparison between	Subject	P-value
Conversational Groups	P1	0.643
	P2	0.614
	P3	0.357
Turn Taking/Keeping	P1	0.00547 **
	P2	0.00239 **
	P3	0.0213 *

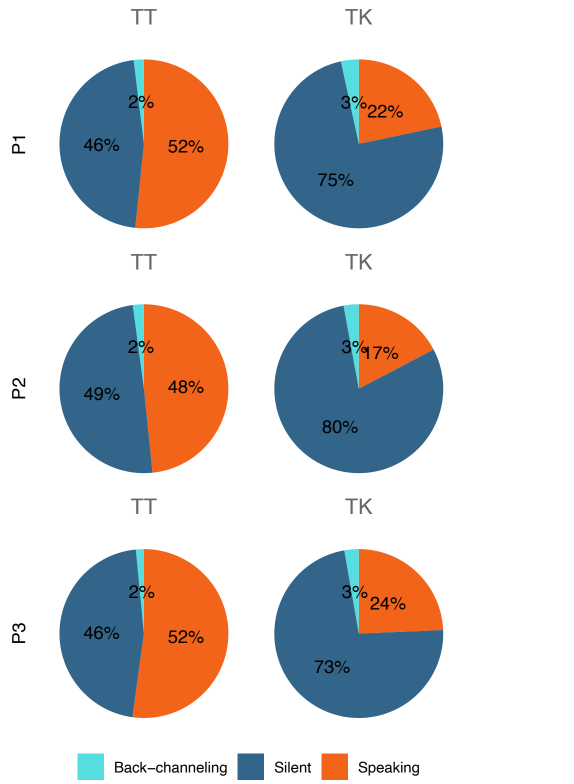

Figure 6: Plotting of the percentages of interlocutors' ISVs in our dataset. TT denotes turn-taking, and TK denotes turn-keeping.

Plotting of the percentages of interlocutorsâ€™ ISVs in our dataset. TT denotes turn-taking, and TK denotes turn- keeping.

6 TURN PREDICTION MODELS

In this section, we describe the details of our transformer-based turn prediction model (Section 6.1), which makes turn predictions according to the IPU preceding the turn. We also compare our model with a baseline model - a classical Hidden Markov Model (HMM) based turn prediction model. As such, we will also describe the details of the baseline model (Section 6.2).

6.1 Transformer-based Turn Prediction

The architecture of the proposed turn prediction model, based on the transformer framework [4], is illustrated in Figure 7. The model takes both the ISVs and the GTVs of an IPU as input features (see Section 4). We describe the details of its key modules below.

Our model includes an embedding layer to map discrete feature inputs to continuous representations. A positional embedding layer with the same dimensions is used to incorporate positional information into the model.

The Transformer Encoder consists of 2 Transformer Encoder Layers, each of which includes a self-attention module that performs multi-head attention over the input sequence. Each Transformer Encoder Layer also consists of a feed-forward neural network layer, which applies a linear transformation followed by a non-linear activation function, and two normalization layers, which perform layer normalization over the output of the self-attention and feed-forward layers, respectively. Both the self-attention and feed-forward layers in each Transformer Encoder Layer are followed by a dropout (set to 0.1), which is used to regularize the model and prevent over-fitting. The output of the Transformer Encoder is then passed through a fully-connected layer with weights and biases learned during training. The fully-connected layer maps the output of the Transformer Encoder to determine whether the IPU results in turn-taking or turn-keeping.

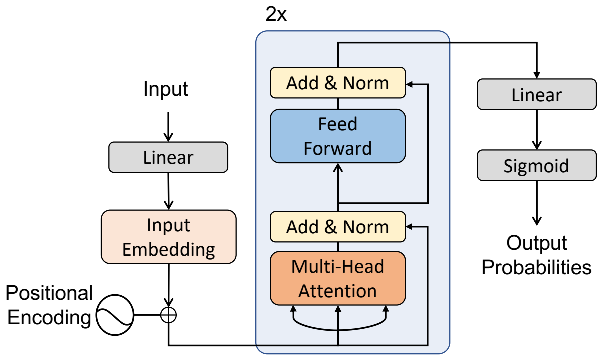

Figure 7: The architecture of the transformer-based turn prediction model

The architecture of the transformer-based turn prediction model

6.2 HMM-based Baseline

We also developed a Hidden Markov Model (HMM) based model as the baseline to compare it with our transformer-based turn prediction model. The HMM model [32] is widely known for its ability to model time series data. The baseline model consists of ten hidden states. To ensure a comprehensive analysis of state transitions, we employed the ergodic topology (EG) for the HMM model, which is fully connected, less restrictive, and capable of learning a vast set of state transitions from data. The probability density function for observation was modeled using a mixture of Gaussians. We utilized standard Baum-Welch re-estimation and Viterbi algorithms for training the parameters of the HMM and inferring the optimal sequence of hidden states. For the training process, we empirically set the maximum number of iterations to 100.

7 RESULTS AND EVALUATIONS

To evaluate the effectiveness of our model, we randomly divided our dataset into two sets: 90% for training and 10% for testing. Then, we applied the training set with 9-fold cross-validation and calculated the accuracy, recall, and F-measure of the test set to assess its performance. Additionally, we performed a cross-group validation, separating seven different groups of conversational behaviors into a 5-1-1 format, using five groups for training, one for validation, and one for testing. Through the cross-group validation, we are able to accurately estimate the model's generalizability and robustness.

Moreover, this helps to ensure that our trained model is not biased toward a particular subset of the data.

Table 4 presents the quantitative comparison results of our transformer-based model, a baseline model, and a recent related work [17]. Note that while the work in [17] used different datasets, our setups and labeling methods are generally similar to those in [17]. The key distinction lies in our utilization of 3D features in the datasets. In contrast, the work in [17] used 2D features. As shown in this table, our transformer-based approach can significantly outperform both the HMM-based baseline model, and the previous work by Ishii et al. [17], achieving precision 0.812, recall rate 0.845, and F-measure 0.805. As shown in Table 5, the obtained p-values: 1.859e-06, 1.163e-4, and 1.099e-4, indicate strong statistical significance of the transformer model over the baseline model in all metrics evaluated.

Table 4: Quantitative comparison results of turn prediction by three different methods: our transformer-based model, the HMM-based baseline model, and the previous work by Ishii et al. [17].

	Metrics		
Model	Precision	Recall	F-Measure
Ishii et al. [17]	0.763	0.775	0.769
Baseline Model	0.778	0.794	0.785
Transformer-based Model	0.812	0.845	0.805

Table 5: Two-sample t-test results, comparing the baseline and transformer-based model.

Model Tested against baseline	**Metric**	***t***	***p***
Transformer Encoder	Precision	6.5513	1.859e-6 **
	Recall	9.8427	1.163e-4 **
	F1	5.1142	1.099e-4 **

7.1 Cross-group Validation

Table 6 shows the quantitative results of cross-group validation using our model. The results were obtained using the 5-1-1 format and each of the seven conversational groups in our data was used as a test set. Since the data capture participants in different conversational groups did not overlap, the participants in each test set were not observed or included in the training or validation sets. The results in Table 6 demonstrate the robustness of our model, and it can be effectively generalized to new multiparty conversations by novel subjects, compared to the results presented in Table 4.

7.2 Ablation Study

We also conducted an ablation study to compare our model's performance given different combinations of input features.

As shown in Table 7, our model, which utilizes both ISV (Interlocutor State Vector) and GTV (Gaze Transition Vector) as input features, achieves the highest performance across all metrics: Precision is 0.812, recall rate is 0.845, and F-measure is 0.805. Results from our ablation study suggest that although both ISV and GTV contribute significantly to the turn prediction model, ISV (i.e., when interlocutors are speaking, back-channeling, or silent) proves to be a more effective cue than GTV (i.e., when interlocutors look at other interlocutors) for our turn prediction model.

Table 6: Quantitative cross-group validation results by our model

	Metrics		
Test Group	Precision	Recall	F-Measure
1	0.804	0.816	0.784
2	0.777	0.778	0.734
3	0.825	0.874	0.837
4	0.841	0.883	0.848
5	0.845	0.864	0.853
6	0.830	0.863	0.827
7	0.859	0.883	0.868

Many previous studies have emphasized the importance of gaze transition patterns in turn prediction, which our findings corroborate. However, our research is the first to experimentally highlight the critical role of interlocutor states in turn prediction for multiparty conversations.

Table 7: Ablations study results by our model with different combinations of input features. Here ISV refers to the Interlocutor State Vector, and GTV refers to the Gaze Target Vector (see their definitions in Section 4).

	Metrics		
Features	Precision	Recall	F-Measure
ISV + GTV	0.812	0.845	0.805
ISV	0.806	0.842	0.800
GTV	0.703	0.838	0.765

8 DISCUSSION AND CONCLUSION

In this paper, we introduce an innovative IPU-based method for turn analysis and prediction in multi-party conversations, specifically focusing on three-party interactions. Our approach is unique in that it eschews the less accurate, manually annotated 2D video data utilized in previous studies, opting instead for a large-scale, multimodal (mainly 3D) dataset of multiparty conversational behaviors. This dataset accurately captures simultaneous 3D conversational gesture signals from all participants, including 3D gazes, 3D head movements, 3D body movements, and speech. The high precision of our large-scale 3D dataset allows us to examine turns in multi-party conversations at an unparalleled level of accuracy.

Capitalizing on our large-scale, multi-modal dataset, we propose a new metric called the Relative Engagement Level (REL) to quantify the relative involvement effort of each participant within an IPU. We discovered that the REL differences between turn-taking and turn-keeping IPUs are statistically significant, and this distinction persists across different conversational groups. These results offer initial evidence that the REL metric could serve as a novel, quantitative lens through which to investigate turns in multi-party conversations.

Employing both the interlocutor state vector and the gaze target vector extracted from the IPUs (categorized into turn-taking IPUs and turn-keeping IPUs), we further devised a transformer-based turn prediction model. This model achieves a performance greater than 80% in all three evaluation metrics (precision, recall, and F measure) and outperforms the baseline model based on the Hidden Markov Model (HMM) and a related previous study. Furthermore, our results of cross-group validation underscore the robustness and generalizability of our approach.

Our ablation study reveals that the interlocutor state vector contributes more significantly to our turn prediction model than the gaze target vector. While many prior studies have emphasized the importance of gaze patterns for turn prediction in multi-party conversations, our research is the first, to the best of our knowledge, to identify and experimentally validate the critical role of interlocutor states in IPUs for turn prediction.

Despite the promising results of our current approach, there are some limitations. First, our method does not incorporate linguistic analysis or individual word features, which could potentially enhance the performance of our model. Second, while we have explicitly included verbal back-channeling into our features, we do not consider nonverbal cues like head nodding or shaking. Moreover, given that our turn prediction model is based on IPUs, its application is restricted to offline scenarios. In the future, we plan to refine our turn prediction model by integrating additional features and leveraging large language models. Lastly, we intend to investigate patterns and relationships not only within groups or sequences, but also across different modalities, as a part of our ongoing research.

ACKNOWLEDGMENTS

This work is supported in part by the NSF IIS grant 2005430.

REFERENCES

[1] Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. 2018. Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task. In *Proceeding of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6159–6163.

[2] Peter Auer. 2018. Gaze, addressee selection and turn-taking in three-party interaction. In *Eye-tracking in interaction: Studies on the role of eye gaze in dialogue*. John Benjamins Amsterdam, 197–231.

[3] Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. 2017. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics* 28, 3 (2017), 449–483.

[4] Francois Chollet. 2021. *Deep learning with Python*. Simon and Schuster.

[5] Iwan De Kok and Dirk Heylen. 2009. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 international conference on Multimodal interfaces*. 91–98.

[6] Yu Ding, Yuting Zhang, Meihua Xiao, and Zhigang Deng. 2017. A multifaceted study on eye contact based speaker identification in three-party conversations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3011–3021.

[7] Starkey Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology* 23 (1972), 283–292.

[8] Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns. *Language in Society* 3, 2 (1974), 161–180.

[9] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceeding of IEEE International Conference on Acoustics, speech, and signal processing*, Vol. 1. IEEE Computer Society, 517–520.

[10] Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. 2014. Eye gaze for spoken language understanding in multi-modal conversational interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 263–266.

[11] Roy S. Hessels. 2020. How does gaze to faces support face-to-face interaction? A review and perspective. *Psychonomic Bulletin & Review* 27, 5 (may 2020), 856–881.

[12] Koki Ijuin, Ichiro Umata, Tsuneo Kato, and Seiichi Yamamoto. 2018. Difference in Eye Gaze for Floor Apportionment in Native- and Second-Language Conversations. *Journal of Nonverbal Behavior* 42 (03 2018), 1–16.

[13] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 99–106.

[14] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2319–2323.

[15] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2018. Analyzing gaze behavior and dialogue act during turn-taking for estimating empathy skill level. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 31–39.

[16] 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).

[17] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato. 2013. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 79–86.

[18] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems (TIIS)* 6, 1 (2016), 1–31.

[19] Aobo Jin, Qixin Deng, and Zhigang Deng. 2020. A Live Speech-Driven Avatar-Mediated Three-Party Telepresence System: Design and Evaluation. *PRESENCE: Virtual and Augmented Reality* 29 (2020), 113–139.

[20] Aobo Jin, Qixin Deng, and Zhigang Deng. 2022. S2M-Net: Speech Driven Three-party Conversational Motion Synthesis Networks. In *Proceedings of the 15th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 2:1–2:10.

[21] Aobo Jin, Qixin Deng, Yuting Zhang, and Zhigang Deng. 2019. A deep learning-based model for head and eye motion generation in three-party conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2 (2019), 1–19.

[22] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–30.

[23] Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto. 2010. Turn-alignment using eye-gaze and speech in conversational interaction. In *Eleventh Annual Conference of the International Speech Communication Association*.

[24] Kristiina Jokinen, Masafumi Nishida, and Seiichi Yamamoto. 2009. Eye-gaze experiments for conversation monitoring. In *Proceedings of the 3rd international universal communication symposium*. 303–308.

[25] Tatsuya Kawahara, Takuma Iwatate, and Katsuya Takanashi. 2012. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*.

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

[27] 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. *Language and Speech* 41, 3-4 (1998), 295–321.

[28] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *Proceeding of 2019 International Conference on Multimodal Interaction*. 226–234.

[29] Gene H. Lerner. 2003. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society* 32, 2 (2003), 177–201.

[30] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics* 11 (2023), 250–266.

[31] Kazuhiro Otsuka. 2011. Multimodal conversation scene analysis for understanding people's communicative behaviors in face-to-face meetings. In *Proceedings of Symposium on Human Interface 2011*. Springer, 171–179.

[32] Lawrence Rabiner and Biinghwang Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP magazine* 3, 1 (1986), 4–16.

[33] Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 67–74.

[34] Clarissa Weiss. 2018. When gaze-selected next speakers do not take the turn. *Journal of Pragmatics* 133 (2018), 28–44.

[35] Elisabeth Zima, Clarissa Weiß, and Geert Brône. 2019. Gaze and overlap resolution in triadic interactions. *Journal of Pragmatics* 140 (2019), 49–69.