Online Multimodal End-of-Turn Prediction for Three-party Conversations

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

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

ABSTRACT

Predicting end-of-turn in multiparty conversations is crucial to increase the usability and natural flow of spoken dialogue systems, offering substantial enhancements to conversational agents. We present a novel window-based method to predict end-of-turn moments in real-time in multiparty conversations, by leveraging the capabilities of cutting-edge pre-trained language models (PLMs) and recurrent neural networks (RNN). Our method fuses the distilBERT language model with a Gated Recurrent Unit (GRU) to accurately predict end-of-turn points in an online fashion. Our approach can significantly outperform conventional Inter-Pausal Unit (IPU)-based prediction methods that often overlook the nuances of overlap and interruption during dynamic conversations. Potential applications of this study are significant, particularly in the domains of virtual agents and human-robot interactions. Our accurate online end-of-turn prediction model can be facilitated to enhance the user experience in these applications, making them more natural and seamlessly integrated into real-world conversations.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Collaborative interaction**.

KEYWORDS

multi-party conversations; multimodal interaction; non-verbal gesture; human-human interaction; turn prediction

ACM Reference Format:
Meng-Chen Lee and Zhigang Deng. 2024. Online Multimodal End-of-Turn Prediction for Three-party Conversations. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3678957.3685742

1 INTRODUCTION

There has been a growing fascination with comprehending and modeling the intricate interplay and correlations among various modes of multimodal multiparty communication in recent years. These encompass facial expressions, gaze, hand gestures, and linguistic content. This growing interest has captured the attention of researchers in diverse fields, including communication, human-computer interaction, graphics, robotics, and the expansive realm of multimodal interaction communities [18, 25, 33, 46].

In real world, characterized by the growing prevalence of conversational human-computer interactions, accurate prediction of end-of-turn points in multiparty conversations has assumed central importance [2]. The intricacies of turn-taking in dialogues have also attracted the attention of scholars in various disciplines [45]. The capacity to foresee when a participant will wrap up their statement or seek a response essentially shapes the natural flow and rhythm of a conversation.

Predicting the end of a turn in a multiparty conversation is not just a technical detail; it has profound implications for the usability and naturalness of spoken dialog systems [2]. Applying such predictions can significantly improve the overall effectiveness of conversational agents [32]: deciding when to interject, when to wait, or when to facilitate a smooth exchange of ideas. Multiparty interactions, as the name suggests, involve multiple participants participating in a conversation. However, it is not only the number of participants that makes these interactions intricate; it is the nuanced social dynamics that come into play [13]. In such dialogues, a dialog system must navigate the challenge of detecting the conclusion of the utterances of the participants while deciphering the complex web of social signals [10]. This involves determining the intended audience of an utterance, understanding the current state of the conversation, and predicting when a participant might interject or remain silent.

To tackle these complex challenges, in this paper we present a novel methodology to predict end-of-turn occurrences including turn-taking, interruption, and speech-overlapping, in multiparty conversations. This comprehensive approach considers various factors, including the evolving state of the conversation, participants' gaze, prosody features, and textual cues. Significantly, it also accounts for uncertainties originating from computational delays inherent in dialog systems.

Contributions. The main contributions of this work include:

- We introduce a holistic PLM-GRU multimodal fusion system designed to achieve accurate end-of-turn prediction in multiparty conversations. Our approach stands as the first continuous model supporting end-of-turn prediction in the context of multi-party conversations.
- This work addresses a critical gap by introducing annotation methods that encompass all categories of end-of-turn in multiparty conversations, including often overlooked aspects such as interruption and overlapping.

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685742

2 RELATED WORK

IPU-based Models. Some previous studies have adopted an IPU (Inter-Pausal Unit) [31] based approach to predict the end of a turn in dialogues. Figure 1 (left) shows the paradigm of IPU-based prediction. Specifically, in the data we first identify any silences lasting longer than 200 milliseconds and then assign the utterance preceding each such pause as an IPU.

Early approaches to turn-taking detection include rule-based methods, such as the semantic parser [1] and data-driven techniques like decision trees [42]. These models considered various cues, such as semantics, syntax, dialogue state, and prosody, to classify pauses longer than 750 ms. Later works [36, 43] explored similar models but with shorter silence thresholds. In recent years, Ishii and colleagues [22–24] conducted a series of research studies to understand of end-of-turn prediction in conversation. Their works highlight the importance of analyzing intricate gaze transition patterns [24] as a more effective approach compared to relying solely on a single line of gaze, incorporating human gaze and respiratory behavior to forecast turn changes [22], and incorporating head movements toward the end of an utterance [23]. Recently, Lee et al. [34] proposed an IPU-based computational model to predict turns in multiparty conversations, based on a concept called Relatively Engagement Level (REL).

Although these methods have provided valuable insights, they are not without drawbacks. IPU-based prediction often suffers from a noticeable delay in making predictions, which hinders the flow of real-time interactions [32]. Additionally, these methods may struggle to determine the exact moment of interruption in a conversation, which could lead to less precise turn-taking decisions [32].

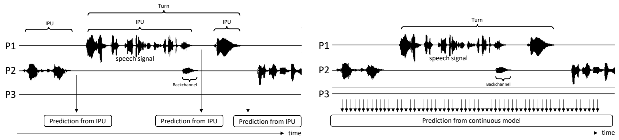

Figure 1: (left) IPU-based models. (right) Continuous models.

Continuous Models. Recognizing the constraints of IPU-based prediction, researchers have turned their attention to alternative methodologies, particularly those centered on continuous prediction techniques [19, 32, 39, 44]. Figure 1 (right) illustrates the framework of a continuous model designed to predict end-of-turn events continuously (e.g., at 500ms intervals). These approaches are geared toward achieving real-time turn prediction, effectively mitigating the latency challenges inherent in IPU-based methods.

Skantze [44] presented a comprehensive continuous turn-taking model that uses various prosody features. They used an LSTM architecture to predict speech activity for both speakers within a future 3-second window. Roddy et al. [38] adopted a similar approach by exploring various characteristics of speech conducive to prediction, and introduced a novel multiscale LSTM architecture that integrates different modalities at different timescales [39]. These continuous prediction models have mainly focused on *two-party* dialogues. They excel at predicting the end of a turn within a dyadic conversation but face challenges when extended to multiparty interactions. The intricacies of multiparty conversations, where multiple participants can engage concurrently or wait for their turn, introduce additional complexities Note that the previous study in [7] that addressed multimodal end-of-turn prediction in multiparty meetings has encountered significant difficulties in achieving successful results, underscoring the inherent challenges associated with this topic.

Since evaluations in previous studies were often conducted on distinct internal datasets; therefore, direct comparisons of numerical values should be approached with caution, as they may lack the necessary contextual relevance. The critical insight from these prior works is in demonstrating the viability of leveraging various verbal and non-verbal cues to predict turn changes and the next speaker during dynamic conversations spanning multiple participants.

End-of-turn Cues. Numerous studies have focused on identifying cues at the end of IPUs. These cues serve as vital indicators for listeners to distinguish turn-taking from turn-keeping. Duncan and colleagues [14] systematically examined these cues within face-to-face American English dyadic conversations. They discerned several multi-modal cues denoting turn-completion, encompassing phrase-final intonation, cessation of hand gesticulation, and completion of a grammatical clause. Subsequent studies have expanded on these findings, using larger data sets, automated methodologies, and robust statistical analyses [8, 17, 21, 31]. These investigations generally affirm the additive nature of turn-taking cues despite significant redundancy.

Verbal Cues. As dialogues progress, linguistic elements encompassing spoken words and their semantic and pragmatic subtleties assume a central role in orchestrating turn transitions. The fulfillment of a syntactic unit serves as a prerequisite for designating a turn as "completed." For example, when confronted with an unfinished phrase such as "I would like to order a...," the listener is compelled to await the sentence's conclusion. The predictability of syntactic units, coupled with the ability to anticipate their culmination within the conversational context, elucidates the precise coordination of turn-taking [15, 40].

Prosody. The role of prosody has attracted considerable attention in conversation analysis. Prosody encompasses nonverbal speech elements such as intonation, loudness, speaking rate, and timbre, serving various functions in conversation, including marking prominence, disambiguating syntax, conveying attitudes, expressing uncertainty, and signaling topic shifts [47]. Across different languages, studies have revealed that level intonation, occurring near the end of an IPU within the speaker's fundamental frequency range, tends to function as a turn-holding cue. This observation holds for different languages [14, 17, 31]. In English dialogues, Gravano and Hirschberg [17] observed that speakers tend to lower their voices when approaching potential turn boundaries, while turn-internal pauses exhibit greater intensity. Similar patterns were found in Japanese dialogues [31], where low or decreasing energy was associated with turn change. Compared to verbal cues, prosody may play a more prominent role from the perspective of conversational systems.

Gaze. The role of gaze in communication is undeniably crucial. Previous research has revealed that when engaging in conversation, speakers tend to direct their gaze toward their conversation partners more frequently towards the conclusion of their speaking turns rather than at the outset [9]. Direct gaze, where speakers maintain eye contact, can signal their readiness to yield the floor to their interlocutors, signaling end-of-turn transitions. On the other hand, averted gaze may convey the desire to retain control over the ongoing conversation, subtly expressing the desire to maintain the current conversation turn [20].

3 DATA ACQUISITION

For this study, we developed a comprehensive conversational behavior dataset involving three participants. As outlined below, our approach involved an in-house built data acquisition system, combining advanced technologies such as optical motion capture, eye trackers, and high-definition microphones. Figure 2 offers a snapshot of the actual data acquisition process in action.

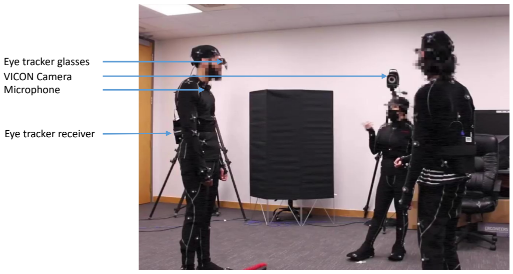

Figure 2: A snapshot of our three-party conversational motion data acquisition experiment

Motion Capture. A state of the art optical motion capture system, comprising eight strategically positioned cameras, was meticulously installed within a controlled laboratory setting. The system consisted of four upper and four lower cameras, ensuring comprehensive coverage to capture intricate details. To capture the complete spectrum of body movements, each participant was outfitted with a motion capture (mocap) suit embellished with optical markers, meticulously recording their motions spanning head, hands, torso, and lower body actions (refer to Figure 2).

Gaze and Audio. We employed three Ergoneers Dikablis Glass 3 eye trackers for eye tracking. We evaluated their impact on participants’ natural conversational behavior, including their gazes, by soliciting their feedback. The consensus among the participants was that the eye trackers had a minimal influence on their behavior. Notably, the eye trackers used in our research closely resembled conventional eyeglasses in terms of size and weight, ensuring its unobtrusive presence during our data capture experiments. Additionally, we equipped each participant with a wireless microphone discreetly attached to their necklines for audio recording.

Capture Design. We recruited 21 volunteers from a university campus for our data capture experiment, randomly dividing them into seven groups each of which comprises three participants. Note that the 21 participants were unfamiliar with each other prior to our experiment. Among them, 12 were men and 9 were women, 20 to 30 years old. Most of the participants (70%) had a major in computer science and were native English speakers. In our study, we recorded 3-5 sessions of three-party conversations for each group. The recorded session lengths for the seven groups were as follows: 46 minutes and 18 seconds (46’18”), 44’47”, 46’17”, 42’49”, 46’31”, 44’27”, and 48’19”, totaling 319 minutes and 28 seconds of recorded data.

During the data collection process, we instructed the three participants in each conversation to maintain fixed positions, forming an equilateral triangle with approximately 1 meter of separation between them. This spatial configuration for interlocutors aligns with similar setups used in previous studies that focused on three-party conversations, such as [4, 12, 26–28], and has similarities with the setup in [37], which ensured that all five participants formed a regular pentagon. Participants were allowed to discuss topics of interest during the data collection process. To synchronize the start times of different data sources, a clapperboard was used. As a result, all collected data, including 3D motion capture, eye tracking, and speech data, were temporally aligned and subsequently downsampled to a rate of 60 frames per second.

4 FEATURES EXTRACTION

Interlocutor States. We derived interlocutor state feature vectors from each frame through the following procedure. Initially, audio content was transcribed into texts using Google’s speech-to-text technology. Subsequently, a manual review and necessary corrections were performed. To detect instances of back-channeling utterances, we adopted the criteria established in [29], encompassing expressions such as “Yeah," “Um," “Cool," “Oh," “Okay," "Right," and "Uh." These expressions were then marked within the transcriptions. Consequently, for every frame, a 1×3 vector with ternary variables was generated. This vector, referred to as the *Interlocutor State Vector* (ISV), was created to depict the conversational states of the three interlocutors involved: the i-th element’s value (i ranging from 1 to 3) signifies whether the i-th interlocutor was actively speaking (=2), engaging in back-channeling (=1), or experiencing silence (i.e., neither speaking nor back-channeling) (=0).

Prosody Features. We extracted prosody features, specifically pitch and intensity, from the aforementioned audio data. In each frame, which spans approximately 16 milliseconds, we generated a distinct feature vector to serve as input for our neural networks. Our focus is primarily on capturing momentary prosodic characteristics, with an emphasis on the prevailing pitch level and intensity, both of which are highly interpretable. Consequently, for every frame in the audio data involving three interlocutors, we generated a numerical vector of dimensions 1 × 6.

Gaze Targets. Drawing inspiration from prior research [27], we extended our exploration to include the computation of an interlocutor’s Direction-of-Focus (DFoc), achieved through the integration of both torso-head orientation and eye gaze direction vector (V_E). Specifically, given the representations R_{hips}, R_S, and R_H for the 3D rotations of the hip, spine, and head, and using G_h, O_S, O_H, and O_E to symbolize the global positioning of the hip along with the offsets of the spine, head, and eyes, respectively, we formulated the following equation to derive DFoc. Notably, to ensure accuracy, we addressed anomalies and filled in the gaps by employing shape-preserving cubic interpolations in piecewise order for all instances of V_E.

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

Incorporating the aforementioned DFoc calculation, we performed an additional step of generating a higher order feature called *Focus of Attention* (FoA). This distinctive attribute serves to indicate which of the other interlocutors an individual interlocutor is directing their attention to. To this end, we devised a 1×3 vector with ternary variables for every frame, known as the *Gaze Target Vector* (GTV). This vector succinctly encapsulates the FoA for the three interlocutors. Specifically, the value of its i-th component corresponds to the index of the interlocutor who captures the attention of the i-th interlocutor during that specific frame.

4.1 Gestural Backchanneling

Besides verbal backchanneling, nonverbal gestures can indicate listeners' attention and sentiments toward speech. Our study focuses on backchanneling gestures in American culture, considering participants from the United States. We identify a gestural backchaneling response when a listener nods or shakes his/her head. We use head movement data, including roll, pitch, and yaw angles, and employ an algorithm inspired by previous work for detection.

Following the concept from [30], we classified frames as stable, extreme, or transient, based on the change of the yaw angle. Stable frames indicate minimal headshake, extreme frames indicate significant movement, and transient frames indicate ongoing head movement. To detect headshaking, we require at least two extreme frames with a yaw angle difference exceeding 3 degrees between them between two stable frames. To this end, for each frame in a data observation, we create a 1×6 gestural back-channeling vector by combining two 1×3 binary-valued sub-vectors. The first sub-vector indicates whether each participant is shaking their head, and the second sub-vector represents nodding. Gestural backchanneling responses are exclusive to listeners, signifying their attention to ongoing speech. The current speaker does not contribute to these responses, even if our algorithm detects the corresponding head motions.

5 METHODOLOGY

Label Annotations. Compared to IPU-based annotations, which typically annotate at the end of an IPU, our approach provides annotations at much finer temporal granularity. Specifically, we set the step size to 100 milliseconds, that is, we predict a label every 100 milliseconds. In our work, we categorize the end-of-turn event into three distinct sub-categories: "interruption," "overlapping," and "turn-taking." From a technical standpoint, our objective involves predicting the moment when the speaker finishes his/her speech. However, framing it this way would result in a significant class imbalance within the labels. To mitigate this imbalance and forecast possible ends, we adopt a strategy aligned with previous research [32], where we classify an end-of-turn instance as a time point indicating that the end of the speaker's turn will occur within the next 500 milliseconds.

Figure 3 illustrates all three sub-categories. (i) *Interruption*, also called "barge-in," is a prevalent phenomenon frequently occurring in fast-paced dialogues, where the next speaker anticipates the conclusion of the current speaker's turn and initiates his/her response before the current speaker finishes. This phenomenon is particularly common in rapid exchanges. Note that apart from the 500 milliseconds before the exact end-of-turn, we have also designated the intersection of interruption as a positive instance. (ii) *Overlapping*, on the other hand, is common in multiparty conversations, characterized by simultaneous contributions from multiple interlocutors following extended silences, resulting in segments of speech that overlap. It is important to note that back-channeling, while a common feature, does not signal the end of a speaker's turn, as it typically does not prompt the current speaker to yield the floor. Overall, the label "overlapping" not only denotes the turn before overlap but also specifies when the overlapping is expected to conclude. (iii) Lastly, *turn-taking*, a broader category within traditional IPU-based annotations, refers to instances where a shift in the conversational floorholder occurs, typically defined by a change before and after a silence period lasting more than 200 milliseconds.

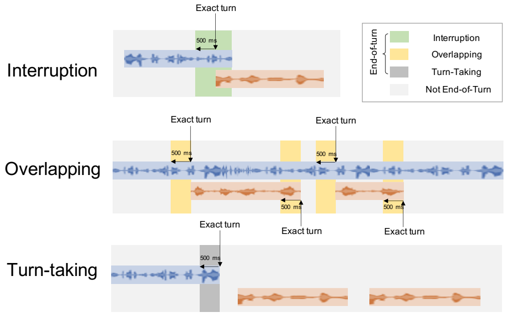

Figure 3: Illustration of three categories for end-of-turn: interruption, overlapping, and turn-taking.

Our approach not only provides a more precise and comprehensive method for predicting end-of-turn events, covering all types of turn changes, but also overcomes a major limitation of traditional IPU-based approaches, which frequently experience delays in realtime prediction. Furthermore, compared to conventional IPU-based annotation, our approach significantly enhances the dataset, elevating the label count from 4,248 (if IPU-based approaches are used) to 167,042. This substantial dataset expansion proves especially advantageous for training deep learning models. Instead of adhering to a binary classification paradigm ("end-of-turn" vs. "not end-of-turn"), we design a four-fold classification system by considering the following distinct categories: "Interruption", "Overlapping", "Turn-taking", and "Not end-of-turn." The statistics of these sub-categories in our dataset are presented in Table 1.

Table 1: Statistics of end-of-turn sub-categories in our dataset

Categories	Not End-of-Turn	End-of-Turn		
		Interruption	Overlapping	Turn-Taking
Instances	134,715	18,509	6,992	6,826
Ratio	80.6%	11.1%	4.2%	4.1%

5.1 Window-based Prediction Model

Our primary objective is the design of an efficient approach for online (or even real-time) end-of-turn prediction in multiparty conversations. To accomplish this, we design a window-based continuous model tailored for handling multi-modal features. For text-based features, we choose to use DistilBERT [41] as an alternative to the larger BERT model [11] due to its smaller size and faster inference capabilities. Instead of the commonly used LSTM model for other time series features, we opt for GRU (Gated Recurrent Unit) as a more efficient alternative, benefiting from its reduced number of gates and parameters, leading to faster processing while maintaining comparable performance [6].

We empirically set the window size to be 500 milliseconds (ms), and set the step size to 100 ms, which implies that our approach makes predictions for the end-of-turn moment every 100 ms using the features within a 500 ms window. However, to extract text features, we extend the window to include texts from 100 seconds before the exact prediction time. This extension allows us to capture a more comprehensive history of the dialogue content. We empirically choose 100 seconds based on our observation: Selecting shorter time frames for text features could result in higher probabilities of missing words or encountering silence. Such instances are not conducive to our system's ability to comprehend the contributions and mechanisms between interlocutors.

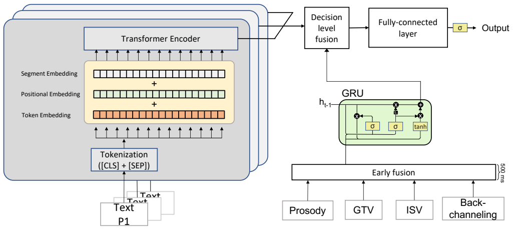

Figure 4: The architecture of our window-based prediction model

Figure 4 provides a pipeline overview of our approach, in which we employ DistilBERT for the extraction of semantic features without fine-tuning its parameters during training. Our rationale for this decision lies in the multi-modal nature of our approach, as fine-tuning solely on text data could potentially lead to overfitting. The input texts are transformed into embedding vectors that represent a word in the sequence. DistilBERT with its transformer encoder captures contextual information for each word. These contextual embeddings are consolidated into a single vector, effectively encapsulating the conversation's semantic information. By avoiding fine-tuning, we aim to guard against overfitting, recognizing the richness of features introduced by including multiple modalities, such as visual and gestural cues, which collectively enhance our method's adaptability and generalizability across diverse data sources.

In our approach, the treatment of additional features entails a well-defined process. Initially, we employed a Min-Max Scaler to normalize numerical features, ensuring that they conform to a consistent and manageable range. Based on this, we embrace the concept of early fusion, in which information from various modalities is integrated at the initial stage of processing. Specifically, we employed a range of fusion techniques to enhance the integration of modalities. Initially, we adopted a straightforward approach by concatenating all modalities into a single vector. Subsequently, drawing inspiration from [16], we implemented a more intricate strategy. This involved combining all representations through addition and L2 normalization. Let us consider a fused representation, denoted as $\mathcal{F}_n$, with a total of n frames. We have four modalities: M_n^I for ISV (Interlocutor State Vector), M_n^P for prosody features, M_n^G for GTV (Gaze Target Vector), and M_n^B for gestural back-channeling, each corresponding to a frame. We employ separate frame embedding layers for each modality, namely $\mathcal{E}_I$, $\mathcal{E}_P$, $\mathcal{E}_G$, and $\mathcal{E}_B$. The fused representation is calculated using the following equation:

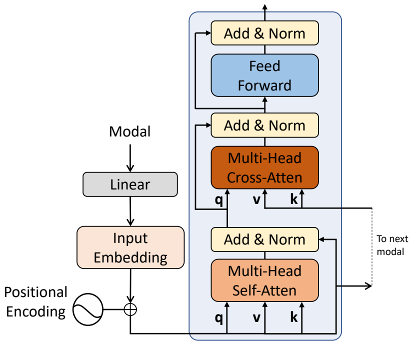

Figure 5: The architecture of cross-modal fusion

$$\mathcal{F}_n = \frac{\mathcal{E}_I(M_n^I) + \mathcal{E}_P(M_n^P) + \mathcal{E}_G(M_n^G) + \mathcal{E}_B(M_n^B)}{\left\| \mathcal{E}_I(M_n^I) + \mathcal{E}_P(M_n^P) + \mathcal{E}_G(M_n^G) + \mathcal{E}_B(M_n^B)) \right\|_2}. \qquad (2)$$

This method, referred to as the 'addition-embedder,' derives its name from its core principle of summing up the contributions from various modalities. Moving on to a third technique shown in Figure 5, we introduced cross-modal attention, further enriching the fusion process. These diverse approaches were systematically investigated to provide a comprehensive evaluation of their influence on both the overall model performance and fusion efficacy. Comparative results are presented in Section 6.5. The fusion process precedes the input into a one-layer GRU network, configured with a hidden size parameter set to 256. The utilization of the GRU architecture, renowned for its update and reset gates, is essential in our design. This architecture excels in its capacity to capture intricate sequential data patterns, making it an effective tool for our model's overall functionality. By integrating numerical and categorical features through early fusion and processing them with the GRU network, our model gains the ability to uncover nuanced patterns and dependencies within the data, contributing to its robust performance.

Following the feature extraction phase, we seamlessly integrate the features derived from the language model with the ultimate hidden states originating from the GRU layer. These combined features undergo further refinement as they traverse a fully connected layer, ultimately culminating in generating probabilities facilitated by applying a sigmoid activation function. To fortify the model's generalization capacity and preempt overfitting, we prudently implement a dropout rate of 0.1 during the training process, coupled with the deployment of an early stop technique for vigilant overfitting monitoring and prevention. Additionally, we employ Binary Cross Entropy Loss (BCELoss) as our designated loss function. Our initial learning rate, thoughtfully set at 0.001, is complemented by the use of the Adam optimizer, which guides the training process.

Table 2: Distribution of IPU instances.

Categories	Not End-of-Turn	End-of-Turn (Turn-Taking)
Instances	5,975	1,105
Ratio	84.4%	15.6%

5.2 IPU-based Prediction Model Baseline

The IPU-based model shares the same architecture and features as the window-based model. The primary distinctions between the two models lie in the annotation method and the length of each instance, as discussed in Section 5. First, the process involves segmenting only the clauses preceding a 200-millisecond silence, resulting in relatively smaller instances. Second, unlike the window-based method that employs a fixed window size for each sample, the extracted IPUs exhibited a range of durations, with the shortest IPU spanning 32 frames (equivalent to 0.53 seconds) and the longest extending to 1,755 frames (29.25 seconds). On average, IPUs had a length of 201 frames (3.35 seconds). For consistency in model training, we applied padding to all IPUs, extending them to the maximum length of 1,755 frames.

From the identified IPUs, we further categorize them into two different types: those marked as end-of-turn (indicating a turn-taking event resulting in a speaker change) and those categorized as not end-of-turn, indicating a scenario where the speaker remains the same. In total, we extracted 7,080 IPUs from seven groups of three-party conversations in our dataset. Among these, 5,975 IPUs corresponded to turn-taking instances, while 1,105 IPUs represented scenarios where the turn did not conclude. It is crucial to note that the IPUs were carefully processed to ensure no temporal overlap, since any instances of overlapping speech segments were discarded. A detailed breakdown of the extracted IPUs is provided in Table 2.

6 RESULTS AND EVALUATIONS

To gauge the effectiveness of our model, we meticulously designed and executed a rigorous evaluation methodology that ensures the reliability and completeness of our assessment. This methodology is pivotal in ascertaining our model's performance under diverse scenarios and across various conversational groups. First, we employed a 10-fold cross-validation procedure on our training dataset. This approach is renowned for providing a robust evaluation by systematically partitioning the data into ten equally sized subsets. In this process, nine subsets are used for training, while the remaining one is reserved for testing. This cycle repeats ten times, each time using a different subset as the test set. By aggregating the results from these iterations, we calculated key performance metrics, including precision, recall, and F-measure, using a macro-average approach. This method ensures that every class or conversational behavior receives equal weighting in the evaluation, offering a comprehensive view of our model's overall performance. Note that, in all presented results, only those in Section 6.1 involve 4-class classifications (Table 4). In contrast, all other results are based on binary classifications.

Table 3: Quantitative comparison of turn prediction between our window-based prediction model and the IPU-based prediction model. The bottom table displays the results of a two-sample t-test.

	Metrics		
Model	Precision	Recall	F-Measure
IPU-based model	0.513	0.521	0.406
Window-based model	0.874	0.872	0.873

Metrics	*t*	*p-value*
Precision	112.164	1.505e-27 **
Recall	114.730	3.065e-27 **
F1	154.145	1.514e-29 **

In Table 3, we provide a detailed quantitative comparison between our window-based prediction model and the conventional IPU-based baseline model, in the context of predicting end-of-turn events in multiparty conversations. Besides, we applied a two-sample t-test between the two models to rigorously assess and quantify the statistical significance of observed differences in their performances. This evaluation serves as a critical benchmark for evaluating the effectiveness of our proposed approach. It is imperative to recognize that the IPU-based model shares the same architecture as the window-based model but is trained on a relatively smaller dataset, constrained by the IPU definition that segments based only on the clause preceding a 200-millisecond silence (refer to Section 5). However, during testing, both models were evaluated within the same window-based framework, predicting the results every 100 milliseconds. This way ensures a fair comparison, particularly in the context of real-world applications. As revealed in Table 3, our window-based approach demonstrates a remarkable superiority over the IPU-based model in terms of all the metrics studied, including precision, recall, and the F-measure. Specifically, our model achieves a precision score of 0.874, indicating a high accuracy in correctly identifying actual end-of-turn instances. The recall rate of 0.872 highlights our model's capability to effectively capture a substantial portion of the actual end-of-turn events in multiparty conversations. The F-measure, which balances precision and recall, achieves an impressive value of 0.873, underlining the well-rounded performance of our window-based prediction model. Moreover, the obtained p-values: 1.505e-25, 2.633e-27, and 3.015e-30, indicate strong statistical significances of our window-based model over the IPU-based model in all metrics evaluated.

Table 4: Performance metrics for the four-class categorization, highlighting individual category performance.

	Metrics		
Class	Precision	Recall	F-Measure
Not End-of-Turn	0.957	0.950	0.954
Interruption	0.782	0.849	0.814
Overlapping	0.687	0.607	0.645
Turn-Taking	0.606	0.613	0.610
Macro-Average	0.758	0.755	0.756

6.1 End-of-turn Sub-category Prediction

In Table 3, end-of-turn events cover three subcategories (i.e., turn-taking, interruption, and overlapping). A binary classification method would be insufficient to capture the nuanced intricacies inherent in conversational dynamics, including interruption, overlapping, and standard turn-taking. In recognition of this constraint, we implemented a four-class categorization framework. In this framework, we adjusted the activation function to the softmax function and the loss function to CrossEntropyLoss. The results of this modification are presented in Table 4.

The results in Table 4 provide a detailed breakdown of our model's performance for each category. Notably, we observe variations in precision, recall, and F1-score across categories. For example, our model exhibits higher precision (0.782) and recall (0.849) for the Interruption category, indicating its ability to identify instances of interruption events correctly. However, when dealing with other conversational dynamics, such as overlapping and turn-taking, our model had sub-optimal performances, as evidenced by slightly lower precision and recall values. The macro-average performance metrics, which consider our model's effectiveness across all classes, comprehensively evaluate its ability to handle diverse conversational dynamics. In this case, the precision is 0.758, the recall is 0.755, and the F1 score is 0.750. These values provide a holistic assessment of our model's performance, highlighting the challenges in distinguishing the sub-categories of end-of-turn.

Table 5: Quantitative cross-group validations by our model

		Metrics	
Test Group	Precision	Recall	F-Measure
1	0.835	0.776	0.801
2	0.812	0.775	0.791
3	0.887	0.844	0.863
4	0.844	0.775	0.802
5	0.871	0.813	0.837
6	0.873	0.882	0.878
7	0.886	0.822	0.849

6.2 Cross-group Validation

We also conducted cross-group validation in seven distinct conversational groups, adhering to a structured 5-1-1 format. This arrangement designated five groups for training, one for validation, and one for testing. This systematic cross-group validation method yielded valuable insights into our model's generalizability and its robustness across a spectrum of conversational behaviors. Notably, it safeguarded against potential bias towards any specific subset of the data, thereby bolstering the reliability of our research findings. Table 5 presents the quantitative results of our cross-group validation. The results shown in Table 5 underscore the robustness of our model, indicating its capacity for effective generalization to new multiparty conversations involving novel participants, compared to the results presented in Table 3.

6.3 Ablation Study

To gain profound insights into the integral components and features that influence our model's performance, we conducted an ablation study. Figure 6 (left) presents the results of this study, which mainly focused on assessing the impact of various input features on our model's predictive capabilities regarding end-of-turn instances. The "All features" configuration, encompassing a fusion of gestural back-channeling, prosody, gaze target (GTV), interlocutor state (ISV), and textual information, yielded the highest precision (0.874), recall (0.872), and F-measure (0.873). This outcome underscores the critical importance of an integrated approach involving all these features to achieve optimal predictive performance. Our results consistently demonstrate that the exclusion of specific features leads to a noticeable performance decline. For example, the removal of GTV or prosody resulted in a reduction in precision, recall, and F-measure, highlighting the pivotal role played by gaze-related and prosodic cues in end-of-turn prediction. Similarly, the omission of gestural back-channeling or textual information led to a decrease in performance metrics, indicating the greater importance of contextual cues and textual content in improving model accuracy. Notably, the exclusion of the ISV feature yielded the poorest performance when compared to other single-feature-exclusion approaches. This observation firmly underscores the pivotal role played by the ISV feature among all the considered features, as the ISV feature directly provides invaluable insights into conversational dynamics and status. Remarkably, the "Text Only" configuration exhibited the least favorable performance. This finding accentuates that relying solely on textual information may prove insufficient in providing the necessary cues for accurate end-of-turn prediction in multi-party conversational contexts, reinforcing the critical need for a multimodal approach.

		Metrics	
Features	Precision	Recall	F-Measure
All features	0.874	0.872	0.873
No GTV	0.826	0.770	0.793
No Prosody	0.823	0.749	0.777
No GBack	0.815	0.744	0.771
No Text	0.816	0.735	0.765
No ISV	0.749	0.533	0.516
Text Only	0.586	0.644	0.535

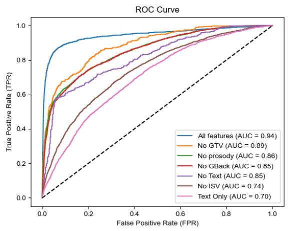

Figure 6: (left) Ablation study results by our model. Here ISV refers to the Interlocutor State Vector, GTV refers to the Gaze Target Vector, GBack refers to Gestural Back-Channeling; (right) ROC curves for different feature combinations

The ROC curves in Figure 6 (right) further show the discriminative powers of these feature combinations. The area under the ROC curve (AUC-ROC) provides a quantitative measure of our model's ability to distinguish between true positives and false positives. The "All features" configuration achieved the highest AUC-ROC of 0.94, indicating strong discriminatory capabilities. Notably, the "No GTV," "No Prosody," "No GBack," and "No Text" configurations also displayed competitive AUC-ROC values of 0.89, 0.86, 0.85, and 0.85, respectively, suggesting that while these features contribute to our model significantly, their exclusion does not severely affect our model's overall discriminative ability.

In summary, these results underscore the importance of a multimodal approach, incorporating linguistic, prosodic, gaze-related, and contextual information for robust end-of-turn prediction in multiparty conversations. While individual feature exclusions have a negative impact on performance, ISV plays a particularly important role.

Table 6: Runtime statistics of our approach for different architecture combinations

Language model	RNN	Inference time (ms)	Precision	Recall	F-Measure
DistilBERT	RNN	1.6 ± 0.20	0.816	0.724	0.756
	GRU (selected)	2 ± 0.06	0.874	0.872	0.873
	LSTM	2.12 ± 0.15	0.886	0.849	0.866
BERT	RNN	7.54 ± 0.86	0.810	0.728	0.758
	GRU	8.04 ± 0.15	0.874	0.882	0.877
	LSTM	8.22 ± 0.21	0.887	0.844	0.863
RoBERTa	RNN	7.92 ± 1.06	0.844	0.888	0.815
	GRU	8.29 ± 0.19	0.893	0.879	0.886
	LSTM	8.92 ± 1.07	0.898	0.883	0.890
GPT-3	RNN	355.82 ± 53.08	0.853	0.733	0.771
	GRU	355.86 ± 53.07	0.903	0.843	0.872
	LSTM	355.91 ± 53.07	0.918	0.902	0.910
PaLM	RNN	503.89 ± 30.85	0.826	0.728	0.762
	GRU	503.92 ± 30.83	0.893	0.854	0.873
	LSTM	504.03 ± 30.85	0.916	0.897	0.907

6.4 Comparisons with Alternative Models

To demonstrate both the effectiveness and the real-time prediction capabilities of our model, we conducted an in-depth analysis of its runtime performance. This analysis extends beyond our model's predictive performance and delves into its computational efficiency compared to other prominent large language models (LLMs) and RNNs. In our evaluation, we have benchmarked our model against some of the most renowned large language models, including both open-sourced (BERT, RoBERTa [35]) and closed-sourced (OpenAI GPT-3 [3], and Google PaLM [5]) LLMs. These comparisons offer insights into how our model performs in terms of runtime efficiency when compared to heavyweight language models that are known for their language understanding capabilities. Additionally, we have assessed our model's runtime performance against a more complex recurrent neural network architecture, specifically simple RNN and LSTM (Long Short-Term Memory) models. By doing so, we aim to provide a comprehensive view of our model's computational efficiency concerning different neural network architectures. In this experiment, we conducted tests on open source LLMs using our system, running on a standard desktop computer with the following specifications: an Intel i9-10850K CPU operating at 3.6 GHz, 40GB of memory, and an NVIDIA GeForce RTX 3070 GPU. For closed-source LLMs, we chose a local implementation but accessed the LLM through cloud servers using provided APIs, resulting in significant delays. This delay is noteworthy, especially in real-time applications, where the system's inability to achieve real-time implementation becomes a considerable limitation. The inference times and performance metrics for various combinations of LLMs and RNNs are reported in Table 6. Notably, the combination of GPT-3 and LSTM exhibited the highest performance; however, it is crucial to emphasize that this configuration led to increased inference time. The choice of configuration should be influenced by the user's specific needs and the trade-off between inference time and performance, taking into account the constraints imposed by available hardware. Future research can explore different setups to align with their particular research goals and hardware limitations.

Table 7: Runtime statistics for different fusion methods

Fusion method	Inference time (ms)	Precision	Recall	F-Measure
Simple Concate.	2 ± 0.06	0.874	0.872	0.873
Addition-embedder	2.37 ± 0.06	0.875	0.854	0.864
Cross attention	2.40 ± 0.09	0.873	0.856	0.864

6.5 Comparison of Early Fusion Methods

As described in Section 5.2, in order to harness the full potential of multi-modal capabilities, we implemented a range of early fusion techniques. Table 6 presents a detailed breakdown of runtime and performance metrics for each method. Despite the simplicity of the "Simple Concate." method, it consistently emerges as the most effective approach among the explored fusion techniques. The efficient inference time of 2 ± 0.06 ms, coupled with superior precision, recall, and F-measure values, underscores the robust performance of this straightforward fusion method. While more complex techniques like "Addition-embedder" and "Cross attention" introduce nuanced variations in trade-offs and performance metrics, the straightforward nature of "Simple Concate." showcases its remarkable efficacy in optimizing multi-modal integration for the given task. This underscores the importance of considering both simplicity and performance when selecting an early fusion technique, with "Simple Concate." standing out as a pragmatic and high-performing choice in this context.

7 CONCLUSION AND DISCUSSION

In this paper, we present a novel deep window-based approach for predicting end-of-turn events, including the moment of interruption and overlapping, within three-party conversations. Our key contribution lies in the use of an PLM-GRU-based multimodal model with early fusion designed to accurately predict the end-of-turn. Extensive experiments have demonstrated the effectiveness of our approach in generating accurate predictions for such complex multiparty conversations.

Although our current approach marks a significant step in this research direction, it is essential to acknowledge its limitations. First, we employ relatively straightforward fusion techniques for handling multimodal inputs, rather than fully harnessing the potent attention mechanism of the transformer model. Future work can be explored to design more intricate architectures, such as cross-attention mechanisms, to integrate each modality's information seamlessly. Second, we have not yet incorporated other important features, such as facial expressions and hand gestures, into our prediction model, which could further improve the prediction accuracy.

Looking ahead, our research agenda encompasses several promising avenues. We aim to extend and generalize our current three-party conversation model to accommodate more complex multiparty interactions. Lastly, although our current approach has the capability of an online approach for existing data, we have not yet developed a comprehensive top-down system that can seamlessly integrate all modalities while making predictions in real time.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Linda Bell, Johan Boye, and Joakim Gustafson. 2001. Real-time handling of fragmented utterances. In *Proc. NAACL workshop on adaptation in dialogue systems*. 2–8.

[2] Dan Bohus and Eric Horvitz. 2011. Decisions about turns in multiparty conversation: from perception to action. In *Proceedings of the 13th international conference on multimodal interfaces*. 153–160.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] 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.

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[7] 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.

[8] Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* 82, 3 (2006), 515–535.

[9] Ziedune Degutyte and Arlene Astell. 2021. The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings. *Frontiers in Psychology* 12 (2021), 616471.

[10] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54 (2021), 755–810.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*. Association for Computational Linguistics, 4171–4186.

[12] 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.

[13] Nia MM Dowell, Tristan M Nixon, and Arthur C Graesser. 2019. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior research methods* 51 (2019), 1007–1041.

[14] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology* 23, 2 (1972), 283.

[15] Cecilia E Ford and Sandra A Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics* 13 (1996), 134–184.

[16] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. 2022. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16102–16112.

[17] Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25, 3 (2011), 601–634.

[18] 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.

[19] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. *Listener* 162 (2018), 364.

[20] Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one* 10, 8 (2015), e0136905.

[21] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. A multimodal end-of-turn prediction model: learning from parasocial consensus sampling. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. Citeseer, 1289–1290.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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*.

[30] S. Kawato and J. Ohya. 2000. Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. 40–45. https://doi.org/10.1109/AFGR.2000.840610

[31] 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.

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

[33] Meng-Chen Lee, Wu Angela Li, and Zhigang Deng. 2024. A Computational Study on Sentence-based Next Speaker Prediction in Multiparty Conversations. In *Proceedings of ACM International Conference on Intelligent Virtual Agents 2024*.

[34] Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal Turn Analysis and Prediction for Multi-party Conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 436–444.

[35] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[36] Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech & Language* 28, 4 (2014), 903–922.

[37] 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.

[38] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Investigating speech features for continuous turn-taking prediction using lstms. *arXiv preprint arXiv:1806.11461* (2018).

[39] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale rnns. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 186–190.

[40] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*. Elsevier, 7–55.

[41] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[42] Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. 2002. Learning decision trees to determine turn-taking by spoken dialogue systems.. In *INTERSPEECH*.

[43] David Schlangen. 2006. From reaction to prediction: Experiments with computational models of turn-taking. *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking* (2006).

[44] Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 220–230.

[45] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.

[46] 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.

[47] Nigel G Ward. 2019. *Prosodic patterns in English conversation*. Cambridge University Press.