# A Survey of Recent Advances on Turn-taking Modeling in Spoken Dialogue Systems

**Galo Castillo-López** **Gaël de Chalendar** **Nasredine Semmar**
Université Paris-Saclay, CEA, List, Palaiseau, France
{galo-daniel.castillolopez, gael.de-chalendar, nasredine.semmar}@cea.fr

## Abstract

The rapid growth of dialogue systems adoption to serve humans in daily tasks has increased the realism expected from these systems. One trait of realism is the way speaking agents take their turns. We provide here a review of recent methods on turn-taking modeling and thoroughly describe the corpora used in these studies. We observe that 72% of the reviewed works in this survey do not compare their methods with previous efforts. We argue that one of the challenges in the field is the lack of well-established benchmarks to monitor progress. This work aims to provide the community with a better understanding of the current state of research around turn-taking modeling and future directions to build more realistic spoken conversational agents.

## 1 Introduction

Conversational agents adoption is rapidly growing. The ubiquity of dialogue systems in recent years has increased the realism (i.e. human-likeness) expected from them. One major trait of realism is the way spoken dialogue systems take turns in dialogues (Ter Maat et al., 2011). Silence between utterances in human-human conversations take 200ms on average (Levinson and Torreira, 2015). However, current spoken dialogue agents initiate turns after long gaps (700-1000ms) (Li et al., 2022), which results in unnatural, less realistic and non-fluid conversations. Thus, realistic turn-taking behavior is still a challenge to be addressed. The main goal of turn-taking modeling is determining when the system should take the turn to speak. Simplest attempts rely on the use of audio-based Voice Activity Detection (VAD) systems and silence thresholds (Raux, 2008; Raux and Eskenazi, 2009). Long silence thresholds derive systems that wait more time than expected, while shorter thresholds tend to interrupt users in the middle of their turns. In contrast, most recent systems use neural approaches to model turn-taking by minimizing speech overlaps and silence gaps. These systems use all types of available input data such as timing, linguistic, speech and visual information. For instance users' gaze and voice intonation at the end of utterances have been found to be relevant features to predict their end-of-turns (Oertel et al., 2012; Gravano and Hirschberg, 2011; Duncan, 1972). Turn-taking cues generation, interruptions handling, and other tasks are also of interest in the management of turns (Skantze, 2021).

Modeling turn-taking in multi-party conversations (MPCs) has not been widely explored in comparison to *dyadic* scenarios (i.e. one single user at a time), but has gained more attention in recent years. Multi-party conversations consist of conversations where more than two participants are involved, e.g. two users and a conversational agent. These types of dialogues include additional complexities to the management of turn-taking (Ganesh et al., 2023). For instance, recognizing who the addressee(s) of a user's utterance is does not appear to be as trivial as in the dyadic case. Challenges in MPCs are detailed in Appendix A. Examples of dialogue systems intervening in such types of conversations include agents providing assistance at hospital receptions (Addlesee et al., 2024a,b) or autonomous public buses (Axelsson et al., 2024), serving as healthcare coaches (Kantharaju and Pelachaud, 2018), guiding games at museums and hospitals for visitors entertainment (Skantze et al., 2015; Schauer et al., 2023), or taking place as attractions at thematic parks (Paetzel-Prüsmann and Kennedy, 2023).

While Skantze (2021) provided an exhaustive overview of turn-taking modeling, significant developments since then call for an updated survey. This paper contributes beyond the prior work in three key ways: (1) it offers the



*Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 254–271
May 27–30, 2025. 

first comprehensive review of datasets used in the field, providing critical insights into data modality and language; (2) it discusses overlooked limitations in turn-taking models, identifying gaps and challenges that are crucial for future work; and (3) it examines new ideas and approaches that have emerged in recent years, reflecting the latest trends and innovations. This survey is designed to serve both newcomers and experienced researchers in the field of turn-taking modeling. For those unfamiliar with the topic, it provides a clear introduction to fundamental concepts in turn-taking management, laying the groundwork for understanding the area. For expert readers, it offers a detailed examination of recent advancements, including new datasets, approaches, and unresolved challenges, with the goal to make it a valuable resource for anyone looking to stay current with the latest developments in the field, complementary to previous reviews. This survey describes relevant research on turn-taking modeling, with a special attention on studies published after 2021. Readers seeking more detailed information on earlier work are invited to refer to (Skantze, 2021). In Appendix B we explain the paper selection criteria we adopted in this survey.

The structure of this paper is organized as follows. In Section 2, we review fundamental concepts of turn-taking management. Section 3 provides an exhaustive description of the corpora used in the field. In Sections 4 and 5, we describe works on the two main subtasks in turn-taking modeling: end-of-turn prediction and backchannel prediction, respectively. We report studies related to MPCs in Section 6. Finally, Section 7 discusses the main open challenges and future opportunities in the field.

## 2 Turn-taking Management

Turn-taking in conversations can be defined as the coordinated successive exchange of speaking roles between multiple subjects to speak, listen, and respond (Fusaroli et al., 2014). Cooperative verbal communication is not unique to humans, as other animal species have also shown certain forms of turn-taking behavior (Pika et al., 2018; Takahashi et al., 2013). Although the coordination of turns feels natural in most human dialogues, it requires training at early stages of childhood (Nguyen et al., 2022; Donnelly and Kidd, 2021; Cosper and Pika, 2024). This suggests that there is a level of cognitive effort we need to perform to fluently manage turns. Such coordination lies on its dynamic temporal structure where listeners have to foresee the end of the speaker's utterance to anticipate their take of turn (Sacks et al., 1978). Figure 2 in Appendix C shows how turn-taking is handled in dialogues and illustrates various elements associated to turn management.

Fluency in the organization of turns is commonly assessed by the amount of overlaps and gaps between turns. A high number of these events indicate a poor ability of anticipating the end of turn by listeners (Heldner and Edlund, 2010). Overlaps occur when the listener starts speaking before the speaker completes their utterance. Gaps take place when long silences precede the take of turn of the next speaker. To optimize the organization of turns, listeners rely on cues provided by the speaker while holding or releasing their turns to determine when it is adequate to take the turn. Similarly, speakers rely on cues generated by listeners to know when any listener desires to take the turn, in order to decide whether to hold or release the floor. These cues can include verbal signals, gestures, and others. We describe in detail turn-taking cues in Appendix D.

Modeling turn-taking comprises multiple subtasks such as end-of-turn detection, interruptions handling and others. Although certain attempts of turn-taking modeling have proposed to simultaneously tackle several subtasks (Nguyen et al., 2023), each of them have been mostly treated as independent problems. In addition, turn-taking in MPCs has received little attention compared to two-party dialogues. The study of turn-taking modeling in MPCs has been mainly conducted in the field of human-robot interaction (Sato and Takeuchi, 2014; Bohus and Horvitz, 2010; Skantze et al., 2015), since it is difficult to organize turns in MPCs without the visual channel (Skantze, 2021).

## 3 Datasets

In this section, we detail the datasets used in all the works we review in this survey, i.e. described in sections 4, 5, and 6. Although research on turn-taking modeling have been mainly developed on dialogues in English, there are a few dialogue corpora in other languages. We separate datasets according to languages in English and other languages. A summary of all datasets is shown in Table 1.

| Dataset | Language | Modality | Duration | Nb. dialogues | Nb. turns | Multy-party |
|---|---|---|---|---|---|---|
| Switchboard (Godfrey et al., 1992) | en | sp, txt | 260h | 2.4K | 106.6K | ✗ |
| HarperValleyBank (Wu et al., 2020) | en | sp, txt | 24h | 1.4K | 25.7K | ✗ |
| HCRC Map Task (Anderson et al., 1991) | en | sp, txt | 15h | 128 | - | ✗ |
| Mahnob Mimicry Database (Bilakhia et al., 2015) | en | sp, vid | 11h | 54 | - | ✗ |
| Fisher Corpus (Cieri et al., 2004) | en | sp, txt | 1960h | 11.7K | - | ✗ |
| NoXi Database (Cafaro et al., 2017) | en, es, fr, de, it, ar, id | sp, txt, vid | 25h | 84 | 1.7K | ✗ |
| Japanese Travel Agency Task (Inaba et al., 2022) | ja | sp, txt, vid | 15h | 330 | 111.7K | ✗ |
| SSC of Japanese (Maekawa et al., 2000) | ja | sp, txt | 661h | 3.3K | - | ✗ |
| HKUST/MTS Corpus (Liu et al., 2006) | zh | sp, txt | 200h | 1.2K | 248.9K | ✗ |
| JaNoXi (Onishi et al., 2023) | ja | sp, txt, vid | 7h | 19 | - | ✗ |
| EALC (Yoshino et al., 2018) | ja | sp, txt | 200h | 60 | 28.0K | ✗ |
| ICSI Meeting Corpus (Janin et al., 2003) | en | sp, txt | 72h | 75 | - | ✓ |
| AMI Meeting Corpus (Kraaij et al., 2005) | en | sp, txt | 100h | 175 | - | ✓ |
| CEJC (Koiso et al., 2022) | ja | sp, txt, vid | 200h | 577 | - | ✓ |

Table 1: Spoken dialogue corpora for turn-taking modeling tasks. **sp:** speech, **txt:** transcripts, **vid:** video.

### 3.1 English Corpora

One of the most used datasets for turn-taking modeling is the **Switchboard Corpus** (Godfrey et al., 1992). This dataset is a collection of audio and transcripts from 2.4K dyadic fully spontaneous telephone call dialogues by 500 speakers. The **Fisher Corpus** (Cieri et al., 2004) consists of 11.7K topic-oriented telephone conversations among randomly paired recruited participants. Similarly, the **HarperValleyBank Corpus** (Wu et al., 2020) contains 24 hours of simulated telephone dialogues between participants playing the roles of bank agents and customers. Dialogues are labeled according to the customers' intentions and utterances are assigned a sentiment class and dialogue act. Both the HarperValleyBank Corpus and Fisher Corpus count with audio data and speech transcripts. The **HCRC Map Task Corpus** (Anderson et al., 1991) was collected to study linguistic phenomena in cooperative dyadic interactions between young speakers. This dataset adds up to 128 conversations where an instruction giver indicates an instruction listener how to reproduce a route in a map, which is only known by the instruction giver. The **Mahnob Mimicry Database** (Bilakhia et al., 2015) is a set of 54 audiovisual recordings of socio-political discussions and tenancy negotiations. The corpus includes visual annotations such as gestures, body movement and facial expressions.

### 3.2 Corpora in Other Languages

The **NoXi Database** (**NO**vice e**X**pert **I**nteraction database) (Cafaro et al., 2017) is a set of audiovisual recordings designed to study social behavior in seven languages: English, Spanish, French, German, Italian, Arabic and Indonesian. Skeleton data, action units, head position and other types of data were collected along 25 hours of dyadic conversations where interlocutors discussed about a large variety of topics. A Japanese version of the NoXi corpus is compiled in the **JaNoXi** dataset (Onishi et al., 2023), where 6.8 hours of dialogues were recorded in similar settings as the NoXi Database. The **Japanese Travel Agency Task** dataset (Inaba et al., 2022) and the **Spontaneous Speech Corpus of Japanese** (Maekawa et al., 2000) are other datasets of two-party conversations in Japanese with over 15 and 661 hours of speech, respectively. The first compiles audio, video and transcripts of tourism consultation dialogues between a customer and an agent through the online meeting platform Zoom. The second mostly corresponds to annotated monologues in spontaneous Japanese. The corpus comprises morphologically annotated transcripts, as well as segmental and intonation labeling for mainly studying speech recognition. The **Elderly Attentive Listening Corpus** (EALC) is a 200h text and speech corpus designed for modeling various dialogue tasks in conversations with elderly people (Yoshino et al., 2018). Mandarin conversations were compiled in the **HKUST Mandarin Telephone Speech Corpus** (Liu et al., 2006), which includes speech data, transcripts and speaker demographic information, e.g. age, gender, education background, etc. In total, 1,206 ten-minute natural Mandarin telephone conversations about multiple topics were recorded to study topic detection, speaker recognition and others.

### 3.3 Multi-party Corpora

The **AMI Meeting Corpus** (Kraaij et al., 2005) and **ICSI Meeting Corpus** (Janin et al., 2003) are two well-known datasets of multi-party conversation

audio recordings in English. The former corresponds to 175 sessions of four participants in scenario-oriented meetings. These recordings also contain collected data from devices such as digital pen and whiteboard usage, as well as video recordings. The ICSI corpus consists of 72 hours of meetings not elicited by a scenario, i.e. meetings would have taken place in any case. Speech transcripts are available for both datasets. The **Corpus of Everyday Japanese Conversation** (CEJC) dataset includes videos, audios and transcripts of spontaneous Japanese dialogues occurring in everyday scenarios (Koiso et al., 2022). The CEJC contains 200 hours of speech from 577 conversations, where around half of them are MPCs.

## 4 End-of-turn Prediction

Detecting the end-of-turn is the most well-studied problem in turn-taking modeling. End-of-turn prediction, also referred to as *end-of-utterance (EOU) detection*, is usually defined as a binary classification task. Its goal is to determine if the system should take the turn or not, depending on the dialogue context. Methods for EOU prediction can be grouped in three categories: silence-based, IPU-based and continuous (Skantze, 2021). Silence-based methods rely on Voice Activity Detection (VAD) tools, where a silence threshold (e.g. 700ms) is set to determine whether the system should take the turn. These methods result in poor user experience due to lack of naturalness (Aldeneh et al., 2018; Ekstedt and Skantze, 2022). IPU-based and continuous approaches differ on the time when predictions are made along the dialogue. While IPU-based[1] methods evaluate if the turn should be taken after every inter-pausal unit, i.e. after a silence, continuous models constantly evaluate the occurrence of an end of turn regardless of silences– e.g. every 50ms of speech. In this survey work we focus in continuous and IPU-based methods.

### 4.1 Continuous Methods

Continuous models either periodically evaluate end of turn at different time frames (Skantze, 2017), or incrementally perform predictions as utterances are built token by token (Coman et al., 2019). Predictions are executed regardless of whether silences of certain duration are observed. Some of the first attempts to build continuous methods for modeling turn-taking is performed in (Skantze, 2017), where a model that predicts future speech activity at every new frame of 50ms is proposed. A LSTM model (Hochreiter and Schmidhuber, 1997) is trained to predict the occurrence of a turn-shift from acoustic input features, including voice activity, pitch, speech intensity, and spectral stability, as well as Part-of-speech (POS) tags. An extension of (Skantze, 2017) revealed that there are significant performance benefits to modeling linguistic features at a lower temporal rate, and in a separate sub-network from acoustic features (Roddy et al., 2018b). Other early attempts explored reinforcement learning to model turn-taking (Zhao et al., 2015; Khouzaimi et al., 2016, 2018).

Roddy et al. (2018a) observed that POS tags only enhance model performance to discern whether an utterance will be short, e.g. backchannel. Hara et al. (2018) found that introducing backchannel and filler predictions as auxiliary tasks improved turn-taking prediction. Several studies have shown that the simultaneous use of both prosodic and word features outperforms the independent use of each separately (Wang et al., 2024; Li et al., 2022; Liu et al., 2017), which is in line with previous research that tends to confirm that combined turn-taking cues in human communication have an additive effect (Hjalmarsson, 2011). More recently, studies have investigated how ASR can be utilized for turn-completion time estimation (Kanai et al., 2024; Zink et al., 2024). Kanai et al. (2024) showed that fine-tuning wav2vec 2.0 (Baevski et al., 2020) for ASR to introduce linguistic features outperforms the use of solely acoustic features. Instruction fine-tuning (Wei et al., 2022) in a multitask setting has also been explored on LLMs in combination with HuBert (Hsu et al., 2021) through a fusion layer to model turn-taking from linguistic and acoustic features in (Wang et al., 2024). Likewise, Chang et al. (2022) feed a RNN Transducer (Graves, 2012) with audio streams and previous tokens to predict turn-taking-related wordpieces. Gaze direction, head pose and other non-verbal features have also been studied in combination with speech information. Onishi et al. (2023) found that action units are crucial input information for turn-taking and backchannel prediction. Results in these works exhibit that words prosody, timing, linguistic, and other types of features jointly provide better signals

[1]IPU: Inter-Pausal Units, see Appendix C.

for predicting EOU.

While many works focus on audio signals as main inputs, recent methods have shown how syntactic completeness obtained from transcripts alone can be used for turn-taking modeling. Ekstedt and Skantze (2020) introduced TurnGPT, a language model based on GPT-2 (Radford et al., 2019) and fine-tuned on various dialogue datasets to predict turn-completion based on text features only. They represent dialogues as sequences of concatenated utterances, separated by special tokens associated to turn-shift, to learn probabilities of turn-completion. Their results demonstrate that turn-shift prediction performed as a language modeling task outperforms previous work due to the strong representation of context that prior models miss. In (Jiang et al., 2023), response candidates are also considered as a proxy to determine whether a turn-shift is plausible in a given dialogue, arguing that the decision of taking a turn also depends on what the next speaker wants to say. Their results indicate that response-conditioning is especially useful when the utterance is a question and it semantically matches with the response. Further works showed that adapting TurnGPT to two separate streams of lexical content improves EOU prediction by capturing temporal dynamics (Leishman et al., 2024).

Most recent advances propose models based on Voice Activity Projection (VAP), whose main objective is essentially to predict future voice activity of every interlocutor in the conversation (Inoue et al., 2024b; Ekstedt and Skantze, 2023; Onishi et al., 2024). These models incrementally process the interlocutor speech to mimic humans' abilities to infer what the speaker is going to say to simultaneously prepare a reply and reduce response delay (Schlangen and Skantze, 2011). Ekstedt and Skantze (2022) propose a VAP self-supervised learning model to predict distinct turn-taking events and evaluated on zero-shot settings in four tasks: shift vs. hold prediction at mutual silence, shift prediction at voice activity presence, upcoming backchannel prediction, and backchannel vs. turn-shift prediction. The proposed base model consists of a frame-wise speech and VA encoder followed by a sequence predictor. VAP models have been found to perform better in Japanese when trained in English and fine-tuned with Japanese data than models directly trained in Japanese (Sato et al., 2024a). Inoue et al. (2024a) investigated multilingual VAP models to predict turns-shifts in English, Mandarin, and Japanese. While their results indicate that models evaluated in cross-lingual settings do not perform well, Sato et al. (2024b) demonstrated that aligning the criteria for speech segmentation labels across datasets is crucial to provide proper evaluation and to effectively use VAP models in cross-lingual scenarios.

We note an emerging trend in continuous methods using VAP models. An important opportunity in this direction for future work is the examination on how this type of models can be integrated with multi-modal data (e.g. video signals), as they have only been explored on audio inputs. We argue that although promising results have been observed when using utterance-level labels such as dialogue acts, the lack of availability of these types of annotations in real-world scenarios is a key limitation. We also find that even though LLMs have shown impressive results in a series of NLP tasks, recent studies demonstrate their inefficiency to detect opportunities to take turns at mid-utterance in spoken dialogue (Umair et al., 2024).

### 4.2 IPU-based Methods

Turn-taking models based on IPUs assume that turns cannot be taken while the user speaks (Skantze, 2021). Hence, predictions are performed every time a silence is detected from user's channel. Early works used LSTM-based architectures to model turn-taking from prosodic, phonetic, and lexical sequential features (Masumura et al., 2017, 2018; Hara et al., 2019). On the other hand, models based on CNN have been observed to be effective when introducing visual cues such as eye, mouth and head motion (Kurata et al., 2023). Experiments on multi-task learning have shown that using speech acts in auxiliary tasks for turn-taking modeling improves system performances. Aldeneh et al. (2018) observed that using speaker intention prediction (e.g. asking a question, uttering a backchannel, etc.) as a secondary task enhance turn shift prediction performance. Sakuma et al. (2022) found that integrating dialogue act information for response time estimation allows systems to efficiently capture dialogue context with smaller amounts of data than other methods.

Recent works have explored syntactic completeness to model turn-taking. Ekstedt and Skantze (2021) used TurnGPT to introduce

speaker shift tokens as in (Ekstedt and Skantze, 2020), but on an IPU-based approach. At the end of every IPU, they project possible continuations from dialogue context to obtain the ratio of continuations containing shift tokens, to be used as an approximation of the actual probability of EOU. Syntactic completeness is also studied in (Sakuma et al., 2023) to determine response time from a multimodal Japanese dialogue corpus. They build a unidirectional LSTM language model to compute the probability of a special EOU token appearing in the next $M$ tokens, outperforming the Gated Multimodal Fusion method proposed in (Yang et al., 2022) on similar features. Inspired by (Morais et al., 2022), the use of self-supervised learning based on Up- plus Down-stream models has also been investigated on audio and text data for end of turn detection (Morais et al., 2023).

The use of syntactic completeness has demonstrated relevant improvements in IPU-based methods for turn-taking modeling by effectively leveraging linguistic cues to predict turn-completion points. In addition to the efforts made by (Sakuma et al., 2023), future work may consider exploring how syntactic completeness can be integrated into multimodal methods using non-LSTM architectures as done in (Kurata et al., 2023), as well as in multi-task training settings. Moreover, although promising results using LLMs combined with VAD systems have been reported in (Pinto and Belpaeme, 2024), studies on this direction still have to be widely studied. We observe that IPU-based approaches have received less attention than continuous methods, as the latest continuous models are more aligned to human-like EOU prediction.

## 5 Backchannel Prediction

Overlaps in dialogues occur when multiple participants produce IPUs at the same time. These overlaps may take place in the proximity of the end of a turn if the listener desires to start their turn, or in the middle of the speaker utterance. In the latter case there are three possible scenarios: (1) the listener desires to interrupt and grab the floor, (2) the listener intents to provide a feedback to the speaker without the aim of taking the turn (backchannel), and (3) the listener produces non-lexical sounds such as coughing, which can be misinterpreted as an interruption. Classifying an overlap as a backchannel or an actual interruption (or noise) is an important subtask in turn-taking modeling. Backchannel prediction is generally defined as a binary classification task, where the aim is to classify an IPU as a backchannel or non-backchannel.

One of the first attempts to model backchannel prediction using neural networks was reported in (Mueller et al., 2015), where only speech features were used. In (Skantze, 2017), backchannel detection was addressed by predicting if a speech onset of 500 ms corresponded to a short (less than 500ms, i.e. backchannel) or a long utterance (more than 2500 ms), using handcrafted acoustic features and POS tags to feed a LSTM model. Yokoyama et al. (2018) considered backchannels as an intention label to build an intention recognition model. Other early works used word2vec (Mikolov et al., 2013) to combine word embeddings as linguistic features with acoustic features (Ruede et al., 2017). Adiba et al. (2021) took delays in ASR into account to propose a prior prediction model, as words are available some time after these have been uttered. Speaker and listener embeddings to encode interlocutor interactions have also been considered to predict backchannels (Ortega et al., 2020, 2023).

Recent advances have examined auxiliary tasks to predict backchannels (Choi et al., 2024; Wang et al., 2024), showing improvements over single-task methods. These tasks include sentiment classification, dialogue act prediction, and others (Liermann et al., 2023; Jang et al., 2021). Müller et al. (2022) used audiovisual data to introduce agreement estimation in a multitask setting to detect backchannels. Park et al. (2024) proposed a Context-Aware Backchannel Prediction model to enhance predictions in Korean and English corpora. They encoded features using text embeddings from BERT (Devlin et al., 2018) and acoustic inputs represented by wav2vec embeddings. Finally, a multi-head attention mechanism is employed to build an attentive context embedding that holds relevant information of the current utterance. Voice Activity Projection (VAP) models, presented in Section 4.1, have gained special attention for backchannel prediction. Onishi et al. (2024) found that integrating non-verbal features on VAP models enhances turn-taking events prediction, including backchannels. Pre-training on large dialogue data and fine-tuning on a specialized backchannel corpus has also shown improvements on VAP model's generalizability (Inoue et al., 2024c).

## 6 Multi-party Turn-taking Modeling

In this section we describe methods proposed for turn-taking modeling in multi-party conversations. We outline the main complexities in MPCs in human-human dialogues in Appendix A. Turn-taking prediction becomes more challenging in multi-party scenarios, where various sub-tasks arise such as conversation disentanglement, addressee recognition, and others (Ganesh et al., 2023). Multi-party conversation modeling addresses the issues on *Who says What to Whom* (who speaks, says what, and addresses whom) (Gu et al., 2022). Modeling turn-taking in MPCs has received much less attention than dyadic interactions. Although work has been done on the topic for a long time (Traum, 2003; Laskowski, 2010; Bohus and Horvitz, 2011; De Kok and Heylen, 2009; Thórisson et al., 2010), only recently have the methods designed begun to yield promising results, due to advancements in available technologies.

Fujie et al. (2021) proposed a Timing Generating Network, which incorporates a first-order lag system to estimate how much other speakers in the dialogue expect the system to take the turn in Japanese. Their approach is in contrast to the conventional framing of turn-taking modeling as an end-of-turn detection problem, which assumes that the system should take the turn right after the previous speaker releases it. They integrate response obligation recognition as an auxiliary task to improve estimation. de Bayser et al. (2019) and de Bayser et al. (2020) studied next speaker identification from dialogue logs to model turn-taking prediction in MPCs. Gaze-transition patterns and timing information have been investigated to predict the next speaker and the time at which each utterance will be made (Ishii et al., 2016; Lee et al., 2023). Experiments on a Transformer-based architecture using 3D gazes, 3D head and body movements, and speech showed that speech signals play a more critical role than gaze patterns for turn-taking prediction (Lee et al., 2023). Multimodal fusion has also been studied for turn-taking prediction, where multiple event types are predicted simultaneously (Lee and Deng, 2024). Johansson and Skantze (2015) argue that there are different states in which turn-taking could be obliged or optional. They proposed annotating utterances into a scale of four classes according to the appropriateness for an agent to take the turn. They observed that dialogue acts as turn change predictors in MPCs need a special treatment compared to two-party settings.

Turn-taking modeling in MPCs has notably been understudied in comparison to dyadic scenarios. Work in MPCs where no visual data are used has been overlooked, as most studies have been conducted in the field of human-robot interaction where visual cues are captured by sensors. We think that introducing response obligation detection in any form is crucial for modeling turn-taking in MPCs. Determining whether the agent should take the turn or not after the floor is released mitigates poor performance on Out-of-Scope utterances, a common phenomenon for dialogue systems in such scenarios. We discuss more about open challenges in multi-party conversations, including Out-of-Scope utterances, in section 7.4.

## 7 Challenges and Future Directions

In this section we describe some relevant open challenges and suggest opportunities for future work.

### 7.1 System Evaluations

The lack of comparative evaluations in works is one of the most important challenges in the field. We find that only 28% of the reviewed papers in this survey compare their methods with systems presented in prior works. Even some of the studies where comparisons are made, do not use the same data to compare methods or use different input features than originally proposed. Consequently, comparisons are not fair. We also note that one third of the reviewed works on end-of-turn detection do not conduct experiments on any public corpus. This represents an issue for reproducibility and properly monitoring progress in the field.

To address this challenge, we suggest future work should focus on the creation of a standardized benchmark for each turn-taking modeling task. This benchmark should include a diverse set of publicly available corpora. We note that the Switchboard Corpus is the most popular resource for turn-taking modeling evaluation in dyadic dialogues, as used in 69% and 41% of the surveyed papers on backchannel and end-of-turn detection, respectively. An important aspect to take into consideration for end-of-turn detection is the definition of turns and IPUs. Previous works have used distinct silence thresholds between 50ms and

1s to delimit IPUs or define turns. We believe that considering multiple cutoffs for evaluation is necessary, as done in (Skantze, 2017; Sakuma et al., 2023).

### 7.2 Groups with Varying Needs

Another challenge for spoken dialogue systems is the interaction with people who present more complex behaviors on turn management such as senior adults or individuals with mental health disorders. These types of interplays require systems to conduct human-like turn-taking behaviors (Addlesee and Eshghi, 2024; Bell, 2024). Prior work has proposed mechanisms to address these scenarios (Lala et al., 2017; Hara et al., 2018; Kawahara et al., 2016). LLMs have been recently proposed to build dialogue systems that interact with individuals with mental health disorders. (Addlesee and Eshghi, 2024) studied the recovery from interruptions in dialogues with people with dementia. Although the previously mentioned efforts are valuable, the body of work in this subject is still scanty. We also observe that most of studies in turn-taking modeling in these scenarios have not been conducted in multidisciplinary environments. We believe that integrating domain-knowledge and insights from experts in other fields beyond Dialogue Systems, would be beneficial for the research community.

### 7.3 Multilinguality

Although multilingual aspects in dialogue systems have been addressed in other sub-tasks such as natural language understanding (Firdaus et al., 2023; Gupta et al., 2021; Gerz et al., 2021), dialogue state tracking (Lee et al., 2024; Yu et al., 2023; Zuo et al., 2021), or response generation (Wu et al., 2024), research in turn-taking modeling is limited (Ward et al., 2018; Inoue et al., 2024a). End of turn prediction is more difficult in some languages than others even for humans (Stivers et al., 2009). Inherent phenomena from spoken dialogues, which are not found in other dialogue system sub-tasks, such as backchannels or hesitations, make end of utterance detection more complex. For instance, backchannels use varies from one culture to another (Clancy et al., 1996; Tartory et al., 2024).

### 7.4 Multi-party Conversations

In Section 6 we briefly discussed about the limited amount of work in turn-taking modeling on MPCs, which is in line with the low amount of available MPC corpora we described in Section 3. We found that most works in turn-taking modeling on MPCs use visual information to detect end of turns, since predicting EOUs without the aid of the visual channel is a complex task. However, there are multiple scenarios where using visual inputs is not feasible or useful. For example, agents assisting participants in online meetings do not count with such sort of cues and should mostly rely on linguistic and audio inputs. Future work should take this challenge into consideration.

Additionally, we note that most works in MPCs propose systems where agents actively participate in dialogues. Nevertheless, that is not always the case in real-world applications. For instance, task-oriented dialogue agents such as Alexa or Siri, generally play the role of listeners in dialogues, switching their role when they have something to say –commonly when a wake-word is spoken. Skilled assistants should not only base their turn-taking decisions on wake-words, but should be effective on determining when to intervene in conversations to assist on a given task. In other words, virtual assistants should be able to detect when they can contribute in dialogues in scenarios where they are not expected to have an active participation. One major challenge in these scenarios is managing Out-of-Scope (OOS) utterances, as users may discuss about a diverse set of topics where only a few utterances are task-related. The study of intent recognition in MPCs is a possible direction on this subject, as user intentions may suggest the need for intervention of an agent. One limitation is that there are no corpora with intent recognition annotations in spontaneous MPCs with a focus on OOS utterances. A corpus with these characteristics in scripted MPCs (dialogues from TV shows) is proposed in (Zhang et al., 2024). Addlesee et al. (2023) used GPT-3.5-turbo to detect user goals in MPCs, which can be seen as a surrogate task for turn-taking modeling. They argue that users' goals in MPCs can be addressed by virtual agents as well as other human participants, hence they propose the task of goal-tracking to detect solved tasks and determine the relevance of agent intervention. Intent recognition was also used as an auxiliary task for turn-taking prediction in (Aldeneh et al., 2018). To the best of our knowledge, these are the only studies where intent recognition is considered for modeling turn-taking. However, none of such

works take into consideration OOS utterances.

## 8 Conclusions

Turn-taking modeling is a key component of spoken dialogue systems. Effective methods for modeling turn taking are crucial for developing systems that can be perceived as realistic. This survey provides an overview of recent advancements in turn-taking modeling in spoken dialogue systems. We provided the first detailed review of the corpora used in the field. We observed that the majority of works have been conducted on English and Japanese corpora, with almost no efforts in other languages. We also described recent works in end-of-turn prediction and backchannel classification. Finally, we discussed several overlooked open challenges in current turn-taking models and key directions indicating how future work could push the field. For instance, we noted a tendency in the reviewed works not to compare their proposed methods with previous works, which might affect monitoring progress in the field. Addressing these challenges and improving cross-linguistic research and method comparisons will be essential for advancing turn-taking models and making spoken dialogue systems more natural and effective.

## Acknowledgments

We warmly thank our anonymous reviewers for their time and valuable feedback. This work has been partially funded by the EU project CORTEX2 (under grant agreement: N° 101070192).

## References

Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández Garcia, Nancie Gunson, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2024a. Multi-party multimodal conversations between patients, their companions, and a social robot in a hospital memory clinic. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 62–70.

Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández García, Nancie Gunson, Weronika Sieińska, Marta Romeo, Christian Dondrup, and Oliver Lemon. 2024b. A multi-party conversational social robot using llms. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1273–1275.

Angus Addlesee and Arash Eshghi. 2024. You have interrupted me again!: making voice assistants more dementia-friendly with incremental clarification. *Frontiers in Dementia*, 3:1343052.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernandez Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with LLMs: Comparing pre-training, fine-tuning, and prompt engineering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 229–241, Prague, Czechia. Association for Computational Linguistics.

Amalia Istiqlali Adiba, Takeshi Homma, Dario Bertero, Takashi Sumiyoshi, and Kenji Nagamatsu. 2021. Delay mitigation for backchannel prediction in spoken dialog system. *Conversational Dialogue Systems for the Next Decade*, pages 129–143.

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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6159–6163. IEEE.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Peter Auer. 2018. Gaze, addressee selection and turn-taking in three-party interaction. *Eye-tracking in interaction: Studies on the role of eye gaze in dialogue*, 197:231.

Agnes Axelsson, Bhavana Vaddadi, Cristian Bogdan, and Gabriel Skantze. 2024. Robots in autonomous buses: Who hosts when no human is there? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1278–1280.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Grace Madeline Bell. 2024. Prosodic speech rate, utterance duration, interruption rate, and turn-taking latency in autistic and neurotypical adults. Master's thesis, Brigham Young University.

Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. 2015. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61.

Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8.

Dan Bohus and Eric Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, pages 98–109.

Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359.

Shuo-Yiin Chang, Bo Li, Tara Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. 2022. Turn-Taking Prediction for Natural Conversational Speech. In *Proc. Interspeech 2022*, pages 1821–1825.

Yong-Seok Choi, Jeong-Uk Bang, and Seung Hi Kim. 2024. Joint streaming model for backchannel prediction and automatic speech recognition. *ETRI Journal*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Patricia M Clancy, Sandra A Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in english, japanese, and mandarin. *Journal of pragmatics*, 26(3):355–387.

Andrei C. Coman, Koichiro Yoshino, Yukitoshi Murase, Satoshi Nakamura, and Giuseppe Riccardi. 2019. An incremental turn-taking model for task-oriented dialog systems. In *Interspeech 2019*, pages 4155–4159.

Samuel H Cosper and Simone Pika. 2024. Human turn-taking development: A multi-faceted review of turn-taking comprehension and production in the first years of life. https://osf.io/yjad2/download.

Anne Cutler and Mark Pearson. 2018. On the analysis of prosodic turn-taking cues. In *Intonation in discourse*, pages 139–156. Routledge.

Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. 2019. Learning multi-party turn-taking models from dialogue logs. *arXiv preprint arXiv:1907.02090*.

Maira Gatti de Bayser, Melina Alberio Guerra, Paulo Cavalin, and Claudio Pinhanez. 2020. A hybrid solution to learn turn-taking in multi-party service-based chat groups. *arXiv preprint arXiv:2001.06350*.

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*, pages 91–98.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Seamus Donnelly and Evan Kidd. 2021. The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, 92(2):609–625.

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

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online. Association for Computational Linguistics.

Erik Ekstedt and Gabriel Skantze. 2021. Projection of turn completion in incremental spoken dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 431–437.

Erik Ekstedt and Gabriel Skantze. 2022. Voice activity projection: Self-supervised learning of turn-taking events. In *Proc. Interspeech 2022*, pages 5190–5194.

Erik Ekstedt and Gabriel Skantze. 2023. Show & tell: Voice activity projection and turn-taking. In *24th International Speech Communication Association, Interspeech 2023, Dublin, Ireland, Aug 20 2023-Aug 24 2023*, pages 2020–2021. International Speech Communication Association.

Mauajama Firdaus, Asif Ekbal, and Erik Cambria. 2023. Multitask learning for multilingual intent detection and slot filling in dialogue systems. *Information Fusion*, 91:299–315.

Shinya Fujie, Hayato Katayama, Jin Sakuma, and Tetsunori Kobayashi. 2021. Timing generating networks: Neural network based precise turn-taking timing prediction in multiparty conversation. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 3771–3775. International Speech Communication Association.

Riccardo Fusaroli, Ivana Konvalinka, and Sebastian Wallot. 2014. Analyzing social interactions: the promises and challenges of using cross recurrence quantification analysis. In *Translational recurrences: From mathematical theory to real-world applications*, pages 137–155. Springer.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2023. A survey of challenges and methods in the computational modeling of multi-party dialog. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7468–7475, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

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

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. 2022. Who says what to whom: A survey of multi-party conversations. In *IJCAI*, pages 5486–5493.

Akshat Gupta, Xinjian Li, Sai Krishna Rallabandi, and Alan W Black. 2021. Acoustics based intent recognition using discovered phonetic units for low resource languages. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7453–7457. IEEE.

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

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Turn-taking prediction based on detection of transition relevance place. In *INTERSPEECH*, pages 4170–4174.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Anna Hjalmarsson. 2011. The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Michimasa Inaba, Yuya Chiba, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2022. Collection and analysis of travel agency task dialogues with age-diverse speakers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5759–5767.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11873–11883, Torino, Italia. ELRA and ICCL.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024b. Real-time and continuous turn-taking prediction using voice activity projection. *arXiv preprint arXiv:2401.04868*.

Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2024c. Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. *arXiv preprint arXiv:2410.15929*.

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):70.

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):1–31.

Joseph Jaffe and Stanley Feldstein. 1970. Rhythms of dialogue. *(No Title)*.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. Bpm_mt: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3452.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023. Response-conditioned turn-taking prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12241–12248, Toronto, Canada. Association for Computational Linguistics.

Martin Johansson and Gabriel Skantze. 2015. Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 305–314.

Takanori Kanai, Yukoh Wakabayashi, Ryota Nishimura, and Norihide Kitaoka. 2024. Predicting utterance-final timing considering linguistic features using wav2vec 2.0. In *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–5. IEEE.

Reshmashree Bangalore Kantharaju and Catherine Pelachaud. 2018. Towards developing a model to handle multiparty conversations for healthcare agents. In *ICAHGCA@ AAMAS*, pages 30–34.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G Ward. 2016. Prediction and generation of backchannel form for attentive listening systems. In *Interspeech*, pages 2890–2894.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Kobin H Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical transactions of the royal society B*, 378(1875):20210473.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. 2016. Reinforcement learning for turn-taking management in incremental spoken dialogue systems. In *IJCAI*, pages 2831–2837.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. 2018. A methodology for turn-taking capabilities enhancement in spoken dialogue systems using reinforcement learning. *Computer Speech & Language*, 47:93–111.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the corpus of everyday Japanese conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5587–5594, Marseille, France. European Language Resources Association.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*.

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems. *Proc. Interspeech 2023*, pages 2658–2662.

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 127–136.

Kornel Laskowski. 2010. Modeling norms of turn-taking in multi-party conversation. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 999–1008.

Kornel Laskowski, Mattias Heldner, and Jens Edlund. 2012. On the dynamics of overlap in multi-party conversation. In *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012 Portland, OR; United States; 9 September 2012 through 13 September 2012;*, pages 846–849. Curran Associates, Inc.

Andrew H Lee, Sina J Semnani, Galo Castillo-López, Gäel de Chalendar, Monojit Choudhury, Ashna Dua, Kapil Rajesh Kavitha, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, et al. 2024. Benchmark underestimates the readiness of multi-lingual dialogue agents. *arXiv preprint arXiv:2405.17840*.

Meng-Chen Lee and Zhigang Deng. 2024. Online multimodal end-of-turn prediction for three-party conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 57–65.

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*, pages 436–444.

Sean Leishman, Peter Bell, and Sarenne Wallbridge. 2024. Pairwiseturngpt: a multi-stream turn prediction model for spoken dialogue. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

Siyan Li, Ashwin Paranjape, and Christopher Manning. 2022. When can I speak? predicting initiation points for spoken dialogue agents. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–224, Edinburgh, UK. Association for Computational Linguistics.

Wencke Liermann, Yo-Han Park, Yong-Seok Choi, and Kong Lee. 2023. Dialogue act-aided backchannel prediction using multi-task learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15073–15079.

Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2017. Turn-taking estimation model based on joint embedding of lexical and prosodic contents. In *Interspeech*, pages 1686–1690.

Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff. 2006. Hkust/mts: A very large scale mandarin telephone speech corpus. In *Chinese Spoken Language Processing: 5th International Symposium, ISCSLP 2006, Singapore, December 13-16, 2006. Proceedings*, pages 724–735. Springer.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka. 2017. Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. In *Interspeech*, volume 2017, pages 1661–1665.

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Ryuichiro Higashinaka, and Yushi Aono. 2018. Neural dialogue context online end-of-turn detection. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 224–228.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Edmilson Morais, Matheus Damasceno, Hagai Aronowitz, Aharon Satt, and Ron Hoory. 2023. Modeling turn-taking in human-to-human spoken dialogue datasets using self-supervised features. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926. IEEE.

Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques. In *Human-Computer Interaction: Interaction Technologies: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II 17*, pages 329–340. Springer.

Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. Multimediate'22: Backchannel detection and agreement estimation in group interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7109–7114.

Kazumasa Murai. 2011. Speaker predicting apparatus, speaker predicting method, and program product for predicting speaker. US Patent 7,907,165.

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*, 11:250–266.

Vivian Nguyen, Otto Versyp, Christopher Cox, and Riccardo Fusaroli. 2022. A systematic review and bayesian meta-analysis of the development of turn taking in adult–child vocal interactions. *Child Development*, 93(4):1181–1200.

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

Catharine Oertel, Marcin Włodarczak, Jens Edlund, Petra Wagner, and Joakim Gustafson. 2012. Gaze patterns in turn-taking. In *Thirteenth annual conference of the international speech communication association*.

D Kimbrough Oller. 1973. The effect of position in utterance on speech segment duration in english. *The journal of the Acoustical Society of America*, 54(5):1235–1247.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 13–21.

Kazuyo Onishi, Hiroki Tankka, and Satoshi Nakamura. 2024. Multimodal voice activity projection for turn-taking and effects on speaker adaptation. *IEICE Transactions on Information and Systems*.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE.

Daniel Ortega, Sarina Meyer, Antje Schweitzer, and Ngoc Thang Vu. 2023. Modeling speaker-listener interaction for backchannel prediction. *arXiv preprint arXiv:2304.04472*.

Maike Paetzel-Prüsmann and James Kennedy. 2023. Improving a robot's turn-taking behavior in dynamic multiparty interactions. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 411–415.

Yo-Han Park, Wencke Liermann, Yong-Seok Choi, and Kong Joo Lee. 2024. Improving backchannel prediction leveraging sequential and attentive context awareness. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1689–1694.

Simone Pika, Ray Wilkinson, Kobin H Kendrick, and Sonja C Vernes. 2018. Taking turns: bridging the gap between human and animal communication. *Proceedings of the Royal Society B*, 285(1880):20180598.

Maria J Pinto and Tony Belpaeme. 2024. Predictive turn-taking: Leveraging language models to anticipate turn transitions in human-robot dialogue. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 1733–1738. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Antoine Raux. 2008. Flexible turn-taking for spoken dialog systems. *Language Technologies Institute, CMU Dec*, 12.

Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 629–637.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018a. Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. In *Proc. Interspeech 2018*, pages 586–590.

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

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing backchannel prediction using word embeddings. In *Interspeech*, pages 879–883.

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*, pages 7–55. Elsevier.

Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2022. Response timing estimation for spoken dialog system using dialog act estimation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 4486–4490.

Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2023. Response timing estimation for spoken dialog systems based on syntactic completeness prediction. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 369–374. IEEE.

Ryo Sato and Yugo Takeuchi. 2014. Coordinating turn-taking and talking in multi-party conversations by controlling robot's eye-gaze. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 280–285. IEEE.

Yuki Sato, Yuya Chiba, and Ryuichiro Higashinaka. 2024a. Effects of multiple japanese datasets for training voice activity projection models. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Yuki Sato, Yuya Chiba, and Ryuichiro Higashinaka. 2024b. Investigating the language independence of voice activity projection models through standardization of speech segmentation labels. In *Proceedings of 2023 APSIPA Annual Summit and Conference*.

Laura Schauer, Jason Sweeney, Charlie Lyttle, Zein Said, Aron Szeles, Cale Clark, Katie McAskill, Xander Wickham, Tom Byars, Daniel Hernández Garcia, et al. 2023. Detecting agreement in multi-party conversational ai. *arXiv preprint arXiv:2311.03026*.

Emanuel A Schegloff. 1996. Issues of relevance for discourse analysis: Contingency in action, interaction and co-participant context. In *Computational and conversational discourse: Burning issues—An interdisciplinary account*, pages 3–35. Springer.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

Elizabeth Shriberg, Andreas Stolcke, and Don Baron. 2001. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Interspeech*, pages 1359–1362. Citeseer.

Rein Ove Sikveland and Richard Ogden. 2012. Holding gestures across turns: moments to generate shared understanding. *Gesture*, 12(2):166–199.

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*, pages 220–230.

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

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*, pages 67–74.

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

Jürgen Streeck and Ulrike Hartge. 1992. Previews: Gestures at the transition place. *The contextualization of language*, pages 135–157.

Daniel Y Takahashi, Darshana Z Narayanan, and Asif A Ghazanfar. 2013. Coupled oscillator dynamics of vocal turn-taking in monkeys. *Current Biology*, 23(21):2162–2168.

Raeda Tartory, Sami Al-khawaldeh, Samia Azieb, and Bassam Al Saideen. 2024. Backchannel forms and functions in context and culture: The use of backchannels in arab media discourse. *Discourse Studies*, page 14614456241236904.

Mark Ter Maat, Khiet P Truong, and Dirk Heylen. 2011. How agents' turn-taking strategies influence impressions and response behaviors. *Presence: Teleoperators and Virtual Environments*, 20(5):412–430.

Kristinn R Thórisson, Olafur Gislason, Gudny Ragna Jonsdottir, and Hrafn Th Thorisson. 2010. A multiparty multimodal architecture for realtime turntaking. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 350–356. Springer.

David Traum. 2003. Issues in multiparty dialogues. In *Workshop on Agent Communication Languages*, pages 201–211. Springer.

Muhammad Umair, Vasanth Sarathy, and JP de Ruiter. 2024. Large language models know what to say but not when to speak. *arXiv preprint arXiv:2410.16044*.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *ICASSP 2024*.

Nigel G Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. 2018. Turn-taking predictions across languages and genres using an lstm recurrent neural network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 831–837. IEEE.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Mike Wu, Jonathan Nafziger, Anthony Scodary, and Andrew Maas. 2020. Harpervalleybank: A domain-specific spoken dialog corpus. *arXiv preprint arXiv:2010.13929*.

Sixing Wu, Jiong Yu, Jiahao Chen, Xiaofan Deng, and Wei Zhou. 2024. Improving open-domain dialogue response generation with multi-source multilingual commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19252–19260.

Jiudong Yang, Peiying Wang, Yi Zhu, Mingchao Feng, Meng Chen, and Xiaodong He. 2022. Gated multimodal fusion with contrastive learning for turn-taking prediction in human-robot dialogue. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7747–7751. IEEE.

Katsuya Yokoyama, Hiroaki Takatsu, Hiroshi Honda, Shinya Fujie, and Tetsunori Kobayashi. 2018. Investigation of users' short responses in actual conversation system and automatic recognition of their intentions. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 934–940. IEEE.

Koichiro Yoshino, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo, and Satoshi Nakamura. 2018. Japanese dialogue corpus of information navigation and attentive listening annotated with extended ISO-24617-2 dialogue act tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xiang Yu, Zhang Ting, Di Hui, Huang Hui, Li Chunyou, Ouchi Kazushige, Chen Yufeng, and Xu Jinan. 2023. Improving zero-shot cross-lingual dialogue state tracking via contrastive learning. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 624–625.

Margaret Zellers, Jan Gorisch, David House, and Benno Peters. 2019. Timing properties of hand gestures and their lexical counterparts at turn transition places. In *Proceedings of the FONETIK (Swedish Phonetics Conference) 2019 in Stockholm, June 10–12, 2019*, pages 119–124. Stockholm University.

Margaret Zellers, David House, and Simon Alexanderson. 2016. Prosody and hand gesture at turn boundaries in swedish. In *Speech Prosody*, pages 831–835.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024. MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*.

Tiancheng Zhao, Alan W Black, and Maxine Eskenazi. 2015. An incremental turn-taking model with active system barge-in for spoken dialog systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 42–50.

Oswald Zink, Yosuke Higuchi, Carlos Mullov, Alexander Waibel, and Tetsunori Kobayashi. 2024. Predictive speech recognition and end-of-utterance detection towards spoken dialog systems. *arXiv preprint arXiv:2409.19990*.

Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. Allwoz: Towards multilingual task-oriented dialog systems for all. *arXiv preprint arXiv:2112.08333*.

## A Multi-party Conversations

In this appendix we describe the main difficulties in turn management in multi-party conversations and how they are addressed in natural dialogues. Multi-party conversations consist of dialogues where more than two interlocutors are involved. Sociolinguists have found these types of interactions the most natural form of conversations, arguing that dyadic scenarios and monologues are special cases (Jaffe and Feldstein, 1970). MPCs entail additional challenges for humans to coordinate turns. In dyadic interactions speakers always address the other interlocutor, thus it is trivial determining who the next speaker is. In MPCs the speaker may address anyone, be it a single listener, a subset of the listeners, or all of them. Therefore, deciding who should speak after a turn is yielded in a MPC is not simple given that there are multiple candidates (Schegloff, 1996). Although overlap occurrence in MPCs is similar to two-party dialogues in some cases (Shriberg et al., 2001), overlap duration has been observed to be inversely proportional to the number of simultaneously speaking parties (Laskowski et al., 2012). In general, the dynamics in MPCs differ from dyadic scenarios, hence they need special attention.

Verbal and non-verbal behaviors are adopted to ease turn shifts to overcome turn-taking issues in MPCs. Speakers tend to use cues at the end of turns to select the next speaker, such as naming the addressee. In addition, speakers do not only use gaze to indicate turn yielding as in the dyadic case, but also to address a specific listener, who is obliged to take the next turn (Auer, 2018; Sacks et al., 1978). Mouth-opening patterns also reveal relevant information to predict next speakers in MPCs (Murai, 2011). Ishii et al. (2019) found that the next speaker starts opening their mouth narrowly before change of turns. This phenomenon can be due to both the next speaker's ability to predict the end of turn and current speaker's skills to interpret next speaker's desire to gain the floor.

## B Paper Selection Criteria

In this appendix we describe the procedure we followed to search scientific articles for this survey. We adopted a systematic approach to identify relevant research on turn-taking modeling, with a special attention on studies published after 2021. We began with an extensive search on Google Scholar using a set of targeted keywords, including "turn-taking", "end-of-turn", "end-of-utterance", and "backchannel", combined with terms like "prediction", "detection", and "multi-party". We also conducted manual searches of proceedings from major NLP and dialogue system conferences taking place between 2020 and 2024, such as SIGDIAL, *ACL, Interspeech, IWSDS and ICASSP. To minimize the risk of missing key studies, we employed additional strategies to enhance coverage. We reviewed the Google Scholar profiles of identified scholars active in the field from the pool of articles we previously obtained to find any potentially missed publications. Finally, we also examined recent citations from our pool of papers to identify emerging research. Through these efforts, we aimed to provide a thorough and representative overview of the state of research in turn-taking modeling, ensuring that this survey reflects the latest developments in the field. As a result, this survey describes new methods and corpora included in more than 35 papers published after 2021.

Figure 1 shows the distributions, by subtask, of publication years of the articles we report in this survey. We observe that around 65% of the studies on end-of-turn and backchannel prediction included in this survey were published between 2021 and 2024. On the other hand, less than 30% of the works we found on multi-party conversations were published in the same time span. These findings confirm the lack of contributions and slow progress in MPCs and turn-taking modeling research.

## C Turn-taking Events and Phenomena

In this appendix we define relevant events and phenomena in turn management in natural

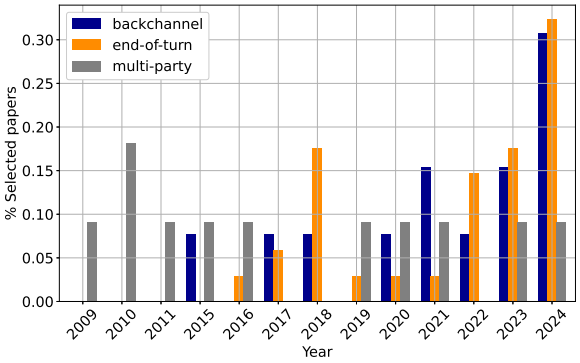


Figure 1: Distribution across the years of the papers we describe in Sections 4, 5 and 6.

dialogues. Figure 2 depicts distinct elements and events occurring in spoken dialogue turn-taking management.

**Inter-Pausal Units (IPUs)** are speech segments preceding a silence of a certain duration. IPUs correspond to the main pieces of information exchanged by interlocutors. **Pauses** are silences between two consecutive IPUs of the same speaker. **Gaps** are silences between two consecutive turns of different speakers. **Backchannel** are lexical or non-lexical sounds provided as a feedback by a listener in the dialogue. These expressions are usually used to indicate to the speaker that the listener understands or acknowledges what the speaker says, without the intention of interrupting. A backchannel is not considered as a turn. An **Overlap** takes place when IPUs from distinct speakers are produced at the same time. Usually, they occur at turn shifts.

## D Turn-taking Cues

In this appendix we explain the cues both speakers and listeners use to anticipate turn completions. A long silence after a speaker's utterance is the most basic form of cues indicating that the speaker has completed their turn. It is not enough though to detect when a turn-shift should occur. In practice, combinations of cues such as gaze, prosody (i.e. voice volume, intonation, etc.), syntactic completeness and body gestures are used to predict when to take the turn. Although there exist some differences in the use of cues across languages due to cultural or grammatical aspects, most languages follow similar patterns (Stivers et al., 2009). Prosody is one of the most studied cues for turn-yielding prediction. The prosodic structure of speech carries turn-taking cues in three dimensions: fundamental frequency, duration and amplitude (Cutler and Pearson, 2018). For example, in English words are uttered with longer duration in phrase-final than in non-phrase-final positions (Oller, 1973). Duncan (1972) and Cutler and Pearson (2018) observed that intermediate fundamental frequency contours maintain a mid-level pitch range, whereas either higher and lower pitch levels are found at the end of utterances. Syntactic completeness is another cue obtained from speech, which involves *what* the speaker says rather than *the form* it is spoken. Syntax and semantics are more relevant than prosody for turn-yield prediction, as it is easier for humans to predict a syntactically complete phrase than prosodic changes in speech (Sacks et al., 1978).

Body language also plays a crucial role to foresee the speaker's end of turn. Speakers tend to gaze away after taking their turns and look back again toward the listener when their speech is completed (Kendon, 1967). Generally, the participant taking the next turn is the one breaking the mutual gaze when beginning to speak (Novick et al., 1996). Hand gestures have also been extensively studied as turn-yielding and turn-holding cues (Sikveland and Ogden, 2012; Streeck and Hartge, 1992; Zellers et al., 2019). Kendrick et al. (2023) found that turns including manual gestures resulted in faster transitions than those without any. They reported that gaps between turns were approximately 150ms shorter on average when hand gestures were used. Similarly, Zellers et al. (2016) noted a relation between turn-shifts and hand gestures produced before the end of turn of Swedish speakers. Despite the previously described works have mostly studied cues in an isolated fashion, cues have been observed to have an additive effect (Hjalmarsson, 2011). In other words, humans use combinations of them to adequately manage turns in dialogues.

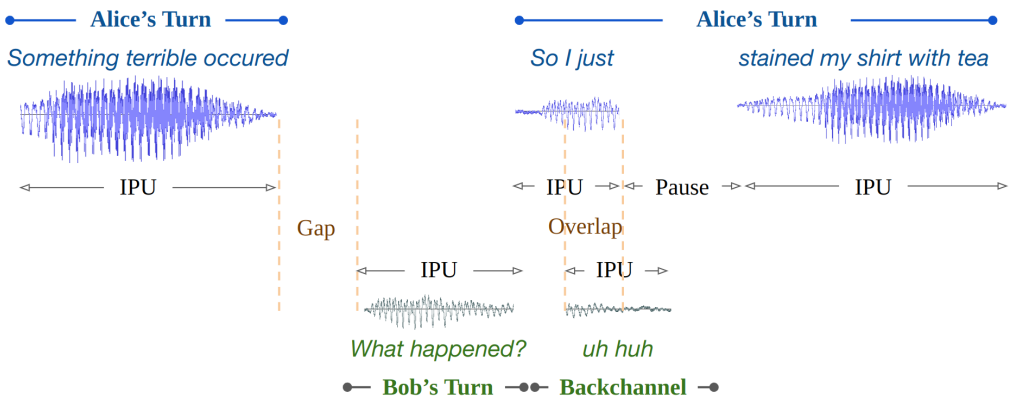


Figure 2: Turn-taking management illustration in a dyadic conversation. IPU: Inter-Pausal Unit.