technologies

MDPI

Review

Turn-Taking Modelling in Conversational Systems: A Review of Recent Advances

Rutherford Agbeshi Patamia *, Ha Pham Thien Dinh, Ming Liu and Akansel Cosgun *

School of Information Technology, Deakin University, Burwood, VIC 3125, Australia; s221475302@deakin.edu.au (H.P.T.D.); m.liu@deakin.edu.au (M.L.)
* Correspondence: r.patamia@deakin.edu.au (R.A.P.); akan.cosgun@deakin.edu.au (A.C.)

Abstract

Effective turn-taking is fundamental to conversational interactions, shaping the fluidity of communication across human dialogues and interactions with spoken dialogue systems (SDS). Despite its apparent simplicity, conversational turn-taking involves complex timing mechanisms influenced by various linguistic, prosodic, and multimodal cues. This review synthesises recent theoretical insights and practical advancements in understanding and modelling conversational timing dynamics, emphasising critical phenomena such as voice activity (VA), turn floor offsets (TFO), and predictive turn-taking. We first discuss foundational concepts, such as voice activity detection (VAD) and inter-pausal units (IPUs), and highlight their significance for systematically representing dialogue states. Central to the challenge of interactive systems is distinguishing moments when conversational roles shift versus when they remain with the current speaker, encapsulated by the concepts of "hold" and "shift". The timing of these transitions, measured through Turn Floor Offsets (TFOs), aligns closely with minimal human reaction times, suggesting biological underpinnings while exhibiting cross-linguistic variability. This review further explores computational turn-taking heuristics and models, noting that simplistic strategies may reduce interruptions yet risk introducing unnatural delays. Integrating multimodal signals, prosodic, verbal, visual, and predictive mechanisms is emphasised as essential for future developments in achieving human-like conversational responsiveness.

Keywords: turn-taking; conversational systems; human–robot interaction; multimodal; predictive modelling

check for updates

Academic Editors: Damith Herath, Janie Busby-Grant and Xing Wang

Received: 11 October 2025
Revised: 5 December 2025
Accepted: 9 December 2025
Published: 15 December 2025

Citation: Patamia, R.A.; Dinh, H.P.T.; Liu, M.; Cosgun, A. Turn-Taking Modelling in Conversational Systems: A Review of Recent Advances. *Technologies* **2025**, *13*, 591. https://doi.org/10.3390/technologies13120591

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Everyday interactions depend on a shared rhythm of turn-taking that signals when to speak or pause, much like the moves in a simple game of Tic-Tac-Toe. Since speaking while also processing someone else's words can be challenging, we rely on cues such as vocal inflection and body language to indicate whose turn it is to communicate. This review aims to investigate the mechanisms of turn-taking in conversational systems, drawing on years of study across various fields that have explored these subtle signals to develop technology that replicates this natural exchange. By aligning computers more closely with our innate communication styles, we can enhance accessibility for everyone, thereby eliminating the need for specialised training.

As a careful and engaged conversational partner, you, the reader, process each statement in the same way a listening audience takes in spoken words. Your extensive experience with interpersonal dynamics allows you to identify when to respond, recognise opportune

moments to interject and understand when it is better to let others speak. You may recall instances when you unintentionally spoke over someone, or perhaps did so on purpose, waiting for a pause to express your viewpoint. You also know the excitement of sharing an interesting discovery or presenting information that might capture another person's interest. A simple question can shift your role from speaker to listener, and slight changes in vocal tone can help you maintain or smoothly yield the conversational lead. You are aware of the usual greeting patterns, those brief "How are you?" exchanges guiding our interactions, whether reconnecting with someone familiar or meeting a new person for the first time. By your fiftieth birthday, you have likely processed nearly 100,000 h of spoken conversation [1], averaging about 16,000 spoken words daily [2]. These countless interactions have sharpened your ability to smoothly transition between being a speaker and a listener, providing a strong basis for exploring how conversation fundamentally influences social interaction.

During verbal interactions, conversation partners typically manage speaking and listening in a synchronised process referred to as turn-taking. This dynamic arises not only among humans but also across a variety of animal species [3,4]. Notably, humans begin to develop turn-taking competencies early in infancy: "proto-conversations," where caregivers and infants alternate vocalisations, emerge long before any formal language structures are in place [5,6]. These basic patterns are remarkably consistent across different linguistic and cultural settings, suggesting an underlying reliance on predictive abilities [7–9]. Such abilities enable rapid exchange, frequently within a window of about 200 ms [10,11], which is notably faster than the time generally required to formulate a complete utterance [10].

While humans navigate transitions in conversations almost effortlessly, many current conversational systems still struggle to handle smooth turn-taking. Common issues include poorly timed responses, such as delayed acknowledgments or premature interruptions, as well as failures to recognise when a speaker's turn begins. These problems can result in overlaps or unfilled pauses, which can frustrate users and lead to miscommunication. For instance, if a virtual assistant interrupts a user mid-sentence, it can disrupt their train of thought; conversely, if the assistant waits too long to respond, the interaction may feel awkward. As artificial intelligence (AI) and social robotics continue to develop, the ability to facilitate seamless spoken interactions has become increasingly important. Recent advances in text-based conversational agents, such as GPT-4 [12] and LLaMA [13], have made significant progress in generating coherent, contextually relevant dialogues. However, transitioning to spoken interaction presents additional challenges. The desire to engage in verbal interaction with sophisticated language models is growing, yet ensuring natural, real-time turn-taking remains an unresolved research issue. Existing studies on turn-taking have explored various aspects, including the detection of turn-end signals, the management of interruptions, and the generation of turn-taking signals. However, several critical gaps still exist. Many current approaches struggle to develop general turn-taking models that can effectively adapt to different interaction styles while remaining robust in human–computer dialogues. Additionally, real-time response timing remains an area of ongoing investigation, requiring models beyond merely minimising pauses and overlaps to capture the nuances of fluid conversation. Further, there is a need for deeper integration of pragmatic understandings, such as recognising the contextual appropriateness of speaking at a given moment, within computational frameworks. Moreover, advancements in predictive modelling, particularly leveraging Large Language Models (LLMs), have opened new possibilities for more sophisticated conversational overlaps beyond simple backchanneling. These limitations in the current research landscape provide the motivation for this review.

In this review, we integrate insights from human communication research, evaluate contemporary approaches to managing turn-taking in conversational technologies

and human–robot interactions, and discuss avenues for ongoing inquiry. At first glance, turn-taking may appear so self-evident that it scarcely merits close study. Yet the next section demonstrates how the seemingly effortless flow of conversation depends on subtle coordination strategies that are easily overlooked without specialised attention. To better organise and contextualise the broad literature on conversational turn-taking, we present a fine-grained survey architecture that categorises existing work into five interlocking domains: (1) Background & Foundations (temporal dynamics, classic dialogue models, and turn-construction theory); (2) Core Phenomena & Cues (predictive timing, overlaps/backchannels, multimodal signals, and multi-party floor management); (3) Data Resources (key corpora such as HCRC Map Task, Switchboard, Fisher, CANDOR, and NoXi); (4) Computational Approaches (from end-of-turn detection and incremental prediction through backchannel modeling, real-time evaluation, foundational deep-learning architectures, and multimodal fusion); and (5) Future Directions (richer multimodal integration, open multilingual corpora, context-aware backchannel frameworks, and human-like latency models). This hierarchical organisation is illustrated in Figure 1. This review article follows a structured organisation to thoroughly explore turn-taking mechanisms in conversational systems. The paper begins with an overview of foundational linguistic concepts related to turn-taking, highlighting its historical and theoretical underpinnings. Next, we will discuss the cues that facilitate turn-taking across multiple modalities, including speech, gestures, and gaze.

Following this, the article delves into four primary aspects of turn-taking research in conversational systems that have garnered significant attention:

- What signals are involved in the coordination of turn-taking in dialogue?
- How can the system identify appropriate places and times to generate a backchannel?
- How can real-time turn-taking be optimised to adapt to human–agent interaction scenarios and evaluated through a user study involving real-world interactions?
- The handling of multi-party and situated interactions, including scenarios with multiple potential addressees or the manipulation of physical objects, will also be addressed.

Finally, the review identifies promising directions for future research, emphasising gaps such as the need for richer multimodal integration, expanded real-world testing, and cross-linguistic studies. While most computational modelling research in this field has been conducted on English data, this review will include some notable exceptions. If the language of study is not explicitly mentioned, it should be assumed to be English. In this review, we define "recent advances" as research on computational modelling of turn-taking published from late 2021 to the time of publication. This time window reflects the emergence of neural, self-supervised, and multimodal approaches that have reshaped the field in recent years. Foundational studies, including classical work in linguistics, phonetics, and conversation analysis prior to 2021, were used primarily to contextualise or motivate contemporary modelling approaches. Our analysis and synthesis of methods in Section 4, therefore, focus exclusively on developments within this time period.

Key Terms

- Transition-Relevance Place (TRP) refers to a moment in conversation where a speaker change can appropriately occur.
- Predictive turn-taking describes the use of cues that precede a TRP to anticipate turn completion and plan response timing.
- Voice Activity Projection (VAP) denotes models that forecast near-future speech activity to coordinate responses or backchannels in real time.

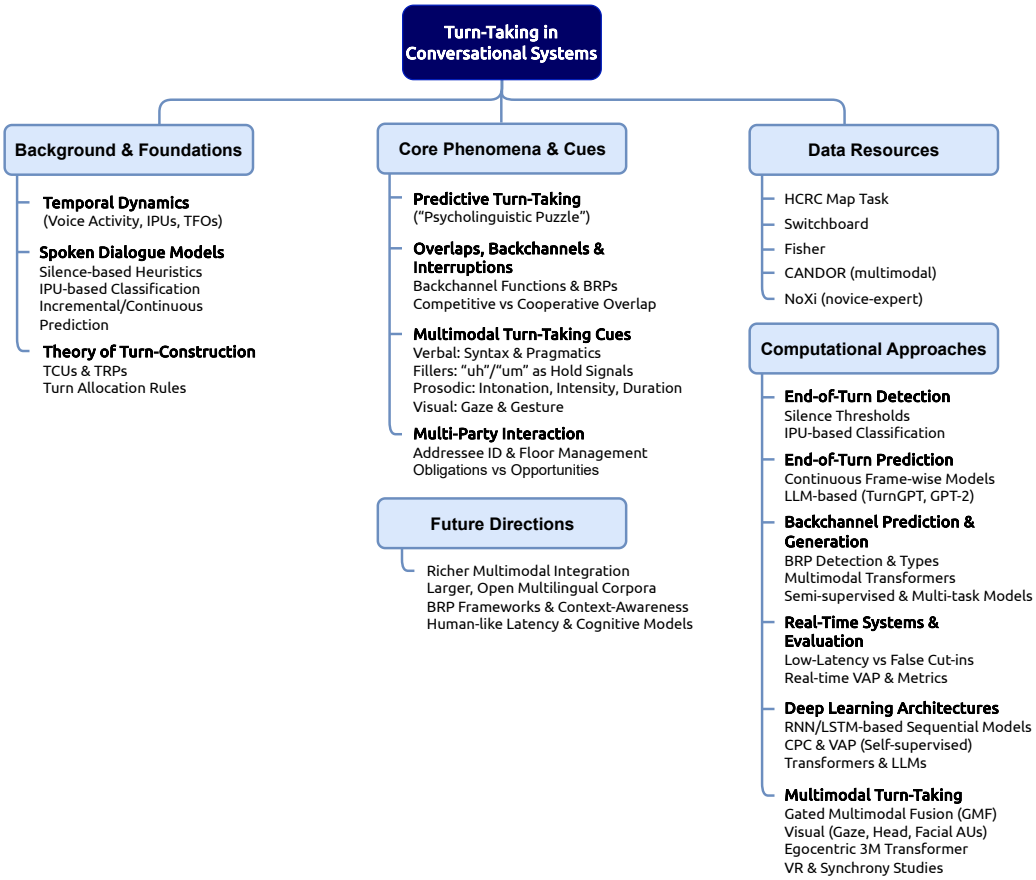


Figure 1. Fine-grained mind-map of turn-taking research in conversational systems, showing how theoretical foundations, observed phenomena, data sets, algorithmic approaches, and future challenges interrelate.

2. Background

A defining characteristic of human social interaction is the rapid alternation between speaker and listener roles, commonly referred to as turn-taking. This mechanism is a fundamental aspect of language use, observed universally across cultures, shaping the interactional environment in which children acquire language and in which language itself is believed to have evolved. Additionally, turn-taking is not exclusive to humans; it has been documented in both vocal and gestural communication among various non-human species [14–16]. In human conversation, the precise timing of turn exchanges raises intriguing questions about the underlying processes. Speakers typically alternate with minimal pauses on average, nearly matching the human response threshold of 200 ms and with minimal overlap (usually less than 5% of the total speech) [10]. This seamless exchange is particularly remarkable given that individual turns can vary unpredictably in length and content, the number of participants may fluctuate, and longer turns (as when narrating a story) must occasionally be accommodated. For example, during a casual conversation among friends, one speaker might briefly share news, prompting quick responses from others. Later, the conversation shifts to storytelling, with one person speaking uninterrupted for several minutes. Despite these variations, the conversational flow remains smooth, underscoring the human capacity to adjust to conversational demands dynamically. For example:

Speaker A: "Hey, did you hear about the concert next weekend?"
Speaker B: "Oh, yeah! I got my tickets yesterday."
Speaker C: "Same here, can't wait!"

(*Later in the conversation*)
Speaker A: "So, this reminds me of when I attended my first concert years ago. It was raining, and we were waiting outside for hours…" (Speaker A continues narrating uninterrupted for several minutes while others listen attentively.)

The theoretical foundation of turn-taking was initially proposed by American sociologist Harvey Sacks. However, the precise definition of a turn has varied among scholars as turn-taking theory has evolved. The foundational model of turn-taking in everyday conversation was introduced in 1974 by Sacks et al. [8], emphasising the structural organisation of dialogue without explicitly defining what constitutes a turn. This model conceptualises turn-taking as a systematic process governed by two primary components: the turn-constructional component and the turn-allocation component [8]. The turn-constructional component defines the linguistic building blocks of a turn, known as turn-constructional units (TCUs), which may consist of words, phrases, clauses, or complete sentences. The turn-allocation component, on the other hand, governs how the next speaker is selected, either explicitly through the current speaker's selection (via direct address or gaze) or implicitly through self-selection, where the first participant to initiate speech gains the turn. If neither selection occurs, the current speaker retains the floor and continues speaking, reapplying the rules to prevent excessive silence [17–20]. The Transition-Relevance Place (TRP) marks the point at which a turn is likely to end, and the speaker's transition becomes relevant. These TRPs align with syntactic, prosodic, and semantic boundaries, facilitating smooth exchanges [17,21–23]. Research has shown that TRPs often coincide with intonational phrase boundaries, reinforcing their importance in turn coordination. However, variations occur depending on conversational context, as specific discourse structures, such as storytelling or explanatory sequences, require extended turns [24]. Beyond these core components, Sacks and his colleagues identified fourteen features characterising turn-taking organisation, illustrating the intricate balance between holding, yielding, and abandoning turns during conversation. These findings underscore that turn-taking is a rule-governed yet flexible system designed to minimise silence and avoid overlapping speech, ensuring conversational coherence and efficiency. The definition of a turn has also evolved, with [25] describing it as a speaker's continuous utterance within a conversation, concluding when the speaker-listener roles shift, all participants remain silent, or a designated signal indicates a transition. This widely accepted definition aligns with the perspective of [26], which distinguishes between the potential to assume the speaker role and the actual verbal output of the turn, despite differences in theoretical framing. A consistent thread across these perspectives highlights turn-taking as both a structural and an interactive process, essential for maintaining conversational order and efficiency.

The following section explores notable turn-taking phenomena during the initial 15 s of a telephone call between two unfamiliar speakers https://catalog.ldc.upenn.edu/LDC2004S13 (accessed on 8 December 2025). In general, most conversations begin with a greeting phase that serves as a foundational exchange. This greeting phase typically comprises mutual salutations, self-introductions, and brief acknowledgements, which are examples of adjacency pairs [8]. Unlike written exchanges, where a user explicitly sends a complete message (e.g., pressing "enter"), spoken interactions allow participants to speak simultaneously or interrupt each other's turns. Determining when a speaker has finished an utterance can therefore be far from trivial.

Figure 2 displays the spectrograms and corresponding waveforms of two participants, labelled Speaker A and Speaker B. Each spectrogram shows frequency distribution (vertical axis) over time (horizontal axis), with brighter regions indicating higher acoustic energy. The waveforms, aligned with each spectrogram, highlight amplitude variations of each

speaker's voice. Textual labels placed near the waveforms indicate the approximate timing and content of each utterance.

The conversation opens with both speakers greeting one another almost in unison ("hello," "hello"). Immediately afterwards, Speaker A extends a more elaborate introduction ("Hi, this is Deena"), to which Speaker B responds similarly ("Hi, I'm Donna"). This is followed by A's question ("So do you have pets?"), accompanied by a short laugh and a brief silence. During this pause, B produces a hesitant response ("ah no"), suggesting uncertainty or momentary confusion, as seen in the low-energy portion of B's waveform. Shortly after, A provides additional information, mentioning multiple dogs ("I have three dogs [laughter]"), while B quickly reacts with an acknowledgement ("Oh okay"). The slight overlap in their waveforms indicates that B's response starts before A fully finishes laughing.

From a technical perspective, this brief exchange highlights key challenges in spoken dialogue research. The presence of overlapping talk, evident from the partial superimposition of waveforms, raises questions about how interlocutors (human or computational) detect whether a turn is still in progress. Likewise, B's hesitation noise ("ah") might be interpreted by an automated system as a more substantive utterance, potentially leading to premature interruption or misinterpretation of turn boundaries. Even within the first few seconds, these subtle overlaps and micro-pauses can disrupt a naive turn-taking algorithm designed for strictly sequential exchanges.

Despite the apparent ease with which humans navigate such interactions, these early moments in a conversation exemplify how spoken dialogue systems can falter. Determining whether a hesitation marks the start of a new turn or simply a filler within an ongoing turn remains an unresolved problem. Furthermore, deciding when to respond, especially if the other speaker has not finished a word or is laughing, poses an equally complex challenge. Although humans have an intuitive grasp of these cues, computational models often lack robust methods for managing rapid turn-taking without unnatural delays or cutoffs. These observations underscore the difficulty of designing computational models that handle rapid turn-by-turn exchanges. Even within the first 10–15 s, multiple overlapping signals appear, such as laughter, hesitations, and truncated phrases. Importantly, there is no comprehensive theory that explains precisely how humans achieve these seamless exchanges, and current models often struggle to replicate this efficiency. As a result, developing robust turn-taking mechanisms remains a key objective for advancing spoken dialogue systems.

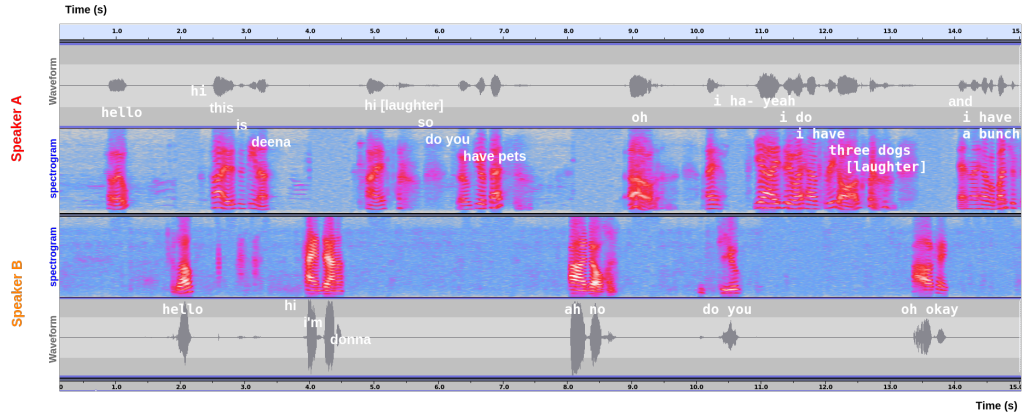


Figure 2. Visualisation of the initial moments in a telephone conversation between two unfamiliar speakers. The top panel displays Speaker A's waveform and spectrogram, while the bottom panel shows Speaker B's. Brighter regions in each spectrogram indicate higher acoustic energy. The spoken words, such as greetings ("hello," "hi") and references to pets ("I have three dogs [laughter]"), are overlaid at their approximate time points. Vertical text placement is adjusted to enhance readability.

2.1. Temporal Dynamics in Turn-Taking

From the perspective of turn-taking, the precise timing of conversational contributions, specifically when individuals begin or end speaking and listening roles, is critical. Timing determines conversational fluidity and directly influences how roles transition smoothly between speakers. The foundational representation of timing information in dialogue research is encapsulated in Voice Activity (VA), a binary encoding indicating speaker activity (active/inactive). Initially conceptualised by [27], voice activity serves as a fundamental tool to examine conversational timing, effectively capturing periods during which a participant contributes vocally.

In practical applications, Voice Activity Detection (VAD) can identify periods of speech activity at varying granularities. Typically, frames of around 20 ms (corresponding to a frame rate of 50 Hz) are utilised. These frames help define Inter-Pausal Units (IPUs), speech segments from a single speaker separated by brief silences of less than 100 ms. This method omits brief intra-speaker pauses, which commonly occur between words, and focuses analysis on more significant conversational segments. Within dyadic conversations, the binary states of voice activity for two speakers combine to produce distinct dialogue states: single-speaker activity (A or B exclusively), overlapping speech (both speakers active), and mutual silence. Following [7], the dialogue state transitions are identified as gaps (silence between IPUs from different speakers), pauses (silence between two IPUs within the same speaker, in other words, silence during speaker holds), overlaps-between (overlaps at speaker transitions), and overlaps-within (overlaps starting and ending with the same speaker). A visual representation of these dialogue states and their transitions enhances clarity (as illustrated in Figure 3).

In natural dialogue, mutual silence frequently signals a pause in active speech, during which conversational roles either shift to a new speaker or remain with the current speaker, referred to as "shift" and "hold," respectively. Differentiating between these scenarios represents a central challenge for interactive turn-taking systems. Analysing these moments can be enhanced by using the concept of Turn Floor Offset (TFO), defined as the interval between the end of one speaker's utterance and the beginning of the next speaker's utterance. Negative TFO values indicate overlapping speech transitions, whereas positive values represent gaps in speech. Studies consistently observe that TFO durations align closely with minimal human reaction times (200 ms), substantially faster than typical speech production times (600–1500 ms) [10,28]. This suggests an underlying biological mechanism, possibly arising from the mutual synchronisation of brain oscillations related to speech rhythm [29]. However, cross-linguistic variations in TFO durations, such as faster transitions in Japanese (0.01 s) compared to slower transitions in Danish (0.47 s), highlight cultural modulation of this phenomenon [30]. Additionally, serial dependencies in TFO durations indicate collaborative coordination between interlocutors rather than individual speaker behaviours [31]. These findings provide valuable guidelines for developing spoken dialogue systems (SDS). Given the dominance of single-speaker activity, SDS should prioritise minimising interruptions by responding primarily during apparent silences. A silence-based heuristic, which suggests waiting about one second before taking the floor, significantly reduces unintended interruptions, although at the risk of appearing sluggish relative to human latency. Additionally, since overlaps frequently occur, especially those signalling a speaker shift, dialogue systems could adopt a policy that treats the emergence of overlap during their speech as a user's intent to speak, prompting the system to yield the turn promptly. Despite these considerations, overly cautious strategies may inadvertently provoke confusion or unnecessary reengagement from users, indicating the need for more sophisticated turn-taking models to effectively mimic the nuanced timing and responsiveness of human interaction [9,32].

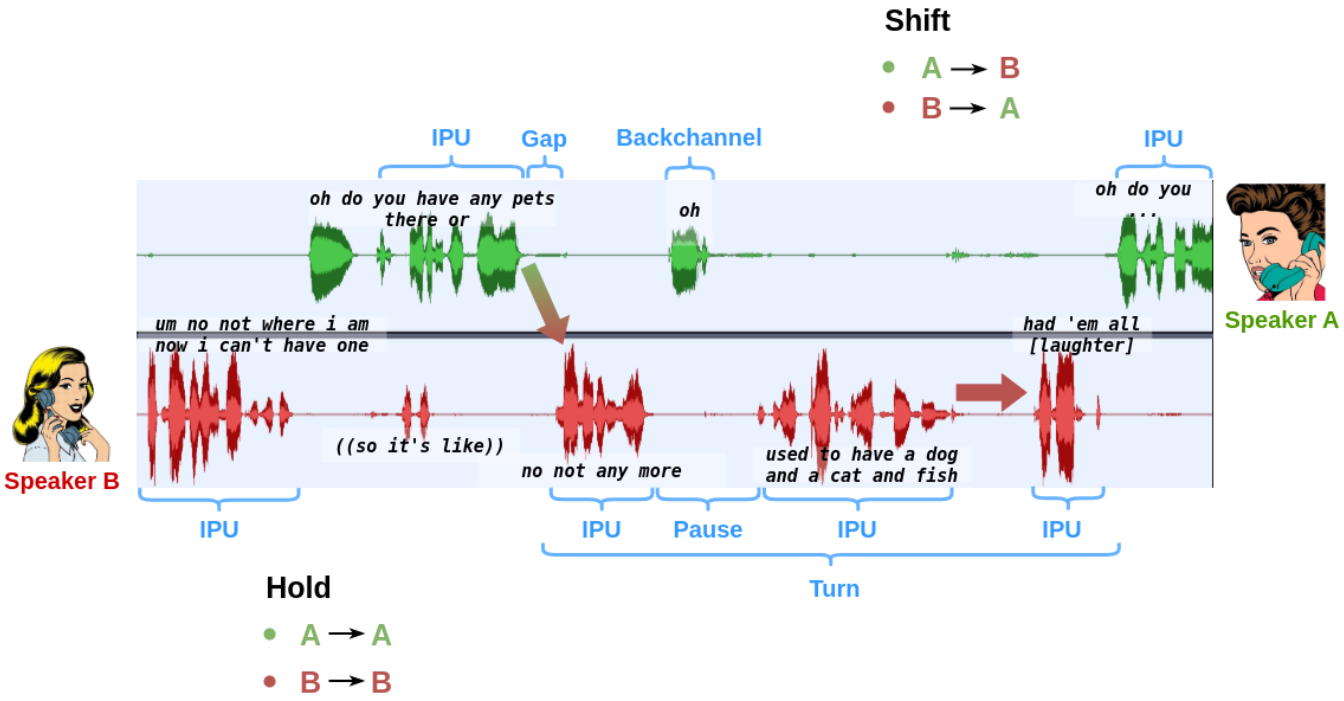


Figure 3. Visualisation of dialogue activity states during a brief conversational exchange between two speakers (Speaker A (green) and Speaker B (red)). The annotated waveforms highlight the practical aspects of timing in turn-taking, specifically showing how subtle timing differences distinguish among turn-yielding, holding, and overlapping speech.

2.2. Implementing Turn-Taking in Spoken Dialogue Systems

The practical implementation of turn-taking in computational models has largely overlooked the concept of predictive turn-taking. Traditionally, spoken conversational systems have been developed similarly to text-based exchanges, where responses are generated only after the preceding contribution is entirely made. In written dialogue, conversations proceed through complete messages exchanged sequentially, with participants responding only after fully receiving a message. This structured format explains why many spoken dialogue systems have traditionally used a silence-based policy to determine the end of a speaker's turn by detecting silence exceeding a predefined threshold. This method is depicted in the upper part of Figure 4. Turn-taking models typically differ in how they identify Transition-Relevance Places (TRPs), or points at which speakers can appropriately exchange turns [8]. However, silence-based models are not well-suited to accurately predict TRPs, as silence alone does not always indicate turn completion. Natural conversation frequently contains pauses that are not necessarily turn-ending signals but can indicate hesitation, emphasis, or the planning of further speech. Consequently, silence-based models often misinterpret natural pauses as TRPs, leading to unnatural interruptions or overly delayed responses. IPU-based models improve upon this by analysing inter-pausal units (IPUs), which are short speech segments separated by brief silences, typically around 200 ms. These models classify each IPU boundary as either an actual TRP or a continuation of the current speaker's turn, as shown in the middle row of Figure 4. Nevertheless, even IPU-based models remain limited by their reliance on predetermined heuristics to identify potential transition points, restricting their ability to predict future TRPs or suitable moments for backchannel responses. The most advanced turn-taking models are incremental or continuous models that actively analyse speech and silence in real time, independently identifying potential TRPs, as depicted at the bottom of Figure 4. By operating continuously, these

models are capable of predicting upcoming TRPs even during ongoing speech, enabling smoother conversational transitions and timely backchanneling.

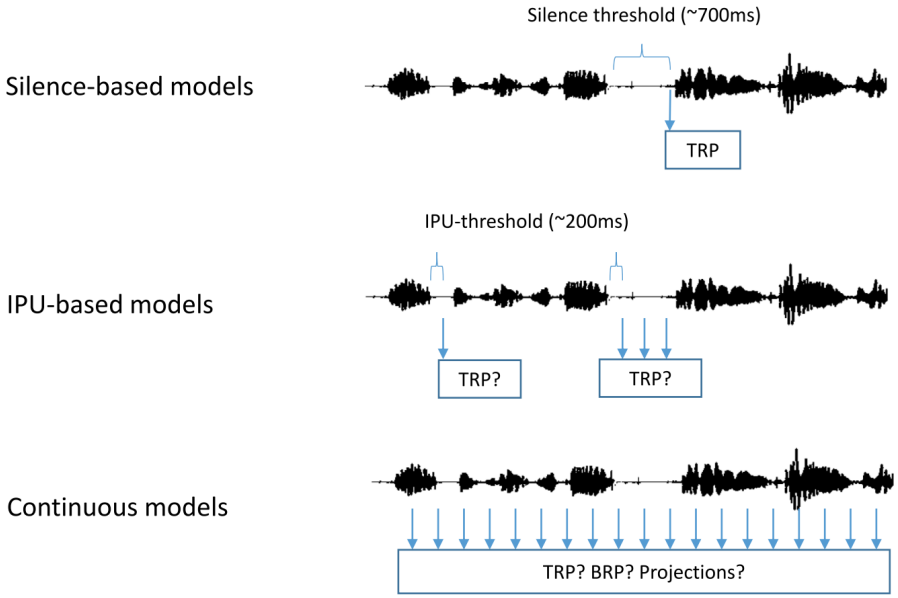


Figure 4. Illustrations of three turn-taking models. (**Top**): A silence-based model employs Voice Activity Detection (VAD) to detect the end of the user's utterance, then uses a predefined silence threshold to decide when to take the turn. (**Middle**): An IPU-based model identifies Inter-Pausal Units (IPUs) as potential turn-taking points via VAD, analysing cues in the user's speech, such as pause length, to determine if the turn is yielded. (**Bottom**): A continuous model processes speech and silence in real time to predict Transition-Relevance Places (TRPs) during both pauses and ongoing speech. This model can also detect backchannel-relevant places (BRPs) and make projections. Source: [9].

2.3. Predictive Turn-Taking

One prominent question in research on conversational turn-taking is the so-called "psycholinguistic puzzle" [10], which highlights the paradox that people often respond to each other in approximately 200 ms. Yet, psycholinguistic estimates place the time needed to formulate and articulate an utterance at 600 ms or more [10]. If listeners waited passively until the speaker's final word to begin preparing their responses, achieving such rapid turn transitions would be nearly impossible, [10,11,33]. Instead, mounting evidence indicates that dialogue participants engage in continuous processing: the listener actively anticipates turn completions, formulates potential replies, and monitors cues such as syntax, semantics, and speech envelope patterns to initiate the next turn [15,34,35]. This aligns with findings that unaddressed listeners sometimes look to the next speaker even before the current turn finishes [36], suggesting they have predicted not only when the turn will end but also the likely content of the utterance. By planning mid-turn responses, listeners effectively overcome the constraint that articulation can take hundreds of ms to begin. Therefore, a more nuanced view combines predictive mechanisms that allow the listener to guess when the speaker is concluding and reactive cues at the end of the turn, confirming that the utterance is complete [7,10]. This dual perspective helps explain how humans resolve the core of the "psycholinguistic puzzle," enabling turn-taking at conversational speeds that far outstrip the essential speech production times [11,15], as illustrated in Figure 5.

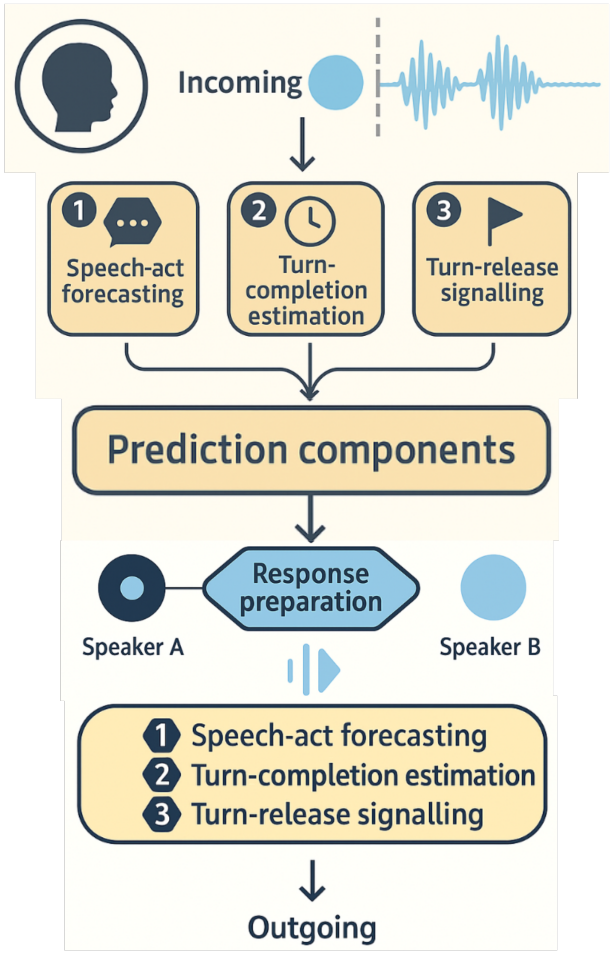


Figure 5. Human speech production during spoken dialogue interactions. While listening to the ongoing speech of the speaker, the listener plans, prepares, and detects a suitable transition point to execute their response.

2.4. Overlaps, Backchannels and Interruptions

Backchanneling plays a crucial role in human conversation, enabling listeners to provide real-time feedback without interrupting the flow. Initially conceptualised by [37], backchannels are brief, non-intrusive responses such as "uh-huh," "mm-hm," or "yeah," which serve to indicate active listening and engagement without signalling an intention to take the turn. Beyond verbal affirmations, nonverbal cues such as nodding, facial expressions, and gaze shifts also serve as backchannel responses, reinforcing the listener's attentiveness. These feedback mechanisms contribute to the smooth progression of discourse by ensuring speakers feel acknowledged and encouraged to continue their turn. As a result, backchanneling has become a significant area of study in spoken dialogue systems (SDS) and human–robot interaction (HRI), where accurately modelling such responses is essential for achieving natural and fluid conversations. Backchannels differ from traditional turn-taking mechanisms in that they do not claim the conversational floor; instead, they operate within a separate communicative track. Ref. [38] categorises these as occurring in "Track 2," running parallel to the primary discourse in "Track 1." This distinction is particularly important in SDS and HRI applications, as failing to identify backchannels correctly can lead to misinterpretations, resulting in a system either interrupting the speaker or failing to provide appropriate feedback. Furthermore, determining what precisely constitutes a backchannel is not always straightforward, as backchannels exist on a spectrum ranging from simple acknowledgements (e.g., "uh-huh") to more extended interjections (e.g., "interesting," "yeah right") that may overlap with discourse markers [39].

Because backchannels frequently occur during overlapping speech, they likely account for a substantial portion of the observed overlap in natural dialogues. However, distinguishing between cooperative and competitive overlaps remains an ongoing challenge in SDS and HRI system development.

Overlapping speech occurs frequently in conversation, but not all overlap constitutes a communication breakdown. Ref. [40] argues that overlap should not merely be seen as "failed" turn-taking, as it often facilitates conversational fluency. Ref. [41] distinguishes between competitive and cooperative overlaps, where competitive overlaps reflect an attempt to seize the conversational floor, while cooperative overlaps serve as collaborative mechanisms that enhance dialogue continuity. Among cooperative overlaps, backchannels (continuers) such as "uh-huh" and "mm-hm" hold a unique position, as they signal continued attention rather than an attempt to take the turn. Other cooperative overlap types include terminal overlaps, where a listener predicts the end of the turn and starts speaking slightly before its completion, conditional access to turn, where a listener helps complete an utterance (e.g., recalling a forgotten name), and choral talk, where speakers jointly produce speech, such as in laughter or simultaneous greetings [42]. The acoustic and prosodic properties of backchannels help differentiate them from full-fledged turns or interruptions. Research by [41,43] has shown that competitive overlaps, instances where a listener attempts to seize the conversational floor, are often marked by increased pitch, intensity, and abrupt timing. In contrast, cooperative overlaps, including backchannels, tend to exhibit softer intensity and align with the speaker's ongoing prosodic patterns. Similarly, competitive overlaps require resolution mechanisms, as participants must decide who retains the conversational floor. Ref. [41] found that most competitive overlaps resolve within one or two syllables, with one speaker ultimately yielding. Unlike competitive overlaps, interruptions introduce an additional layer of complexity. While overlaps can be objectively identified in a corpus, interruptions require interpretation, as they involve a participant violating the speaker's right to speak [44]. Notably, interruptions can occur without any overlap, such as when a speaker pauses mid-turn but has not yielded the floor, and another participant begins speaking [45]. This distinction is particularly relevant for SDS design, as dialogue systems must avoid prematurely taking the turn when a pause does not indicate a Transition-Relevance Place (TRP). Ref. [45] further found that interruptions, whether overlapping or non-overlapping, often feature higher intensity, pitch, and speech rate, characteristics that can help distinguish them from backchannels.

One major challenge in computational backchanneling models is accurately timing responses. Ref. [46] introduced the concept of backchannel-relevant places (BRPs), which refer to specific moments during a conversation when a speaker expects minimal listener feedback without relinquishing their speaking turn. Unlike Transition-Relevance Places (TRPs), which indicate points for speakers to exchange turns, BRPs identify times suitable for brief listener responses, like "uh-huh" or "mm-hmm," without interrupting the main speaker. Computationally identifying BRPs involves analysing speech features like pitch, loudness (intensity), and rhythm. Ref. [47] observed that backchannels in Japanese conversations typically follow approximately 200 ms after low-pitch segments, a pattern also identified in English dialogues by [48], who found rising pitch and increased intensity preceding these minimal responses. Beyond these descriptive findings, several real-time frameworks now exist for automatic BRP detection (see Section 4.2), including Voice Activity Projection variants fine-tuned for "continuer" versus "assessment" backchannels [49], transformer-based audio–visual fusion models [50], and temporal-attention classifiers robust to unbalanced data [51]. Besides prosodic cues, accurately predicting when backchannels should occur is essential. Traditional dialogue systems usually respond only after explicitly detecting that the current speaker has stopped speaking. However, research

by [11] highlights the importance of predictive processing, suggesting listeners actively anticipate when a turn might end and the type of response required. This view aligns with the entrainment hypothesis from [29], which suggests that synchronisation between brain activity and speech rhythms helps listeners anticipate speech timing. Incorporating predictive mechanisms into spoken dialogue systems (SDS) and human–robot interactions (HRI) can make these interactions feel smoother and more natural. Despite advances, backchanneling models still face significant challenges. One major issue is the ambiguity of backchannels, which can change meaning depending on prosody and conversational context. Ref. [52] illustrated how differences in intonation can alter the meaning of responses; for example, an elongated "yeah…" with a falling pitch might indicate uncertainty rather than agreement. Similarly, ref. [53] highlighted that the word "okay" might serve multiple functions, such as a simple acknowledgement or a discourse marker, depending on the situation. Such ambiguity requires more sophisticated, context-aware models for reliable detection and generation. Another critical challenge is integrating multimodal signals. While most current models primarily handle verbal backchannels, non-verbal cues such as nodding, facial expressions, and posture changes significantly influence natural human communication. Research indicates that combining verbal and visual feedback enhances interaction quality [54]; yet synchronising these signals in real-time remains technically challenging due to variations in human response timing [55]. The incorporation of gaze-based methods for eliciting backchannels, as discussed by [54], presents a significant opportunity to enhance interaction quality in SDS and HRI contexts.

2.5. Multi-Party Interaction

Multi-party turn-taking presents a complex challenge in spoken dialogue systems (SDS) and human–robot interaction (HRI) because it requires managing multiple interlocutors simultaneously, unlike in dyadic conversations, where turn-taking transitions involve only two participants: the speaker and the listener. Multi-party interactions require mechanisms to regulate turn allocation, participation roles, and interruptions effectively [56]. These interactions introduce several layers of complexity, including dynamic shifts in speaker roles, overlapping speech, competition for the floor, and the need for efficient coordination mechanisms to maintain coherence and engagement among multiple participants [57].

In dyadic exchanges, turn-taking is often regulated through a combination of verbal, prosodic, and non-verbal cues such as intonation shifts, syntactic completion, eye gaze, and pauses [48]. However, additional mechanisms are required to manage turn allocation in multi-party settings. Speakers must determine not only when to yield their turn but also whom to address next, making the selection of the next speaker a critical component of conversation regulation [58]. One of the primary strategies for resolving this challenge is gaze coordination, in which speakers typically establish eye contact with the intended next speaker before yielding the turn [59]. In HRI, gaze tracking enables conversational agents to anticipate turn transitions and select the appropriate interlocutor based on mutual visual engagement. Moreover, head pose tracking has been recognised as a more robust alternative to gaze tracking, particularly in environments where precise eye-tracking systems are infeasible [60]. Studies have demonstrated that head orientation is a strong indicator of attention direction and turn allocation, particularly in multi-party human–robot discussions [61,62]. When combined with verbal and prosodic cues, head movement data enhances dialogue systems' ability to infer turn-taking intentions accurately.

Obligations vs. Opportunities in Multi-Party Turn-Taking

In multi-party interactions, turn-taking is not simply a matter of either yielding or holding the floor. Conversational agents must distinguish between two types of situations: obligations, instances in which the system is directly addressed and required to respond, and opportunities, instances in which the system can choose to contribute but is not obligated to do so [56]. Recognising these distinctions enables dialogue systems to engage more naturally in multi-party discussions, allowing them to adapt dynamically to the conversation context. For instance, a study on turn-taking in human–robot collaborative games found that the presence of multiple participants necessitated a system to make informed decisions about when to take a turn, based on the nature of the utterance, user engagement, and task relevance [57]. By leveraging probabilistic models, systems can score turn-taking opportunities and obligations, balancing responsiveness with conversational fluidity.

2.6. Datasets

A practical initial approach to reveal significant insights regarding turn-taking is to utilise an extensive database of recordings of individuals participating in spoken dialogue. These resources enable researchers to rapidly isolate and scrutinise specific conversation segments, extracting data pertinent to their investigative goals. Although a dataset that fully encompasses the range of human spoken communication would be ideal, such a comprehensive collection remains unfeasible in practice. Consequently, this section highlights publicly accessible and highly utilised datasets that focus on turn-taking, backchanneling, multi-party interaction, and multimodal turn-taking, which have significantly influenced the study of conversational dynamics. These datasets were chosen for their availability, established use among researchers, and overall contribution to the field of conversation analysis.

2.6.1. HCRC Map Task Corpus

The HCRC Map Task Corpus [63] is a unique dataset of unscripted, task-oriented dialogues designed to study spontaneous speech and the process of achieving communicative goals. It involves 128 dialogues (approximately 15 h) between pairs of participants collaborating to reproduce a route on a map. The corpus systematically manipulates the familiarity of speakers (familiar/unfamiliar) and the presence or absence of eye contact. The landmark names used in the task were chosen to facilitate phonological studies. While split-screen video recordings were made, digital audio recordings and verbatim orthographic transcriptions with detailed markup of spoken phenomena, such as filled pauses and interruptions, were the primary data modalities. The Map Task is highly relevant for examining how turn-taking is organised in a goal-oriented setting and how different contextual factors (familiarity, visual interaction) influence dialogue. It can also offer insights into the role of verbal feedback, similar to backchanneling, in guiding the task. It focuses on dyadic interaction within a specific communicative task. Its strengths include the controlled elicitation and the systematic variation of social cues. However, its task-specific nature might limit the generalizability to entirely natural conversations, and the participant pool was primarily comprised Scottish English speakers. The corpus was initially available on CD-ROM and may be accessible through academic repositories.

2.6.2. Switchboard Corpus

The SWITCHBOARD Corpus [64] is a foundational resource for research in speech processing, particularly in speaker authentication and extensive vocabulary speech recognition. It comprises a vast collection of spontaneous conversational speech and text automatically captured over telephone lines. The corpus includes approximately 2500 conversations from 500 speakers representing various dialect regions across the U.S., amounting to over

250 h of speech and nearly 3 million words of text. A key feature of this corpus is the time-aligned, word-for-word transcription that accompanies each recording. The data acquisition process was automated, which helped ensure consistency and reduce the risk of experimenter bias. Additionally, demographic information about the speakers, such as age, gender, education, and dialect, is stored in a database linked to the call details (including the date, time, and length of the conversation). SWITCHBOARD is particularly relevant for studying dynamics such as turn-taking in telephone conversations, the linguistic realisation of backchanneling, and variations in speech related to turn management among different speakers. Its size and diverse speaker population are significant strengths, making it a cornerstone for training and testing speech algorithms. However, the corpus has some limitations, including the constraints of telephone bandwidth, which can affect audio quality, and the absence of direct multimodal information, such as video or facial cues. The SWITCHBOARD corpus is publicly available through the National Institute of Standards and Technology (NIST).

2.6.3. Fisher Corpus

The Fisher Corpus [65], designed under the DARPA EARS initiative, focuses on English conversational telephone speech. Its primary goal was to provide a large volume of transcribed telephone speech to advance automatic speech recognition (ASR) technology. The collection aimed for 2000 h of conversational speech from many subjects, with individual calls lasting no more than ten minutes. A distinguishing characteristic is its platform-driven collection protocol, in which the system initiated most calls and participants spoke on assigned topics selected at random to encourage a broad vocabulary. The corpus aimed to represent subjects across various demographic categories, including gender, age, dialect, and English fluency. Fisher is relevant for studying turn-taking in the context of specific topics and a diverse speaker base over the telephone. While not its primary focus, the linguistic patterns of backchanneling within these conversations could also be investigated. It focuses on dyadic interaction. The corpus's strength lies in its large size and demographic diversity, designed to maximise inter-speaker variation. However, it shares the bandwidth limitations of telephone speech and lacks multimodal data. The Fisher Corpus was collected by the Linguistic Data Consortium (LDC), and access is typically granted through them.

2.6.4. CANDOR Corpus

The CANDOR (Conversation: A Naturalistic Dataset of Online Recordings) corpus [66] represents a significant advancement in conversational interaction datasets by offering a sizeable multimodal collection of naturalistic conversations recorded over video chat. It encompasses 1656 conversations from 1456 unique participants, resulting in over 7 million words and 850+ h of audio and video. A unique aspect includes moment-to-moment vocal, facial, and semantic expression measures derived through a sophisticated computational pipeline. This includes textual analysis of semantic novelty, acoustic analysis of loudness and intensity, and visual analysis of facial expressions (e.g., happiness) and head movements (e.g., nods, shakes). The corpus also features detailed post-conversation survey data capturing participants' perceptions and feelings. Conversations were unscripted, conducted between strangers, and took place during 2020, offering a unique snapshot of discourse during a tumultuous year. CANDOR is highly pertinent to turn-taking (analysing gaps, overlaps, and turn duration with various algorithms), backchanneling (analysing frequency and potential functions using computational models), and multimodal turn-taking by integrating visual cues with spoken turns. It provides a rich, multimodal view of natural conversation. It explores the interplay between low-level, mid-level, and high-level

conversational features, including psychological well-being and perceptions of conversational skill. Challenges include potential selection bias from voluntary participation and limitations in generalising findings beyond English speakers in the US context. The corpus is publicly available, promoting interdisciplinary collaboration.

2.6.5. NoXi Database

The NoXi (natural dyadic novice–expert interactions) Database [67] is a novel multimodal and multilingual corpus of screen-mediated novice-expert interactions focused on information exchange and retrieval. It contains over 25 h of recordings of dyadic interactions in seven languages (mainly English, French, and German) across 58 diverse topics. The dataset features synchronised audio (room and close-talk microphones), video, depth data, and tracking data for skeleton, facial landmarks, head pose, and action units. A key distinguishing feature is the emphasis on unexpected situations, including controlled and spontaneous conversational interruptions. The corpus includes rich continuous and discrete annotations of low-level social signals (gestures, smiles, head movements), functional descriptors (turn-taking, dialogue acts), and interaction descriptors (engagement, interest). It also provides voice activity detection, communicative state, and turn transition annotations. NoXi offers a valuable resource for studying turn-taking (with explicit annotations), backchanneling (through engagement cues and social signals), and especially multimodal turn-taking in a mediated communication setting. The inclusion of novice-expert dynamics and interruptions provides unique research opportunities. Its strengths lie in its multimodality, multilingualism, focus on a specific interaction type with a knowledge differential, and detailed annotations. Limitations include the screen-mediated setting and the particular context of information retrieval, which might influence conversational patterns. The NoXi Database is freely available through a dedicated web interface for research and non-commercial use.

Having outlined the core corpora and the functionalities they offer (such as dyadic versus multiparty interactions and audio-only versus multimodal formats), we will next analyse the interactional cues that these datasets are designed to highlight. This includes examining textual, acoustic, prosodic, and visual signals that influence turn projections and Transition-Relevance Places (TRPs) during conversations.

3. Analysing Dialogue Content: Cues for Effective Turn-Taking

Building on the foundational concepts of turn-taking discussed earlier in Section 2, this section explores the specific cues that shape conversational dynamics. Understanding turn-taking requires analysing the overall patterns of speaker transitions and considering the signals that guide them in real time. In natural dialogue, interlocutors use a range of signals, including gaze direction, gestures, prosody, and linguistic features (syntactic, semantic, and pragmatic elements), to coordinate transitions [68]. These cues provide critical information that helps listeners anticipate when to remain passive and when to respond appropriately. Although turn-final cues offer a systematic way to identify Transition-Relevant Places (TRPs), they do not capture the full complexity of turn-taking, which involves predictive mechanisms. This predictive aspect is vital in spoken dialogue systems (SDS), where agents must decide when to speak and how to generate appropriate cues for smooth interaction. Computational modelling of these cues is challenging, as it is difficult to distinguish between cues that correlate with turn transitions and those that actively trigger responses. Consequently, researchers advocate supplementing corpus-based analyses with controlled experiments to identify causal relationships better. Even cues that human listeners do not consciously notice can be valuable for SDS design, underscoring the importance of understanding and generating effective turn-taking signals. In what follows, we begin

with verbal/textual signals, then move to fillers and acoustic patterns, discuss prosody, and finally turn to visual modalities, reflecting a progression from more abstract, high-level cues down to the most elemental, frame-wise information.

3.1. Text and Verbal Cues: Syntax, Semantics, and Pragmatics

Understanding verbal cues in spoken dialogue systems (SDS) requires a focus on the lexical content, what is actually said, rather than how it is delivered. This process begins by converting the speech signal into text using methods like automatic speech recognition (ASR) or manual transcription. Once the speech is transcribed, linguistic analysis can concentrate on important features such as syntax, semantics, and pragmatics, which are essential for managing turn-taking and response timing in conversations.

3.1.1. Syntactic & Pragmatic Completeness

A central idea in linguistic turn-taking is the concept of completeness, which helps determine when a speaker's turn is ready for transition. [17] differentiate between syntactic completeness, whether an utterance forms a well-structured linguistic unit, and pragmatic completeness, which assesses whether an utterance functions as a full conversational action in context. Syntactic completeness emerges gradually during speech: an utterance is treated as complete when, in its discourse environment, it can be understood as a full clause whose predicate is either explicitly stated or straightforwardly inferred [17] (p. 143). This interpretation also permits short acknowledgements and feedback tokens (e.g., "mm-hm") to count as complete when they perform an interactionally meaningful role. However, while syntactic completeness is necessary to mark a Transition-Relevance Place (TRP), it is not enough on its own to trigger a turn shift. Consider the following example, where syntactic completeness is marked by "/":

```
A: I booked/the tickets/
B: Oh/for/which event/
A: The concert/
```

In this exchange, "I booked" is not syntactically complete because it lacks a predicate, whereas "the concert" is complete because it conveys a full response. Pragmatic completeness, however, depends on whether the utterance accomplishes the relevant conversational action. For instance, the phrase "for which event" may be syntactically complete yet remains pragmatically incomplete if further elaboration is expected.

3.1.2. Projectability and Turn Coordination

Syntactic and pragmatic completeness contribute to projectability, which refers to how listeners anticipate when an utterance will be completed. Ref. [8] argue that turn-taking relies on predictive mechanisms, where listeners estimate the completion of a turn rather than waiting for explicit signals at the end. This predictive ability explains how speakers often begin to respond within ms (200 ms) of a turn ending, even before the previous turn is fully completed [10]. Projectability also plays a role in collaborative speech behaviours, such as choral speech and sentence completions, where listeners actively participate in constructing an utterance [42]. For example, if a speaker says, "I would like to order a . . . ", the listener can predict that a noun (e.g., "coffee" or "sandwich") will follow, allowing them to delay their response appropriately. Similarly, in question–answer exchanges, the syntactic structure of a question provides clues about when and how the listener should respond, making syntax a key factor in turn regulation.

Despite the significance of syntactic and pragmatic cues, incorporating them into spoken dialogue systems (SDS) remains a formidable challenge. Early turn-taking models often relied on fixed syntactic heuristics, such as examining the final two syntactic category

labels (i.e., part-of-speech tags, which classify words into grammatical categories such as nouns, verbs, and adjectives) [48,69] to determine whether an utterance is complete. For example, an utterance ending in a noun is more likely to be deemed complete, whereas one that terminates with a conjunction or determiner is less so. However, such methods fall short of capturing the inherent unpredictability of spontaneous speech, which frequently includes hesitations, self-repairs, and unfinished phrases. Recent advances in deep learning have enabled dynamic processing of linguistic structures, significantly enhancing turn-taking prediction. For instance, ref. [70] employed long short-term memory (LSTM) networks to analyse word sequences, syntactic category patterns, and phonetic features, demonstrating that lexical cues are critical for improving turn prediction accuracy. More recently, transformer-based models have further shifted the paradigm in turn-taking research. Ref. [71] introduced a transformer-based architecture that outperformed LSTM models in detecting Transition-Relevant Places (TRPs) by leveraging self-attention to capture long-range dependencies within dialogue contexts. Moreover, contemporary studies have integrated contextual embeddings from large-scale language models like BERT and GPT-3 to further refine turn-taking predictions [72,73]. These context-aware models dynamically interpret linguistic cues relative to preceding dialogue, making them more effective than traditional rule-based methods.

3.2. Fillers or Filled Pauses

Filled pauses, commonly known as fillers like "uh" and "um," are prevalent in spontaneous speech and are typically linked to moments of hesitation or uncertainty in the speaker [74]. Research indicates that the use of these fillers may correlate with increased cognitive load, suggesting that speakers employ them more frequently when processing complex information [75]. From a turn-taking perspective, fillers serve as crucial cues for listeners, signalling that the speaker intends to maintain their turn and has not yet completed their thought [76,77]. This turn-holding function is essential for the smooth flow of conversation, as it helps prevent interruptions and ensures that speakers can convey their messages effectively. The intentionality behind the production of fillers has been a subject of debate. Ref. [77] propose that speakers use "uh" and "um" deliberately to indicate minor or significant delays in speech, respectively. Their analysis suggests that "uh" signals a brief pause, while "um" denotes a more prolonged hesitation. However, ref. [78] challenges this view, presenting acoustic data that show no significant difference in the duration of pauses following "uh" versus "um," implying that these fillers may not have distinct meanings. Further studies have explored the linguistic status of filled pauses. Ref. [79] conducted experiments demonstrating that filled pauses can form part of more extensive linguistic representations, affecting sentence acceptability judgments and recall accuracy. This finding supports the notion that fillers function as integral components of language rather than mere disfluencies. Additionally, the prosodic features of fillers, such as intonation and duration, contribute to their communicative function. Research indicates that variations in pitch and length of filled pauses can convey different levels of speaker uncertainty or serve to manage the flow of conversation [80]. Understanding these nuances is crucial for developing spoken dialogue systems that can interpret and generate natural-sounding speech.

3.3. Speech and the Acoustic Domain

In spoken dialogue systems, the acoustic domain is fundamental, as it encompasses all the information conveyed by speech signals. Speech originates from a speaker and propagates as sound waves, pressure fluctuations in the air, captured by microphones or processed by the auditory system in human ears. These pressure variations are recorded as

intensity values, sampled at specific rates (Hertz), and can be visualised as waveforms (as depicted in Figure 6).

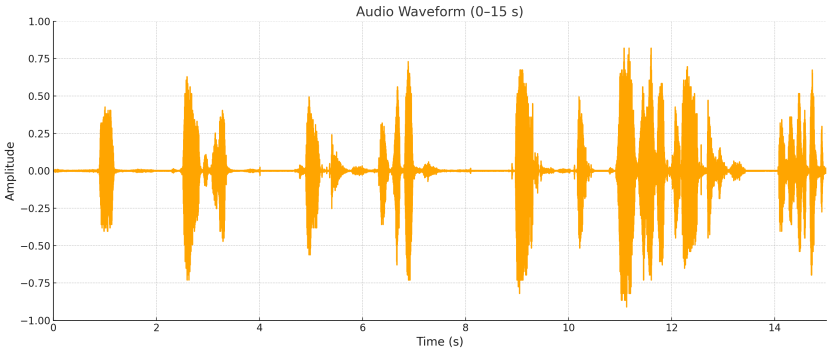
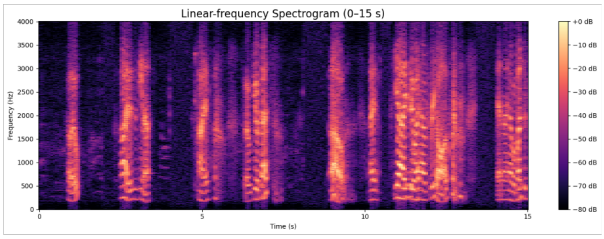


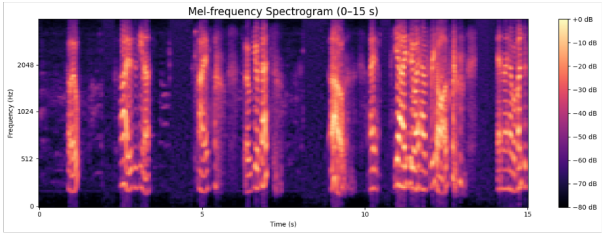
Figure 6. Waveform segment from Section 2 and Figure 2, showing the initial 15 s of the audio.

Spectrogram and Mel-Spectrogram

A spectrogram provides a two-dimensional visualisation of an audio signal, showing how energy is distributed across frequency (y-axis) over time (x-axis), as illustrated in Figure 7. It serves as a compact representation of both temporal and spectral information, making it a valuable basis for analysing acoustic cues relevant to turn-taking, such as pitch movement, intensity variation, and pauses. In practice, a spectrogram is computed by analysing short segments of the waveform to reveal frequency content over time, while a Mel-spectrogram further adjusts the frequency scale to approximate human auditory perception, emphasising the lower frequencies to which humans are most sensitive [81]. In this review, spectrograms are referenced only to contextualise acoustic feature representations used in turn-taking models, rather than to detail the underlying signal-processing procedures.



(a) A spectrogram on a linear scale, giving equal weight to every frequency band



(b) A Mel-scale spectrogram that emphasises the lower-frequency ranges.

Figure 7. Spectrogram visualisations of the speech segment shown in Section 2, Figure 2, illustrating the first 15 s of the audio. The x-axis denotes time, and the y-axis denotes frequency.

3.4. Prosody

The management of conversational dynamics and turn-taking is significantly impacted by prosody, which encompasses non-verbal aspects of speech such as intonation, loudness, and duration [33]. It conveys prominence, resolves syntactic ambiguities, expresses attitudes and emotions, signals uncertainty, and facilitates topic shifts. Research indicates that a stable level of intonation, commonly found in the midrange of the speaker's fundamental

frequency, serves as a signal to maintain the speaking turn in both English and Japanese. At the same time, pitch variations, either rising or falling, frequently suggest the intention of relinquishing the turn [48,82]. Furthermore, research indicates that speakers typically diminish their vocal intensity as they approach turn boundaries, whereas pauses within a turn typically display higher intensity [48,83]. Ref. [84] determined that prolonged pronunciation of the final syllable or the stressed syllable in a concluding clause functions as a turn-yielding signal in English, a conclusion supported by [82].

Additional insights into prosodic behaviour have arisen from studies comparing turn-taking across various languages and contexts. In Swedish, falling pitch consistently indicates turn-yielding, whereas rising pitch does not reliably associate with either maintaining or relinquishing the turn [85,86]. Moreover, analyses of American English dialogues have demonstrated that factors like final lengthening are not limited to turn endings; in fact, this lengthening is occasionally more pronounced within ongoing turns [48].

3.4.1. Intonation and Turn-Taking

Intonation is one of the most thoroughly examined prosodic elements in turn-taking. The pitch contour of an utterance, or the fundamental frequency (F0), functions as a crucial indicator of a speaker's intention to maintain or cede the conversational turn. Investigations in various languages, such as English [48,82,84], German [87], Japanese [83], and Swedish [85], indicate that sustaining a level intonation characterised by a stable pitch within the speaker's natural range serves as a cue for turn-holding. Conversely, falling or rising intonation is typically linked to turn-yielding, indicating that the speaker is prepared to relinquish the conversational floor [48,82,83]. Language-specific distinctions are present: in Swedish, a falling pitch signals the conclusion of a turn, whereas a rising pitch does not consistently correlate with either turn-holding or turn-shifting behaviour [85,86], suggesting that intonational patterns may not function uniformly across linguistic contexts. At the physiological level, F0 corresponds to the rate of vocal cord vibration and influences the perceived pitch of speech. This frequency corresponds to the lowest formant resonance and serves as the foundation for harmonic structures in speech production. Notably, male speakers typically exhibit a lower fundamental frequency (F0) than females, resulting in a lower overall pitch range. Despite the complexity of pitch extraction techniques, computational methods such as the Praat software toolkit [88] provide a standardised approach for analysing and extracting pitch contours, facilitating the integration of intonation-based features into spoken dialogue systems.

3.4.2. Intensity and Turn Regulation

Speech intensity, often called loudness or amplitude, is a key prosodic feature that plays a significant role in managing conversations. It reflects the acoustic energy present in speech and can be measured from spectrograms or raw waveform signals. This measurement corresponds to the waveform's amplitude values or the combined frequency magnitudes in each frame of the spectrogram. In simpler terms, speech at a lower volume results in lower intensity, while higher volume produces greater intensity. Research by [48] has demonstrated that speakers typically lower their vocal intensity when approaching a turn boundary, making it a strong prosodic cue for signalling turn transitions. Conversely, within-turn pauses often exhibit sustained or even increased intensity, reinforcing the speaker's intent to maintain control of the conversational floor. These findings highlight the critical role of intensity variations in shaping TRPs and influencing listener expectations regarding speaker transitions.

Similar patterns have been observed in Japanese conversational studies, where diminishing speech energy is strongly associated with turn shifts, while stable intensity

levels are indicative of turn-holding behaviour [83]. These cross-linguistic findings underscore the universal role of intensity fluctuations in regulating conversational dynamics, making it an essential feature in spoken dialogue systems (SDS). Given that intensity can be continuously and reliably extracted in real time, it has the potential to enhance SDS's predictive capabilities for identifying turn boundaries and managing speaker transitions more effectively.

3.4.3. Speech Rate and Duration as Turn-Taking Cues

The rate of speech and segment duration also serve as critical factors in managing turn coordination, influencing whether a speaker intends to maintain or relinquish the conversational floor. Research suggests that modifications in speech duration, such as extending the final syllable or stressing the terminal clause, can function as turn-yielding cues in English, a phenomenon first identified by [84] and later corroborated by [82]. However, ref. [48] offered a more complex perspective, demonstrating that final lengthening is not exclusively a turn-final phenomenon but occurs across all phrase boundaries. Notably, they found that lengthening is more pronounced in turn-medial inter-pausal units (IPUs) than in turn-final ones, suggesting that duration-based cues may not operate uniformly across all contexts. The role of speech rate and duration in turn-taking is further complicated by cross-linguistic variability. In Japanese task-oriented dialogue, ref. [83] observed that extended duration correlates with turn-holding rather than turn-yielding, contradicting earlier findings in English. Similarly, ref. [48] found that in Swedish, final lengthening does not reliably predict turn transitions, highlighting the need for a language-specific approach to modelling duration-based turn cues.

In spoken dialogue systems (SDS), accurately estimating the rate of speech and the duration of phonetic units, such as phonemes, syllables, or words, is essential to determine when to complete turns in conversation. SDS architectures typically break incoming audio into frames (e.g., 20 ms windows at 50 Hz) and identify phonetic boundaries. However, these boundaries do not always align perfectly with the acoustic minima, so robust feature extraction requires complex segmentation. Unlike human listeners, who can easily adapt to different speaking styles, SDS must normalise raw duration measures against individual speaker tempo averages to facilitate comparisons across various users. This normalisation reduces inter-speaker variation (e.g., some speakers may naturally articulate five syllables per second, while others might produce four). However, this step introduces additional computational overhead and latency. Real-time systems must therefore strike a balance between the window length required for reliable rate estimation and the need to respond quickly to cues indicating the end of a turn. Additionally, duration-based signals can exhibit language-specific patterns: English speakers often use final-syllable lengthening as a cue to yield their turn, while in Japanese, extended duration typically indicates a desire to hold the turn. In Swedish, there is minimal correlation between final lengthening and turn transitions. These variations highlight the importance of developing adaptive SDS models that learn speaker- and language-specific duration norms. Such models would ensure that speech rate and duration features meaningfully contribute to predicting turn-taking across contexts.

3.4.4. Prosody in Computational Models of Turn-Taking

Integrating prosody into spoken dialogue systems (SDS) is essential for attaining natural turn coordination. In contrast to humans, SDS face constraints in forecasting speaker transitions in real time, mainly due to processing delays and inaccuracies in automatic speech recognition (ASR) [7]. Given that prosodic features can be continuously

extracted with high reliability, they may function as significant inputs for enhancing turn prediction in conversational agents.

Numerous computational models have analysed the predictive efficacy of prosodic features in relation to verbal cues. Ref. [83] employed a decision tree to examine syntax–prosody interactions in Japanese task-oriented dialogue, revealing that although individual prosodic features exhibited inferior predictive capability compared to syntactic cues, the exclusion of all prosodic features markedly diminished model efficacy. Ref. [48] employed multiple logistic regression on American English dialogue and determined that textual completion was the most significant predictor of turn shifts, followed by voice quality, speaking rate, intensity, pitch level, and IPU duration. Nevertheless, upon evaluating all features, intonation did not significantly improve overall model performance, suggesting that prosody alone may be insufficient for predicting turn shifts. A visualisation of the Mel-spectrogram, along with prosodic features such as intonation and intensity, is shown in Figure 8. This figure illustrates the Mel-scale spectrogram of the first 15 s of speech, with the speaker's F_0 contour shown in red and the RMS intensity contour in green. Vertical dashed lines indicate the boundaries of individual utterance segments, clearly marking when the speaker is active. This combined view illustrates how prosodic dynamics, changes in pitch and loudness, correspond to spectral energy peaks, underscoring their significance as cues for real-time turn prediction.

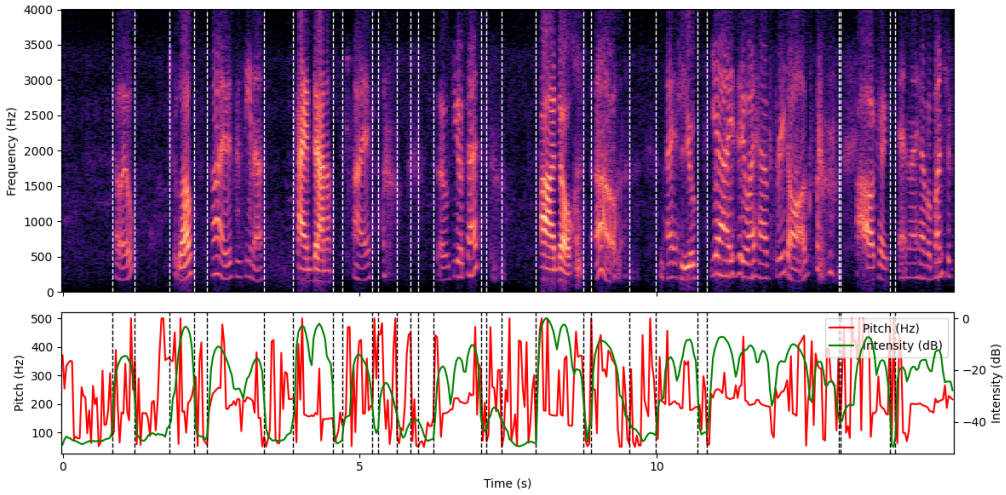


Figure 8. Prosodic contours for the audio in Section 2, Figure 2, with F_0 in red and intensity in green.

3.5. Visual Cues

Non-verbal cues, such as gaze and gestures, are essential for regulating conversation turn-taking, complementing verbal signals to ensure smooth transitions between speakers. In human-human interaction, gaze direction and gesture patterns offer valuable insight into the timing and structure of conversational exchanges, serving as turn-holding and turn-yielding mechanisms. Integrating these non-verbal cues into spoken dialogue systems (SDS) and human–robot interaction (HRI) enhances the naturalness and fluidity of machine-generated communication, making them more responsive to human-like conversational behaviour.

3.5.1. Gaze

Gaze serves as both a communicative and a regulatory mechanism in turn-taking. Early studies by [68,89] demonstrated that speakers tend to avert their gaze at the beginning of a turn and redirect it toward the listener as they near a Transition-Relevance Place (TRP), signalling an imminent speaker change. Similarly, listeners are more likely to maintain eye

contact while the speaker holds the floor, shifting their gaze at the point of transition to indicate readiness to take over [90,91]. Recent research has confirmed that gaze patterns vary across different types of turn-taking scenarios. When speakers yield the floor, they are more likely to direct their gaze toward the listener, reinforcing their intention to hand off the turn. Conversely, in turn-holding situations, speakers tend to sustain gaze aversion, indicating that they intend to continue speaking [54]. This distinction becomes even more critical in multi-party conversations, where gaze helps allocate turns and resolve floor competition. For example, ref. [92] found that participants who successfully claimed the floor achieved mutual gaze with the outgoing speaker before initiating speech, whereas those who failed to secure a turn exhibited misaligned gaze behaviour. Computational models of gaze behaviour have been integrated into embodied conversational agents (ECAs) to enhance the realism of system interactions. Early rule-based models relied on predefined gaze shifts at turn beginnings and completions, while more advanced machine learning approaches dynamically predict gaze behaviour based on contextual cues [93]. Transformer-based models now incorporate gaze alongside verbal and prosodic features, refining turn-taking predictions and improving the responsiveness of conversational agents [94].

3.5.2. Gesture

Gestures, particularly hand movements, act as another crucial modality for regulating conversational flow. Research indicates that speakers use gestures to reinforce, clarify, or anticipate speech content, offering an additional channel of communication beyond verbal language. Ref. [84] observed that specific hand movements, such as outward-directed gestures or static hand positioning, serve as turn-holding cues, discouraging interruptions by signalling the speaker's intent to continue. In contrast, gesture cessation before speech completion can serve as a turn-yielding signal, facilitating smoother transitions [95].

Hand Gestures and Predictive Turn-taking

Beyond turn-holding and yielding functions, hand gestures provide predictive information about upcoming speech content, allowing interlocutors to anticipate conversational structure [96]. Studies have found that representational gestures, which depict semantic content related to speech, frequently precede their verbal counterparts, giving listeners an early indication of meaning before it is explicitly stated. Ref. [97] analysed gesture timing relative to speech and found that gesture strokes the most meaningful segment of a gesture, typically occurring 200–600 ms before their corresponding lexical affiliate, reinforcing their predictive role in the conversation.

Gesture Timing and Turn Coordination

The synchronisation of gesture timing with speech is strategically aligned with Transition-Relevance Places (TRPs). When speakers intend to relinquish the floor, gestures tend to terminate before the end of the speech, serving as a preliminary cue for the next speaker to take over. Conversely, when speakers intend to retain the floor, gestures may extend beyond speech, reinforcing continued engagement [95]. Gesture-speech asynchrony has also been linked to response latency in dialogue. Questions with gestures often receive faster responses, suggesting that listeners utilise gesture-based information to anticipate upcoming turns. These findings highlight the role of gestures in processing conversational structure, reducing gaps and overlaps by providing additional cues for turn coordination [97].

The multimodal nature of human communication underscores the need to integrate gaze and gesture cues into dialogue systems. While gaze direction helps regulate turn allocation and speaker transitions, gestures contribute to semantic reinforcement and predictive timing. The combination of these visual and non-verbal signals enhances turn-

taking accuracy in human–robot interaction (HRI), making artificial conversational agents more natural and intuitive.

4. Computational Models for Natural Turn-Taking in Human–Robot Interaction

To achieve natural interactions between humans and robots, dialogue systems must effectively mimic human conversational behaviours, particularly in managing timing and coordination during turn-taking. This section discusses key components for realising such natural interactions, including the significance of predictive models, multimodal integration, and strategies for handling overlaps and interruptions. Highlighting recent advancements and ongoing challenges, the following subsections outline approaches and considerations that can significantly enhance the responsiveness, efficiency, and overall human-like quality of interactive dialogue systems. In the remainder of this review, we focus on recent computational approaches to turn-taking in spoken dialogue systems and human–robot interaction. Unless otherwise noted, we include work that (i) was published between late 2021 to the time of publication, and (ii) proposes or evaluates models that function in spoken interactions for tasks such as detecting or predicting the end of a turn, timing and selecting types of backchannels, managing real-time system behaviour, or overseeing multi-party conversations. Earlier models will only be referenced when they serve as direct precursors to these recent neural, self-supervised, or multimodal approaches.

4.1. Turn-Taking Detection and Prediction Models

The arguably most studied aspect of turn-taking in conversational systems is determining when the user's turn ends, and the system can begin speaking (i.e., the detection of TRPs). A related aspect is determining when the system should provide a backchannel (i.e., as discussed in an earlier section). This section will examine recent efforts to develop models of turn-taking, highlighting the methodologies, key findings, and datasets used in these studies. Since a review article has already addressed earlier work in this area, this review will focus on more recent developments. Significant progress has been made in the field in recent years, particularly with the introduction of large language models. This serves as the motivation for this review.

There are three broad categories of turn-taking models commonly discussed in the literature. The first, and simplest, is the silence-based approach, in which a Transition-Relevance Place (TRP) is assumed once a pause exceeds a predetermined duration. This method is frequently adopted as a baseline for comparing more advanced models. A second family of approaches relies on Inter-Pausal Units (IPUs), typically segmented using an external Voice Activity Detection (VAD) system. Here, the task is to determine whether the end of an IPU corresponds to a TRP or whether the current speaker is likely to continue. IPU-driven methods have been widely studied across different conversational settings and model architectures [56,69,72,98–102]. These systems generally extract a feature set—such as lexical, syntactic, prosodic, and contextual cues—from the full IPU or its final portion, and then classify the boundary as a hold or a shift. Early work predominantly used decision-tree or random-forest-based classifiers [98,99], as well as SVMs [48,56,69,103], while more recent research has favoured modern machine-learning models [72,100–102,104,105].

The third line of work adopts a continuous, frame-level prediction paradigm in which conversational activity is estimated incrementally throughout the interaction. Rather than waiting for specific events, such as an IPU ending, to trigger a decision, these models continuously compute the probability of TRPs and other turn-taking behaviours. Incremental processing enables deeper, context-sensitive interpretation of the user's utterance and supports real-time decision-making on turn coordination. As a result, such systems

can anticipate upcoming completions, plan responses, identify opportunities for timely backchannels, or even initiate interruptions when appropriate capabilities that IPU-based models cannot offer.

The model variants are divided into two main concepts: end-of-turn detection and end-of-turn prediction models. The key difference between these two concepts is that end-of-turn detection (often referred to as "endpointing") refers to models that react to signals already available, such as silence or completed intonation patterns, to determine whether a speaker has finished speaking. These reactive approaches typically examine cues from the immediate past (e.g., the speaker stopping mid-sentence or pausing long enough to indicate a potential turn completion) before deciding to hand the floor over. In contrast, end-of-turn prediction models strive to anticipate when a speaker's turn is likely to end, often by continuously analysing ongoing speech and prosodic cues to project an imminent completion. Ref. [9] argued that both IPU-based models and silence-based models can be seen as reactive since they respond to past cues to make decisions at the present moment. Thus, these models fall under the category of end-of-turn detection. In contrast, continuous or incremental models are classified as end-of-turn prediction models.

In addition, previous studies have explored combinations of incremental and IPU-based approaches [102,106], as well as models that rely solely on verbal information to forecast turn-taking behaviour before an utterance is complete [107,108]. While these diverse approaches remain relevant, this review focuses on the breakthrough line of work that has shaped the modelling strategy adopted in this research. This emphasis does not imply that incremental processing is the only method currently in use, but rather that it has become one of the most influential and widely applied frameworks in contemporary turn-taking research.

The main contributions of this review trace the development of computational models of turn-taking from earlier, single-cue statistical techniques toward more sophisticated systems that integrate deep learning, self-supervised representations, and real-time incremental processing. In the following section, we outline specific advancements in predictive language mobility to capture and anticipate when a speaker is likely to complete their turn.

4.1.1. End-of-Turn Detection

In human-like spoken dialogue systems, end-of-turn detection, which determines whether a target speaker's utterance has ended, is an essential technology [8,103,109]. It is widely known that heuristic end-of-turn detection based on non-speech duration, as determined by speech activity detection (SAD), is insufficient for smooth turn-taking because speakers often irregularly insert variable-duration pauses into their utterances [110]. So far, various modelling methods have been examined for building speech-based end-of-turn detection [32,48,98,108,111]. The traditional approach is discriminative modelling using the statistics of frame-level acoustic features (such as maximum, minimum, and average values) extracted from the target speaker's current utterance. Earlier seminal works, such as [8], for modelling end-of-turn detection established turn-taking as a systematic yet flexible process. They proposed that conversational turns are governed by Transition-Relevance Places (TRPs), where speakers change seamlessly through current speaker selection or self-selection mechanisms. The paper's central argument was that turn-taking is context-free and context-sensitive, involving a locally managed yet recurrent system of speaker allocation. Although not computational in nature, later empirical studies have repeatedly validated the descriptive adequacy of this framework through observations of tightly timed turn transitions, reinforcing the relevance of TRPs as a basis for modelling conversational timing.

In addition, some recent studies have refined these concepts by addressing their limitations, particularly the challenge of reliably annotating and modelling the dynamics of interruptions, overlaps, and turn completions. For instance, ref. [112] laid the groundwork by focusing on annotating Transition-Relevance Places (TRPs) and interruptions in conversational corpora. It emphasised leveraging prosodic and linguistic cues [72,113] to differentiate between terminal and non-terminal utterances, providing a structured framework for analysing turn-taking behaviour. However, the binary classification approach, categorising events as either smooth transitions or interruptions, limited its ability to capture the nuanced dynamics of overlaps, cooperative interruptions, or turn continuations. Their findings highlighted the need for more flexible models that go beyond static classifications to incorporate contextual variability. While effective for offline analyses, these models are limited in capturing the anticipatory nature of real-time interactions. Importantly, their results also demonstrated that TRPs can be annotated with substantial reliability and that structured annotation exposes systematic patterns, such as smooth transitions occurring even within overlaps, offering valuable empirical grounding for evaluating computational systems. Ref. [73] explores linguistic feature-based token-level segmentation using GPT-2. This approach bypasses traditional acoustic-only methods and emphasises real-time adaptability to live conversations. Such linguistic models address gaps in earlier reactive systems by advancing toward predictive turn-taking, where the focus shifts from detecting TRPs post-facto to anticipating them dynamically. Their findings show that transformer-based models can capture fine-grained linguistic regularities relevant for turn boundaries, achieving stronger performance than recurrent architectures. However, the approach remains dependent on accurate ASR output and does not incorporate prosodic cues, which may limit generalisability in multimodal, spontaneous dialogue scenarios.

Subsequent research by [114] expanded on previous foundational works by examining the temporal dynamics of turn-taking in conversations. They argued that turn duration alone is insufficient for detecting transitions and highlighted the exponential distribution of turn lengths as evidence that temporal markers lack predictive utility. This insight emphasises the need to integrate semantic and prosodic signals into predictive frameworks, a limitation that models such as the Timing Generating Network (TGN) address. The study's empirical contribution is significant, demonstrating that duration information provides virtually no advantage for anticipating turn shifts, thereby challenging longstanding assumptions in traditional systems. Ref. [115] introduced the Timing Generating Network (TGN), a neural network-based model that dynamically predicts the optimal timing for turn initiation in conversation. Unlike traditional systems that rely on end-of-turn detection, the TGN continuously evaluates conversational context to determine when a transition is appropriate. The model also incorporates response obligation recognition as an additional task, ensuring that timing decisions are sensitive to the context. Experimental results demonstrated superior performance in precision and recall compared to traditional pause-based models. However, the reliance on controlled experimental datasets raises concerns about their applicability in less structured, naturalistic settings. Despite this, the model's achievements illustrate the benefits of jointly modelling silence likelihood, pragmatic constraints, and contextual cues, showing clear improvements over rigid, threshold-based timing systems.

Another critical aspect of turn-taking lies in the role of prosodic and contextual cues. Ref. [116] examined the relationship between speech rate, prosody, and turn transitions, highlighting that rapid response times of under 200 ms are based on anticipatory mechanisms rather than reactive processes. This aligns with the findings of [117], which introduced a probabilistic model of Transition-Relevance Place (TRP) durations, emphasising the difference between speaker switches and continuations. The researchers discovered

that shorter TRPs are more likely to trigger speaker switches, supporting the predictions of [8]'s turn-taking model. By introducing a turn-taking propensity function, the study provided practical guidelines for optimising response timing in conversational agents. While the model effectively reduces mistimed responses, it primarily focuses on silence-based transitions, limiting its applicability in scenarios with overlapping speech. Importantly, these studies collectively demonstrate that prosodic cues, particularly final lengthening and pitch movements, enable anticipatory timing, providing strong empirical justification for integrating prosodic prediction into computational turn-taking systems.

4.1.2. End-of-Turn Prediction

One key difference between human turn-taking and spoken dialogue systems (SDSs) is that humans do not simply react to cues indicating that the interlocutor is ready to yield the floor. If humans were to wait for such cues before formulating their responses, psycholinguistic research suggests that response times would range from approximately 600 to 1500 ms [10], which is significantly slower than the response times typically observed in conversation. This discrepancy suggests that humans can predict turn completions in advance, even before the current turn is completed [8,10,11]. A foundational step in modelling responsive turn-taking is to implement an incremental dialogue architecture. This approach continuously processes the user's speech, enabling more timely and fluid decision-making [118].

Ref. [118] presented an early account of a fully incremental dialogue system, although it was constrained to number dictation. While the user was reading the numbers, the system could begin preparing responses and providing very rapid feedback by continuously processing the user's speech (including prosodic turn-taking cues). Empirically, this architecture achieved response latencies of only a few hundred ms at suitable feedback points, substantially faster than typical silence-threshold baselines, demonstrating that fully incremental processing can make system timing perceptibly more human-like [118]. A critical challenge of incremental systems, as identified by [118], is the need for revision, as tentative hypotheses about what the user is saying may change as more speech is processed. Consider a scenario in which the user begins to say "eigh…" but adds additional syllables, ultimately forming "eighteen." If the system prematurely interprets "eigh" as "eight," downstream processes may begin generating output for the number eight. Once the correction becomes apparent, these processes must backtrack, revise their assumptions, and recalculate based on "eighteen" instead. If the spoken output has already started, the system might need to self-correct mid-sentence, much like human speakers do, thereby introducing the complexities of on-the-fly self-repairs, as [119] explored. Similarly, ref. [120] introduced an incremental model designed to forecast the upcoming prosodic contours of dialogue partners. Although the work did not directly evaluate turn-taking decisions, its findings demonstrated that short-horizon predictions of features such as pitch and speaking fraction are feasible and can track broad future prosodic tendencies. However, their results also indicated that prosody alone yields only modest improvements over strong baselines, suggesting that while incremental prosodic modelling is promising, richer contextual signals or multimodal cues are required before these predictions can yield substantial turn-taking gains. Building on this line of research, ref. [70] proposed a continuous, frame-level model for forecasting future voice-activity patterns and explicitly assessed its impact on several turn-taking tasks. Their evaluation showed that the model not only surpassed traditional IPU-based and hand-crafted feature approaches but also exceeded human accuracy in identifying whether post-IPU silences would lead to a shift or a hold. The model also distinguished reliably between brief backchannels and longer contributions, indicating sensitivity to conversational structure that earlier classifiers lacked.

This modelling framework subsequently shaped later work [121–123], which refined the architecture, incorporated stronger linguistic or prosodic representations, and tested the approach across different languages. These studies consistently reported that incremental prediction, whether of prosodic trajectories, voice activity, or linguistic continuation, provides more stable, context-aware estimates of turn-taking than event-based methods. The collective evidence from this research thread underscores a clear trend: models that compute turn-taking cues continuously and probabilistically outperform those that rely on discrete IPU boundaries or silence thresholds, especially in conversational settings with overlaps, rapid transitions, or nuanced backchannel behaviour.

A significant thread running through recent work is the role of context and pragmatic skills in turn-taking. These skills include understanding indirect speech, appreciating humour, and negotiating. Ref. [124] examined the interaction of task performance and pragmatic skills in a Wizard-of-Oz experiment with smart speakers. They demonstrated that pragmatic competence significantly impacts user behaviour, with high-pragmatic systems eliciting more topic development and reducing user disengagement. Their findings underscore the need for systems to possess metalinguistic capabilities, such as processing indirect requests, utilising conversational context, and appreciating subtle verbal cues, such as humour. In practice, users interacting with pragmatically richer systems produced longer, more coherent contributions and showed higher willingness to continue the interaction, indicating that pragmatic skill directly shapes conversational flow. By addressing the interplay between pragmatic skills and user behaviours, the study contributes to the design of conversational agents capable of dynamic, context-aware turn-taking. These findings align with efforts to model nuanced conversational dynamics, such as those captured in the Fine-Grained Turn-taking Action Dataset (FTAD) by [125]. The FTAD provided structural representations of intricate human-to-human interactions, revealing patterns like backchannels, overlaps, and interruptions. Models trained on FTAD showed that even strong neural baselines struggle to fully capture subtle turn-taking actions, highlighting the complexity of pragmatic and contextual cues in natural conversations. Together, these studies emphasise the need for systems to move beyond surface-level turn-taking cues to incorporate deeper contextual and pragmatic understanding. Ref. [126] tackled limitations in rule-based systems, such as repetitive responses, misinterpreted pauses, and overlapping speech. The proposed framework effectively enhances the fluidity and coherence of interactions. Their evaluations showed marked improvements in perceived naturalness and conversational flow compared to traditional rule-based systems, demonstrating that adaptive timing and context-aware response generation lead to smoother, more human-like exchanges.

A deep neural network (DNN)-based turn-taking detector is central to the system, enabling it to accurately identify speaker transitions and dynamically adapt responses, whether through backchanneling, topic changes, or regular conversational turns. This adaptive approach minimised conversational interruptions and enhanced user impressions, bridging the gap between traditional systems and fluid, human-like interactions. However, challenges such as optimising speech overlap avoidance remain a priority for future iterations. Similarly, ref. [127] advanced the field by employing language models to project future turn completions. Their system achieved lower response latencies by predicting upcoming dialogue based on conversational context while maintaining coherence. However, limitations related to domain adaptation and reliance on textual data without prosodic features indicate areas for further development. These innovations highlight the shift from reactive to proactive turn-taking systems capable of anticipating conversational needs. A recurring theme across these studies is the balance between real-time responsiveness and conversational accuracy. While incremental systems benefit from rapid

adaptations, challenges such as overlapping speech, task-specific tuning, and domain generalizability persist.

The challenge of detecting and responding to subtle conversational cues extends to specialised applications. Ref. [128] leveraged turn-taking dynamics for the automatic detection of Reactive Attachment Disorder (RAD) in children, illustrating the diagnostic potential of conversational analysis in clinical settings. By identifying turn-taking patterns indicative of RAD, this work highlights how conversational behaviours can serve as markers for psychological conditions. Theoretical frameworks, such as [129], model turn-taking as a stochastic process, offering insights into the influence of individual speaking tendencies and conversational memory on team dynamics.

The shift to deep learning brought models like recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, and transformers, such as GPT-2, into the realm of turn-taking prediction [101]. Ref. [104] integrated turn-taking prediction into an end-to-end (E2E) automatic speech recognition (ASR) system, addressing the complexities of conversational disfluencies, including hesitations and filled pauses. By embedding a turn-taking detector directly into the E2E model. Their approach was based on an end-to-end Recurrent Neural Network Transducer (RNN-T) architecture, which integrates acoustic features from the encoder and linguistic information from the prediction network. They achieved remarkable precision and recall rates with minimal latency, highlighting the feasibility of combining acoustic and linguistic cues for seamless interaction. In collaborative and robotic domains, ref. [130] demonstrated the potential of spiking neural networks (SNNs) for early turn-taking prediction in human–robot collaborative assembly. SNNs are a class of biologically inspired models that transmit information via discrete "spikes" (events) rather than continuous activations, allowing them to naturally capture fine-grained temporal dynamics and operate with low latency and power consumption. Their approach accounted for human uncertainty and dynamic motion primitives, achieving smoother task transitions; however, its task-specific design restricts its application in more generalised HRC environments. Ref. [131] complements this by using a learning-from-demonstration (LfD) framework to teach robots turn-taking behaviours informed by verbal, prosodic, and gestural cues, emphasising the necessity of context-specific adaptability.

More complex approaches now integrate Large Language Models (LLMs) with audio modelling, leveraging self-supervised techniques. One significant advancement is the Voice Activity Projection (VAP) model, a self-supervised approach that incrementally predicts future voice activity (VA) using raw waveforms. Unlike earlier rule-based models that relied on fixed silence thresholds, VAP learns turn-taking transitions (Shift, Hold, and Backchannel) by leveraging prosodic cues, including pitch, intensity, and duration. This approach enables spoken dialogue systems (SDS) to anticipate conversational shifts more naturally, leading to smoother interactions and reduced latency in speech-based interfaces [123]. The superior performance of the VAP model compared to earlier rule-based models stems from its ability to dynamically adapt and generalise across varied conversational contexts, moving beyond simplistic silence-based thresholds. VAP captures nuanced vocal patterns that signal conversational intent and speaker state changes by directly modelling temporal relationships within raw waveform inputs. This attribute enables it to recognise subtle prosodic patterns, such as pitch contour variations and intensity fluctuations, that earlier models typically overlooked. Consequently, the model improves turn-taking accuracy by providing timely predictions of Transition-Relevance Places, enhancing the responsiveness and naturalness of conversational interactions. To further strengthen turn-taking prediction, ref. [132] explored prosodic perturbations by manipulating pitch (flattened F0) and intensity to examine their effects on turn-taking decisions. Their findings indicate that, while lexical information remains dominant, prosodic variation significantly improves turn-taking pre-

dictions in ambiguous contexts. For instance, when syntactic information is insufficient, pitch dynamics and intensity shifts provide critical signals for distinguishing between Hold and Shift decisions. They also observed that VAP models are sensitive to phonetic cues, indicating that both intonation and intensity are essential factors in turn-taking. Additionally, the study demonstrated that altering the model's frame rate has no substantial impact on performance. This research enhances our understanding of how computational models can utilise prosodic information relevant to turn-taking and underlines the importance of phonetic details. However, their study is constrained by the use of artificially generated utterance pairs and the absence of model training on perturbed data, factors that may not fully mirror the nuanced prosodic patterns found in naturally occurring speech. Ref. [133] adopts a more application-oriented approach by introducing and demonstrating the VAP model as a versatile tool for various spoken dialogue systems (SDS) tasks. Instead of focusing on prosodic analysis, this research demonstrates the model's ability to predict future voice activity incrementally and its utility for managing turn-taking decisions in SDS. The authors emphasise VAP's effectiveness in classifying backchannels, resolving overlapping speech, and assessing the quality of turn-taking signals in text-to-speech (TTS) systems. They highlight that VAP can be used to train TTS systems to produce more human-like turn-taking signals, serving as an automated tool for linguistic research. Applications might include labelling conversational corpora, identifying shifts or interruptions in turns, and analysing the role of fillers in turn-taking. Furthermore, the model demonstrates sensitivity to prosodic cues comparable to that of human recognition. VAP's utility extends to different applications, including evaluating conversational Text-to-Speech (TTS) synthesis, where [134] introduced an automatic evaluation framework for turn-taking cues in TTS. Their results indicate that while commercial TTS systems can generate strong turn-holding signals, they struggle with turn-yielding cues, leading to user confusion and unintended barge-ins. Based on VAP-driven metrics, the proposed evaluation approach provides an automated method for assessing turn-taking behaviours in synthesised speech, paving the way for TTS systems capable of more human-like dialogue coordination.

In the domain of multilingual turn-taking, ref. [122] demonstrated that monolingual VAP models struggle to generalise across languages due to linguistic differences in turn-taking patterns. However, a multilingual VAP model, trained across English, Mandarin, and Japanese, achieves predictive performance comparable to that of monolingual models in their respective languages [135,136], highlighting its ability to generalise across linguistic variations. Furthermore, analyses of intonational shifts at turn-final positions confirm that prosody plays a more critical role in turn-taking decisions in tonal languages, such as Mandarin, whereas syntactic structures dominate conversations in English. Ref. [137] introduced a novel task, target conversation extraction (TCE), highlighting the challenge of isolating a target conversation from interfering speakers and background noise. This problem underscores the need for systems to discern relevant speech based on speaker embeddings and turn-taking dynamics. This is particularly important in noisy environments where machines struggle to isolate target speakers, a task that humans perform with relative ease. The TCE task leverages temporal patterns inherent in human conversations, particularly turn-taking dynamics, uniquely characterising speakers engaged in conversation. Their study demonstrates the feasibility of using neural networks to enhance the signal-to-noise ratio in conversation extraction, yielding significant improvements in both English and Mandarin datasets. However, this work assumes clean speech, speaker ID, and timestamp labels to extract speaker embeddings and avoid speaker repetition for training, which may not always be available in real-world scenarios. Some approaches have sought to improve turn-taking models through domain adaptation. For instance, ref. [138] investigated how spoken dialogue systems might achieve more natural timing

by estimating when an ongoing user utterance is likely to finish, rather than waiting to detect its final boundary. Their approach used wav2vec 2.0 fine-tuned for ASR to obtain linguistically enriched representations, and they compared several feature types, including hidden states, token-level logits, and symbol-sequence outputs, to determine which best supported end-of-utterance prediction. Their results showed a clear trend: models that incorporated linguistic information from wav2vec 2.0 consistently outperformed those that relied solely on acoustic features, such as MFCCs or CNN-based encoders. In particular, hidden-state and logit representations yielded noticeably more reliable estimates of the remaining time before an utterance ended, indicating that contextualised linguistic cues captured by self-supervised models play a substantial role in refining temporal predictions. However, the study also revealed that performance varied across feature types, with symbol-sequence representations performing markedly worse, suggesting that overly discrete representations may lose important prosodic or contextual detail. A practical advantage of this method is that it leverages the internal representations of wav2vec 2.0 without depending on external transcripts, which are often error-prone in spontaneous conversation. At the same time, the authors noted significant constraints: the model was evaluated only within a one-second prediction window, and performance degraded as predictions extended to more distant future time points. These findings indicate both the promise and the current limitations of utterance-final timing prediction using Self-Supervised Learning (SSL) models. While linguistic features significantly boost accuracy, the method still struggles with long-range forecasting and requires further refinement to match the variability of natural conversational timing.

Large Language Models (LLMs), despite their remarkable progress in language generation, exhibit significant limitations in turn-taking prediction. Studies have shown that these models often fail to anticipate Transition-Relevance Places (TRPs) in natural conversation, particularly within-turn TRPs, leading to long silences or inappropriate interruptions. This issue arises from the predominant training of LLMs on written text rather than spoken dialogue, which lacks real-time conversational dynamics [139]. For example, ref. [71]'s TurnGPT, based on a transformer architecture, leverages linguistic cues to predict turn-shift points, showing superior performance compared to traditional methods. However, these models, which are primarily text-based, often struggle to process non-verbal cues such as pauses and intonational shifts, which are crucial for smooth turn-taking dynamics. Addressing this gap, recent work combines neural acoustic models with LLMs. For instance, ref. [140] demonstrates that combining neural acoustic models with large language models (LLMs) improves predictive accuracy, as acoustic cues such as pauses and tonal shifts indicate Transition-Relevance Places (TRPs). They propose a fusion model that integrates LLMs (such as GPT-2 and RedPajama) with HuBERT, a neural acoustic encoder. These findings support the notion that combining linguistic and acoustic information is essential for accurately predicting turn-taking. The unique advantage of integrating linguistic and acoustic information lies in their complementary nature; linguistic cues capture syntactic and semantic signals that inform turn boundaries, while acoustic cues reveal subtle real-time signals such as prosody, stress patterns, rhythm, and pause duration. Acoustic cues, in particular, provide immediate temporal markers indicating when a speaker is likely to yield or retain the floor. By combining these modalities, the resulting model gains a more comprehensive understanding of conversation dynamics, thereby significantly reducing prediction errors, such as premature interruptions or overly prolonged silences. The Lla-VAP model proposed by [141], which integrates LLMs (Llama 3) with Voice Activity Projection (VAP), exemplifies this approach by explicitly modelling the dynamic interplay between verbal content and vocal delivery, enhancing the precision and timing of turn-taking predictions. Their model enhanced turn-taking predictions in both scripted

and unscripted conversations by fusing linguistic context modelling from LLMs with prosodic and temporal information from VAP. Evaluation on the In-Conversation Corpus (ICC) and Coached Conversational Preference Elicitation (CCPE) dataset reveals that this ensemble approach achieves higher accuracy in Transition-Relevance Place (TRP) detection, particularly in task-oriented dialogues. However, the authors also observed reduced performance when predicting TRPs in less-structured exchanges, indicating that model robustness still depends heavily on the conversational domain and data diversity. This highlights a recurring trend across current LLM-based systems: gains in one dialogue setting do not always generalise to others. Similarly, ref. [142] evaluated LLMs such as Gemini, OpenAI's GPT-3.5 & GPT-4, Claude2, and Llama 2, and demonstrated how LLM-powered end-of-turn prediction can enhance conversations between elderly users and companion robots. The combined analysis of LLM and LLM-VAD systems provided a comprehensive perspective on turn-ending prediction augmented by LLMs. The LLM-VAD system showed promise for more nuanced turn-taking management, particularly in scenarios involving elderly users or natural, open-ended conversations. The authors highlighted the variability in model performance across different datasets, emphasising the importance of context and dataset characteristics in model selection. Notably, their results showed that while LLMs improve over simple acoustic baselines, inconsistency across user groups signals that LLM-based turn-taking remains highly sensitive to speech style, tempo, and conversational goals, an essential limitation for real-world deployment. Ref. [139] addresses the challenge of timing turns appropriately in spoken dialogue by focusing on the ability of LLMs to predict within-turn TRPs. The authors collected a novel dataset of human responses to identify within-turn TRPs in natural conversations. By comparing the performance of state-of-the-art LLMs to that of human participants, the study revealed that current LLMs struggle with TRP prediction, suggesting that pre-training on written data is insufficient. This finding highlights a critical bottleneck in applying LLMs to spoken dialogue systems. Their work also suggested that future models may require more spoken dialogue input during pre-training or explicit fine-tuning of spoken dialogue data. A limitation of this study is that the models used had access only to linguistic information, whereas human participants had access to both prosodic and linguistic cues. The authors therefore conclude that multimodal grounding, especially prosody, is essential for closing the gap between machine and human turn-timing ability, reinforcing similar conclusions from [140,141].

Beyond computational applications, linguistic turn-taking behaviours have been examined in clinical settings, particularly in schizophrenia research. Ref. [143] adopted a novel approach by investigating turn-taking patterns in dyadic conversations involving individuals with schizophrenia and control participants, revealing distinct turn-taking behaviours characterised by higher levels of overlap and mutual silence in patient interactions. Notably, these findings indicate that mutual silence and conversational fragmentation correlate with the severity of negative symptoms. At the same time, no significant relationship was found with subjective self-experience anomalies, suggesting that interpersonal coordination deficits in schizophrenia manifest at the structural level of dialogue rather than as a function of conscious self-awareness. The authors employed a semi-automatic analysis method to quantify speech time, speech overlap, and silent pauses, relying on high-quality stereo recordings. While the study provides valuable insights into interactional coordination in schizophrenia, the authors acknowledge its limitations, including its focus on dyadic interactions and its neglect of nonverbal cues. They suggest that future research should explore multimodal features and include participants from across the psychosis spectrum and those with autism spectrum disorder to determine if dialogical patterns are specific to particular psychopathological groups. Critically, this work underscores that conversational impairments can be robustly measured using turn-taking metrics, reinforc-

ing the importance of precise TRP detection not only for artificial systems but also for understanding human communicative disorders.

Recent findings collectively suggest that systems utilising short-horizon voice-activity projections (VAP) provide consistently reliable timing with minimal inference delay, surpassing models that depend solely on text or transformer-based approaches reliant on lexical context [48,83]. Although integrating prosodic and linguistic features tends to improve temporal accuracy, the degree of improvement often declines when pause thresholds or segmentation criteria vary across corpora [85,100]. These insights imply that lightweight, audio-focused predictors are the most reliable foundation for real-time turn-end estimation, with linguistic or semantic signals serving more effectively as supplementary rather than primary predictors.

4.2. Backchannel Prediction and Generation

Backchannel prediction involves computationally determining when and what type of backchannel a listener is likely to produce during a speaker's utterance. In contrast, backchannel generation focuses on creating and deploying appropriate backchannels by artificial agents or systems. When considering backchannels in conversation, researchers primarily focus on their multifaceted role as brief listener responses that significantly contribute to the flow and dynamics of interaction. Key concepts include the timingof backchannels, which must be appropriate to support the speaker without interruption; their type or form, encompassing verbal ("uh-huh", "yeah") and non-verbal cues (head nods, smiles); their function, which can range from signalling continued attention ("continuer") and understanding to expressing agreement or empathy; the multimodal nature of backchannels, involving both auditory and visual signals; the context-dependency of their interpretation and appropriateness; and the influence of individual and socio-cultural variations in their production and perception.

In the backchannel prediction domain, Amer et al. [144] addressed a significant limitation in the literature on backchannel detection: the underutilisation of visual modalities, particularly body pose and facial expressions, for estimating listener agreement during backchannels. Their approach employed transformer architectures that had previously been successful in multimodal fusion and specifically evaluated various transformer configurations, such as One Stream and Cross Attention, applied to visual-only cues from the German group interaction dataset. This study leveraged dynamic visual features derived from body and facial movements, significantly improving the robustness of backchannel detection under compromised audio conditions. Their experiments showed that visual-only transformer models outperformed prosody-based baselines in noise-heavy settings, highlighting the value of visual grounding; however, performance dropped when applied outside German group discussions, indicating that the learned cues were not yet broadly generalisable. Ref. [145] tackled the problem of the time-consuming and non-scalable nature of manual backchannel annotation in existing datasets. They proposed a two-step semi-supervised methodology that automatically identifies backchannel opportunities and signals using listener features, then predicts them based on speaker context. Their findings on an Indian English dataset suggested that training with just 25% manually annotated data could yield performance comparable to 100%, offering a significant benefit to the HCI community and addressing the challenge of data scarcity, especially for low-resource languages. The study demonstrated that a semi-supervised model approach achieves full-supervision accuracy while reducing annotation effort by up to 75%. However, performance remained sensitive to annotation noise and domain mismatch, indicating scalability limits across languages and interaction styles. The study by [146] tackles the challenge of accurately modelling minimal verbal listener backchannels, specifically "Yeah"

and "Uh-huh" in English and their German equivalents, driven by minimal recipiency theory. They emphasise that prior computational models typically adopted a generalised "lumping" approach, treating all backchannels uniformly without accounting for their functional differences, thereby obscuring nuanced conversational behaviours. To address this limitation, they propose a fine-grained "splitting" approach that explicitly distinguishes between minimal verbal backchannels ("Yeah," "Uh-huh") and the absence of backchannels ("No-BC"). Results indicate that embedding either speaker or listener behaviours individually significantly improves prediction accuracy; however, combining these embeddings further enhances performance, thereby empirically validating the importance of speaker–listener interaction (SLI) modelling. Despite these gains, the model struggled to accurately predict "Yeah," likely because acoustic-only features failed to capture lexical-semantic dependencies, suggesting that future models require multimodal or linguistic information to handle semantically rich backchannels. Inoue et al. [147] highlighted the need for continuous, real-time backchannel prediction on unbalanced, real-world datasets, noting the limitations of turn-based or artificially balanced approaches. They proposed a method using a fine-tuned Voice Activity Projection (VAP) model, initially pre-trained on a general Japanese dialogue corpus and subsequently fine-tuned on their collected Attentive Listening Dataset (Japanese Wizard-of-Oz dialogues). Their model predicted both the timing and type (continuer vs. assessment) of backchannels frame by frame. Their results showed that VAP-based approaches outperform traditional prosody-driven methods in temporal precision, particularly for early prediction of continuers; however, performance on assessment backchannels lagged, revealing a limitation in capturing semantically more informed listener responses. The authors acknowledged additional limitations, including that the evaluation was primarily on a Japanese dataset and that there was no real-world evaluation with conversational agents or robots. Wang et al. [55] addressed the challenge of detecting subtle and sparse non-verbal backchannel cues. They proposed a model that incorporates temporal and modality-attention modules to focus on significant moments and their corresponding body parts. This approach led to improved accuracy in the MultiMediate '23 backchannel detection challenge. Their findings showed that attention-based architectures are more effective at capturing microbehaviors, such as eyebrow raises and slight nods, than traditional feature-based models. However, the accuracy of detection still faced limitations due to the infrequent nature of these cues in natural conversations. Ref. [148] highlighted the need to analyse all group members for backchannel detection, moving beyond individual audio–visual cues. They used a graph neural network to model group interaction through implicit and explicit behaviours, achieving strong performance on the MultiMediate '22 challenge for both backchannel detection and agreement estimation. This work showed that relational modelling, capturing who looks where, who reacts to whom, and how behaviours propagate, substantially outperforms single-person models. However, computational complexity and the need for multi-camera setups remain practical limitations. Expanding the categorisation of backchannels, ref. [149] aimed to go beyond simply predicting whether to backchannel by examining the relationship between backchannel type and the speaker's intent (dialogue act). They proposed a classification of backchannels into nine types and created a corpus of speaker dialogue acts and listener backchannel types, revealing a significant dependency of backchannel types on speaker dialogue acts. Their analysis showed clear patterns: for example, "neutral" and "positive" listener responses were more likely during information-provision dialogue acts, providing empirical grounding for context-aware backchannel generation. These findings provide a foundation for dialogue systems to generate more natural and contextually appropriate backchannels. Ishii et al. [150] noted that most prior work focused on single-task backchannel prediction, while backchannels co-occur with other conversation drivers, such

as turn-changing willingness. They proposed a multitask learning model to jointly predict backchanneling and turn-management willingness, suggesting that capturing dependencies between these aspects could enhance backchannel prediction in dyadic interactions. Their multitask model consistently outperformed single-task baselines, demonstrating the benefit of jointly modelling conversational behaviours; however, its performance declined in interactions with high speaker variability, indicating a need for more robust representations of individual differences. Onishi et al. [151] investigated the predictability of various listener backchannel types based on the speaker's multimodal information. They constructed a corpus of two-party dialogues annotated with nine functional backchannel types and built neural network models using visual, acoustic, and linguistic features extracted from the speaker's utterances. Their findings demonstrated that listener backchannel types are predictable from the speaker's multimodal cues, with acoustic features showing strong predictive power. This work provides empirical evidence for generating diverse backchannels in dialogue agents based on real-time speaker behaviour. Their multimodal models outperformed unimodal ones, with full-feature fusion yielding the highest F1 scores. Nevertheless, imbalanced data across backchannel types hindered prediction accuracy for rarer categories, underscoring the need for more balanced datasets.

On the other hand, backchannel generation was explored by [152], identifying a significant gap in existing research: most studies have focused exclusively on dyadic interactions, overlooking the critical role of third-party listeners in multiparty conversations, a gap currently under-explored in the multiparty interaction literature. To address this gap, they conducted a corpus analysis of radio show sessions featuring a human third-party listener and categorised backchannels into three types: responsive interjections, expressive interjections, and shared laughs. They then developed a continuous backchannel generation model that utilised separate models for predicting timing and form based solely on prosodic audio information, enabling real-time functionality. In their subjective evaluation, they compared their model to random, dyadic, and ground truth models. The results indicated that their model outperformed the random baseline and was comparable to the dyadic model (which was trained only on responses). However, the model struggled to generate expressive interjections effectively, which limited its perception of empathy and understanding. A key contribution of this work is the emphasis on the significance and challenges of developing appropriate expressive backchannels in a multiparty context. A limitation acknowledged was the use of augmented data, where the third party was not genuinely interactive with the primary speakers. Building on the significance of context, ref. [153] investigated how conversational contexts, specifically, information-centric versus emotion-centric, along with forms of non-lexical backchannels (such as the Korean terms "ne," "eo," and "eum")—influence user perceptions of robots. They noted a lack of prior research that considers these factors. Through a mixed-participant experiment utilising video stimuli of a robot interacting in different contexts with varying backchannel forms, the researchers found that task attraction and appropriateness were rated more positively in information-centric contexts. Additionally, attentive listening and understanding functions were more pronounced in these contexts. Notably, their mediation analysis indicated that the functions of non-lexical backchannels partially mediated the effect of conversational context on perceived appropriateness. However, contrary to their hypothesis, the specific form of the backchannel did not significantly impact perceived sociability. Their work emphasises the importance of considering conversational context in generating backchannels for robots providing information. A limitation acknowledged by the researchers was the strict manipulation of backchannel timing and the repetitive use of the same form, which could have resulted in an unnatural sound.

Shahverdi et al. [154] examined the nuances of backchanneling by investigating differences in human backchanneling behaviours when interacting with humans versus social robots across happy and sad scenarios. Their primary concern was that models trained solely on human-human interaction (HHI) data may not adequately generalise to human–robot interaction (HRI), given distinct patterns in emotional expression. In an exploratory study, participants listened to emotional stories narrated by either a human or a robot with human-like facial expressions, and their backchannels were coded and analysed for emotional specificity. Their research explicitly reports significant variations in emotionally specific backchannels between HHI and HRI, with human listeners exhibiting richer, more diverse, and contextually nuanced backchannel responses (e.g., expressive nodding accompanied by vocal affirmations like "mm-hm" or "really?") compared to interactions with robots, where backchannels were generally more restrained and limited in expressive gestures and vocal variety (e.g., simple "uh-huh" responses without additional expressive gestures). These findings emphasise the importance of developing specialised backchanneling models explicitly tailored for HRI, capable of capturing the subtler emotional and behavioural nuances distinctively observed in these interactions. However, a notable limitation of this study was its exploratory scope, constrained by a relatively small sample size of only 28 participants. Ref. [155] worked on Chillbot but also explored the concept of the "backchannel" in a different context: private feedback from moderators to users in online communities. They identified a significant problem: moderators lacked the tools to support behind-the-scenes interactions that could help defuse situations and educate users before taking formal action. To address this issue, they designed and deployed Chillbot, a Discord bot that enables moderators to send quick, anonymous feedback nudges to users. Their field deployment across eleven Discord servers showed significant usage, indicating the value of this non-punitive, backchannel approach to shaping community behaviour. While Chillbot does not directly generate backchannels in the same interactive manner as described in other studies, it illustrates the broader importance of private, less-confrontational communication as a means of influencing behaviour and maintaining social norms. This approach aligns with the underlying principle of providing feedback without disrupting the primary interaction. A key contribution of their work is demonstrating the value of designing moderation tools that support more socially engaged practices beyond just content removal. However, there are limitations to consider, including the focus on Discord and the challenges in directly observing the full context of bot usage due to privacy concerns.

Backchanneling is crucial for facilitating smooth, engaging conversations, allowing listeners to signal attention and understanding without interrupting. Properly timed backchannels encourage speakers to continue, elaborate, and feel acknowledged, thereby regulating turn-taking and maintaining a smooth flow of dialogue. Additionally, backchannels subtly convey agreement, empathy, or doubt, which can influence the speaker's direction. In human–robot interactions, effective robot backchanneling enhances perceptions of attentiveness, rapport, and social intelligence, promoting sustained interaction. Thus, nuanced comprehension and generation of backchannels are vital for developing natural and socially aware conversational agents.

Overall, studies in this area show that incorporating both visual and prosodic signals markedly improves the accuracy of backchannel timing and response type prediction [91,92]. Models that integrate dialogue-act or conversational intent awareness yield more context-sensitive responses, whereas purely acoustic approaches are prone to spurious triggers, particularly in spontaneous or multiparty dialogue [93]. Nevertheless, the benefits of multimodal fusion depend on the availability and reliability of visual input, which can degrade under conditions such as camera occlusion or low illumination. Consequently,

hybrid frameworks, anchored in audio features but enhanced by auxiliary visual cues, appear to offer the most balanced trade-off between generalisation and ecological validity.

4.3. Real-Time Turn-Taking Systems

The ability to predict and respond to conversational turns and backchannels in real time is crucial for achieving natural, engaging dialogue across various interactive systems. Real-time turn-taking is essential for avoiding awkward silences or disruptive interruptions, fostering a smoother and more human-like communication flow. Similarly, real-time backchanneling allows a listener (whether human or agent) to provide timely feedback, signalling attentiveness and understanding without disrupting the speaker, thus enhancing engagement and the perception of a natural interaction. Ref. [147] directly tackled this need for real-time responsiveness in conversational agents. They recognised that existing backchannel prediction methods often relied on utterance-based systems or artificially balanced datasets, which are not representative of real-world, continuous interactions. To address this, they proposed a novel method using a fine-tuned Voice Activity Projection (VAP) model that predicts both the timing and type of backchannels (continuers and assessments) in a continuous, frame-wise manner. Their VAP model, pre-trained on a large corpus of general Japanese dialogue and then fine-tuned on a specialised dataset for attentive listening with annotated backchannels, demonstrated robust performance in real-time environments, even with unbalanced data. Notably, their experiments showed that continuier-type backchannels could be reliably predicted with a shorter temporal context. In contrast, assessments benefited from longer input windows, highlighting a functional distinction between backchannel categories and demonstrating that some listener responses are inherently more context-dependent than others. Although the model achieved strong frame-wise accuracy, its performance was still sensitive to domain shifts, indicating that future work must improve cross-domain robustness.

The importance of evaluating turn-taking models in a way that reflects real-time performance was also a central concern for [156]. They argued that conventional metrics such as precision and F1 score fail to capture the real-time trade-off between responsiveness and accuracy, especially the rate of false cut-ins. Their proposed latency–false-cut-in curve showed that generalised deep learning models performed more reliably in loosely structured dialogue types (e.g., attentive listening), while scenario-specific models excelled in tightly constrained tasks such as interviews. This result indicates that turn-taking systems may require different optimisation strategies depending on conversational structure, and that "one-size-fits-all" models may be insufficient for broad deployment. While not explicitly focused on "real-time", ref. [157] developed a multimodal deep learning approach for predicting turn-taking events in human–agent interaction, emphasising efficiency for time-sensitive applications. Their model integrated text, audio, vision, and in-game contextual signals, outperforming audio-only baselines and demonstrating that multimodal cues yield more accurate predictions, particularly in task-oriented settings where visual grounding and contextual information strongly influence turn allocation. A key contribution of this work is the demonstration that multimodal inputs improve prediction stability under noisy audio conditions. However, the authors noted increased computational demands, which raise challenges for low-resource or embedded systems.

Similarly, ref. [158] introduced the original Voice Activity Projection (VAP) model for predicting future voice activity directly from raw audio. Their evaluation demonstrated that VAP can operate in real time on a CPU and outperforms several heuristic baselines in next-frame voice activity prediction. Crucially, VAP's ability to model short-term future activity enabled it to approximate the anticipatory behaviour characteristic of human turn-taking. However, the authors acknowledged that VAP's accuracy degraded with longer

prediction horizons, limiting its ability to model more complex turn-taking phenomena such as early interruptions or delayed holds. Finally, the study by [159] on creativity enhancement through turn-taking feedback provides a complementary perspective on real-time coordination, albeit in human–human interaction rather than human–agent systems. They found that real-time turn-taking feedback produced an instantaneous increase in idea generation, demonstrating that timing-sensitive conversational cues can shape behaviour at sub-second scales. Importantly, their results show that turn-taking signals need not produce gradual or long-term behavioural shifts to be effective; even brief, transient cues can meaningfully modulate conversational dynamics. This finding supports the broader argument that real-time responsiveness is essential for maintaining fluent dialogue, whether between humans or between humans and agents.

Broadly, recent systems demonstrate that adopting incremental inference and continuous prediction improves conversational smoothness and reduces false cut-ins [95,96]. Architectures that couple low-latency ASR with acoustic-based projection modules have achieved more natural overlap management and responsiveness. However, these advantages often come at the cost of increased computational burden and a reliance on robust speech recognition pipelines [57]. Current research trends thus favour adaptive pipelines that prioritise temporal precision while maintaining efficiency, enabling real-time operation without compromising human-like timing behaviour.

4.4. Modelling Turn-Taking in Multi-Party Interaction

The inherent complexity of managing conversational flow among more than two participants distinguishes multi-party turn-taking from its dyadic counterpart, making its study crucial for understanding natural social dynamics and designing effective human–computer interactions. While multi-party interaction can occur through voice-only channels, its natural and primary setting is often face-to-face and physically situated. Consequently, the modelling of multi-party interaction has been extensively explored within the context of human–robot interaction (HRI), as discussed in Section 2.5. However, virtual environments also provide platforms for studying multi-agent, single-user interactions.

Building on the foundational understanding that turn-taking is a fundamental organisational principle in conversation, ref. [160] addressed the problem of predicting turn shifts in four-participant settings by focusing on gaze behaviour, a key nonverbal cue in face-to-face interactions. They defined a set of gaze labels representing changes in speaker and listener gaze, encoded as 2-gram transition patterns. Their Naïve Bayes classifier, trained on the multimodal XJTU corpus, demonstrated that even short gaze sequences hold predictive value for identifying upcoming turn shifts. However, the model's dependence on only two consecutive gaze states represents a substantial limitation. As the authors themselves acknowledge, longer temporal histories of gaze behaviour, as well as interactions with other modalities, may provide richer predictive information, especially given the anticipatory nature of human gaze in multi-party conversation. Later work, such as [161], which found strong evidence that head-turns towards the next speaker occur before turn completion, supports this critique by showing that longer anticipatory cues matter for prediction accuracy.

Extending turn-taking research into human–robot interaction (HRI), ref. [162] investigated how robots can manage fluid multi-party conversations, moving beyond rigid rule-based systems and wizard-of-oz setups. Their analysis of a large human dialogue corpus revealed several critical capabilities robots must acquire: recognising turn-holding cues, identifying turn-yielding signals, and classifying the addressee of an utterance. Their multimodal classifiers for addressee recognition performed reasonably well, even using only audio, with ambiguous cases limited to 3.6%. This contrasts with [160], whose model

relied entirely on visual cues. While audio-only processing offers substantial privacy advantages and preserves functionality in visually constrained environments, it risks missing early nonverbal cues that humans rely on to anticipate turn changes. The authors also showed that pausing alone, a common heuristic in many dialogue systems, is insufficient for managing multi-party turn-taking, leading to inappropriate interruptions. This stands as an explicit critique of earlier, pause-only approaches, positioning their method as more socially aware and better aligned with human interaction norms. Building on this, Moujahid et al. [163] directly compared different turn-taking cue strategies, ranging from neutral behaviour to highly expressive multimodal cues, using a robot receptionist in a multi-party setting. Their results showed that combining verbal and nonverbal cues produced the most favourable impressions of intelligence, social awareness, and politeness. Strategies relying only on nonverbal cues were perceived as impolite, highlighting that nonverbal signals alone cannot sustain socially appropriate behaviour. However, the evaluation relied on pre-recorded demonstrations rather than real-time interactions, which the authors note may limit ecological validity. This methodological limitation echoes the constraints noted in [164], who explored gaze-cue visualisation in multi-party video conferencing. Their system, which represented gaze using arrows or dynamic window-size adjustments, showed that window-size manipulation led to more effective turn-taking and increased verbal participation. However, subjective evaluations were less conclusive, likely due to a small sample size and participants often looking away from the arrow indicators, indicating a mismatch between intended and actual user attention patterns. Unlike [163], which assessed robot behaviour, ref. [164] examined how to support human-human turn-taking in virtual environments; together, they emphasise the central importance of gaze and other nonverbal cues across interaction modalities.

In immersive VR environments, Wang et al. [165] leveraged rich multimodal tracking data to predict turn-taking during open-ended group activities. Their models, informed by speech, motion, and gaze, demonstrated that turn-taking patterns remain predictable even in highly dynamic settings. The study also identified influential factors such as group size and listener personality, offering insights unavailable in physical-world or video conferencing settings. Complementing this, Hadley and Culling [161] found that head turns reliably anticipate upcoming speakers in triadic conversations, providing real-world evidence that listeners signal their engagement before turn boundaries occur, highlighting the importance of modelling anticipatory listener cues rather than just speaker signals. For spoken dialogue systems, ref. [115] introduced the Timing Generating Network (TGN), which predicts precise interjection timing for multi-party conversations, addressing a key limitation of end-of-turn detection approaches that often miss opportunities for natural insertion points. Their results indicate that fine-grained timing substantially improves system responsiveness compared to traditional heuristics. Finally, ref. [166] tackled the challenge of realistic simulation for multi-speaker ASR training. Their generative model, built on detailed utterance-timing statistics, reproduces the temporal interplay in real conversations more accurately than models that treat speakers independently. As shown in their analysis (see detailed turnover overlap statistics in Table 1 of their paper), models that capture inter-speaker dependencies outperform independent-speaker models in both realism and downstream ASR utility. Their work demonstrates that modelling turn-taking interaction structure is crucial for generating valid training data for diarisation and speaker extraction.

Collectively, these studies illustrate both the progress and outstanding challenges in multi-party turn-taking research. While [160] provides foundational insights into short-horizon gaze patterns, later work such as [161–163] emphasises the necessity of modelling richer, longer-range and multimodal cues. Studies differ substantially in the modalities used, the interaction contexts, and the predictive tasks, which leads to varying levels of

performance and generalisability. Approaches that incorporate multimodality, particularly gaze, head pose, and prosody, consistently achieve better prediction accuracy and perceived naturalness. In contrast, models relying on narrow input modalities or simplified temporal assumptions tend to underperform or fail to generalise.

Table 1. Comparative Summary of Datasets (access dates shown as DD–MM–YYYY).

Dataset	Year of Release	Data Types	Sample Size	Use Cases & Link
HCRC Map Task [63]	1991	Text, Audio (Video recorded)	128 dialogues (~15 h), 64 participants	Spontaneous speech, turn-taking analysis https://groups.inf.ed.ac.uk/maptask/ HCRC Corpus (17 April 2025)
Switchboard [64]	1992	Text, Audio	~2500 conversations, ~500 speakers, (~250 h), ~ 3 M words text	ASR, speaker authentication https://catalog.ldc.upenn.edu/LDC97S62 Switchboard Corpus (17 April 2025)
Fisher [65]	2004	Text, Audio	16,000 conversations, 2000 h	ASR, conversational speech processing https://catalog.ldc.upenn.edu/LDC2004S13 Fisher Corpus (17 April 2025)
CANDOR [66]	2023	Text, Audio, Video, Facial Data	1656 conversations, 1456 participants, 850+ h, 7 M+ words	Multimodal interaction, turn-taking, rapport, backchanneling https://guscooney.com/candor-dataset/ CANDOR Corpus (17 April 2025)
NoXi [67]	2017	Text, Audio, Video, Depth, Sensor Data	25+ h, dyadic interactions, 7 languages, 58 topics	Novice-expert dialogues, turn-taking, conversational interruptions https://multimediate-challenge.org/datasets/Dataset_NoXi/ NoXi Database (17 April 2025)

4.5. Multimodal Turn-Taking Approaches

Multimodal turn-taking has emerged as a crucial area of research in understanding and modelling natural human communication, recognising that conversations are rich in verbal and nonverbal information. Face-to-face communication relies heavily on multimodal cues to facilitate effective turn-taking. In this context, our perceptual abilities are engaged as the interaction is highly multimodal [50]. The initial focus on verbal (linguistic and semantic) and acoustic (pauses and pitch variations) cues for predicting turn shifts has

evolved to incorporate the significant role of visual cues, such as gaze, gestures, facial expressions, head movements, and contextual information. A significant challenge in advancing multimodal turn-taking research, particularly in human–robot interaction, has been the scarcity of public multimodal datasets and the limitations of current methods that often rely on unimodal features (because such traditional approaches frequently struggled with precise timing predictions, especially in spontaneous dialogues featuring frequent overlaps and interruptions) or simplistic multimodal ensemble models. Furthermore, real-world scenarios often present issues such as class imbalance, leading to misclassifications of turn endings with short pauses. Multimodal turn-taking is relevant because it seeks to replicate the nuanced dynamics of human-human interaction, thereby enabling more natural and effective spoken dialogue systems and human–robot communication. Our review is limited in its scope, focusing exclusively on the most recent multimodal turn-taking research.

Building on the understanding that communication involves continuous exchanges beyond just words, Yang et al. [167] developed a comprehensive annotated dataset comprising over 5000 authentic human–robot dialogues, covering speech and text modalities and targeting endpointing (user end-of-turn) and barge-in (user interruption) scenarios. Their gated multimodal fusion (GMF) framework, combining OpenSmile-ResNet acoustic features [168], Transformer-based text embeddings, and fine-grained timing features, consistently outperformed strong baselines, with ablation studies showing that multimodal fusion and conversational context substantially boosted predictive accuracy. Notably, contrastive learning and data augmentation mitigated class imbalance, an issue that often degrades endpointing models. However, despite these gains, their results indicated that text-based cues played a disproportionately large role, suggesting that systems heavily dependent on linguistic input may struggle in low-ASR-quality environments or with spontaneous speech. Kurata et al. [49] further demonstrated the importance of visual cues by showing that eye movements were the most influential predictors among gaze, mouth, and head cues. Their end-to-end 3D-CNN (X3D) model outperformed LSTM-based approaches, particularly in capturing complex temporal–spatial changes in facial dynamics. Importantly, their multimodal model combining acoustic, linguistic, and visual cues yielded the highest end-of-utterance accuracy, confirming that visual cues compensate for weaknesses in audio-only predictions, especially under noisy acoustic conditions. Nevertheless, their findings also revealed that head pose and articular point cues contributed minimally, likely due to noisy 3D landmark extraction, highlighting that not all visual modalities are equally informative.

Onishi et al. [51] complemented these findings by examining non-verbal cues (gaze direction, action units, head pose, articulatory points) within a Transformer-based voice activity prediction framework. Using the NoXi dataset [67], they found that non-verbal-only models significantly outperformed audio-only models, with action units and gaze direction contributing the most significant gains. Combining all modalities improved next-speaker prediction, but head pose and articulatory features showed limited predictive utility, suggesting feature-specific differences in reliability. Their work reinforces that non-verbal cues enhance turn-timing prediction, but also that overly broad multimodal fusion can dilute performance when low-quality or weakly informative features are included. Addressing the limitations of third-person data in multiparty prediction, ref. [169] proposed the 3M multimodal transformer architecture using egocentric inputs, achieving marked improvements across short- and long-term horizon predictions on the EgoCom dataset. Their results highlight the advantage of egocentric visual fields, which better reflect the perceptual signals humans use for turn anticipation. However, the computational complexity of multi-stream transformers remains a barrier to deployment in lightweight or

embedded systems. Lee and Deng [170] advanced this line of research with an online window-based PLM-GRU fusion approach for real-time multiparty turn prediction. Their model, integrating prosody, gaze targets, interlocutor states, and PLM-derived linguistic features, outperformed IPU-based baselines, particularly in overlap-heavy scenarios. Their findings emphasise that window-based continuous processing better captures rapid turn transitions than segment-based models. Nonetheless, because their dataset involved only three-party interactions in controlled conditions, the model's generalisability to larger, more chaotic social settings remains uncertain.

Complementing these predictive models, ref. [171] examined speech–gesture synchronisation using cross-wavelet analysis, revealing strong intrapersonal and interpersonal coordination at multiple timescales. Their findings of significant synchronisation across movement–movement, voice–voice, and cross-modal pairings provide foundational evidence that turn-taking emerges through rhythmic coupling of vocal and gestural behaviours. However, the small sample (14 dyads) and lack of cultural diversity limit generalisability, and the method, while analytically robust, does not directly yield actionable predictors for real-time systems. These studies collectively demonstrate that multimodality consistently improves turn-taking prediction, but also reveal nuanced differences in the usefulness of each cue type: gaze and action units strongly enhance detection accuracy, text cues dominate when available, and some visual features (e.g., head pose) provide minimal benefit. The effectiveness of egocentric versus third-person views further highlights that the perceptual framing of input data fundamentally shapes model utility. These cross-study comparisons offer valuable insight into which multimodal approaches hold the most promise for advancing human-like conversational turn-taking, and which limitations remain to be addressed.

Evidence across multimodal studies underscores that visual features, particularly gaze shifts, head motion, and facial activation patterns, add complementary information to audio signals when modelling turn transitions. Fusion strategies that consistently attend to these cues outperform unimodal baselines, although gains diminish under poor lighting or camera occlusion. The field is therefore moving toward selective fusion, in which visual streams refine rather than dominate acoustic prediction, ensuring resilience across interaction settings. To succinctly synthesise the extensive literature examined in this section, Table 2 summarises and compares pivotal findings and computational approaches explored across recent studies, which, due to their length and detail, are split across multiple pages. These tables collectively synthesise key findings, methods, and datasets discussed, enabling quick reference and comparative evaluation across recent studies on human–robot turn-taking. Together, these tables serve as a concise reference to navigate through the breadth of research and inform the design of future conversational agents capable of natural turn-taking.

From a practical standpoint, audio-based voice activity projection (VAP) and other short-term predictive models are the most reliable for real-time turn-taking tasks. They provide the best balance between timing accuracy and computational cost. Multimodal extensions, incorporating gaze, facial action units, or prosodic cues, further enhance performance in face-to-face or socially rich contexts. However, their effectiveness depends on the reliability of the sensors and the quality of the synchronisation. Large language models (LLMs) offer valuable contextual reasoning but still fall short in precise temporal alignment. This makes them better suited as complementary modules rather than standalone predictors. Overall, the most promising approach involves hybrid architectures that combine the temporal precision of acoustic cues with the semantic depth of language-level reasoning. These should be supported by adaptive learning frameworks that generalise across different domains and speakers.

Table 2. Summary of Studies on Turn-taking, Backchannel, Real-time and Multimodal Predictions.

Area of Focus	Articles/Paper	Architecture	Modality	Key Findings	Datasets Used	Limitations
Turn-taking Prediction	[134]	Voice Activity Projection (VAP)	Audio	Commercial TTS systems often produce ambiguous turn-taking cues. TTS systems trained on reading or spontaneous speech exhibit strong turn-holding but weak turn-yielding cues.	MultiWoZ	Ambiguity in synthesized cues affects consistent and accurate interpretation by users.
	[132]	Voice Activity Projection (VAP)	Audio	Prosodic cues, particularly phonetic information is crucial for turn-taking prediction. F0 information is less important for general turn-taking but plays a significant role in distinguishing between turn-shifts and turn-holds.	Switchboard, Fisher	Models not trained on perturbed data; Relied on TTS for ambiguous phrases; Limited linguistic context.
	[139]	Large Language Models (LLMs)	Text	LLMs effectively predict what to say but struggle with accurately predicting turn-taking timing (TRPs)	In Conversation Corpus (ICC) – Novel TRP dataset	Inability to accurately predict within-turn TRPs, highlighting limitations in timing cues.
	[72]	E2E Self-supervised (Wav2Vec 2.0, roBERTa)	Audio, Text	A simple Wav2Vec 2.0 model with a linear classifier outperformed other models on datasets with lower signal-to-noise ratios and greater acoustic and language complexity.	Harper Valley, Switchboard, IBM datasets	Did not incorporate dialogue context and only addressed end-of-turn detection. It focused on binary turn-shift tasks and lacks continuous frame prediction.
	[122]	VAP with cross-attention Transformers	Audio	Monolingual VAP models do not generalize well across languages. Multilingual VAP performs similarly to monolingual models for English, Mandarin, and Japanese.	Switchboard (English), HKUST (Mandarin), Travel Agency Task Dialogues (Japanese)	Monolingual models have difficulty when applied directly across languages without retraining, especially with low-resourced languages and limited pitch variability
	[133]	Voice Activity Projection (VAP)	Audio	VAP effectively predicts voice activity events and models conversational dynamics accurately. (Effectively resolves overlaps.)	Switchboard, Fisher	Vulnerability to acoustic artifacts affecting predictive reliability. (struggles in noisy environments.)
	[137]	Neural Source Separation Network	Audio	Leveraging turn-taking dynamics effectively extracts target conversations from audio mixtures containing interfering speakers and noise.	LibriSpeech, AISHELL-3, RAMC, AMI Meeting Corpus	Performance deteriorates significantly when natural turn-taking timing is disrupted

Table 2. *Cont.*

Area of Focus	Articles/Paper	Architecture	Modality	Key Findings	Datasets Used	Limitations
Turn-taking Prediction	[172]	Voice Activity Projection (VAP)	Audio	Fillers have strong turn-holding effects, especially in yielding contexts. Prosodic features of fillers (pitch, intensity, duration) contribute significantly.	Fisher, Switchboard	Fillers have limited impact due to redundancy with other turn-holding cues.
	[123]	Voice Activity Projection (VAP)	Audio	Modelling-dependent voice activity events improves accuracy for predicting future conversational events.	Switchboard	Requires precise modeling of voice activity dependencies to maintain accuracy.
	[173]	GPT-2, TurnGPT (GPT-2 augmented with explicit speaker embeddings), Transformer	Text	Fine-tuned LLMs find incongruent speaker transitions more surprising than congruent ones but do not exhibit the same tendency during holds. These models do not consider speaker identity as humans do, and further fine-tuning tends to yield diminishing returns.	In Conversation Corpus (ICC)	LLMs fail to replicate normative interactive structure of natural conversation adequately
	[174]	LSTM, syntactic completeness prediction	Audio	Improved response timing significantly enhances user satisfaction in spoken dialogue systems.	InteLLA conversational dataset	Difficulty confirming how timing errors affect user perception clearly.
Backchannel Prediction	[175]	Multi-Task Learning Framework	Audio, Text	Dialogue act information enhances backchannel prediction performance, particularly for assessments.	Switchboard	High dependency on specific task labels; generalizability untested across languages.
	[146]	Neural-Based Acoustic Classifier	Audio	Encoding both speaker and listener behaviour improves backchannel prediction performance, with speaker-listener interaction modelling leading to the best results.	Switchboard, GECO	The analysis did not consider the relationship between listener characteristics and the F1-score. It is limited to minimal responses and lacks a comprehensive interaction analysis.
	[140]	LLM + Acoustic (HuBERT + GPT-2/RedPajama)	Audio, Text	Fusing LLM and acoustic models improves turn-taking and backchannel prediction. Multi-task instruction fine-tuning enhances the LLM's ability to understand task descriptions and dialogue history.	Switchboard	Future research should explore the effects of using automatic transcriptions in place of ground-truth transcriptions. High-quality ASR output is reliant, and the current approach lacks real-time implementation and adaptation to various conversational contexts.
	[147]	Fine-tuned Voice Activity Projection (VAP)	Audio	Fine-tuning the Voice Activity Projection (VAP) model enables continuous and real-time backchannel prediction, including both timing and type. Intensity dynamics play a significant role in backchannel prediction.	LibriSpeech, Switchboard – (Kyoto Backchannel Corpus)	Requires careful tuning of threshold values, sensitivity to dataset imbalance and real-time operational constraints

Table 2. *Cont.*

Area of Focus	Articles/Paper	Architecture	Modality	Key Findings	Datasets Used	Limitations
Backchannel Prediction	[176]	Attention-Based Generative Model	Visual, Audio	Context-sensitive smiles significantly enhance human–agent rapport in mental health dialogues.	RealTalk Dataset (YouTube-based face-to-face interactions)	Results specific to mental health dialogues; limited generalization to broader interactions.
Real-time Turn-taking	[158]	Voice Activity Projection (VAP) with Real-Time Optimization	Audio	The Voice Activity Projection (VAP) model offers high accuracy and minimal inference time for real-time turn-taking prediction. Real-time VAP functions effectively with little performance degradation.	Japanese Travel Agency Task Dialogue dataset, Switchboard, HKUST	The study concentrated on a single modality (audio) and did not account for the effects of other modalities on real-time turn-taking prediction, which is suboptimal with shorter context windows
	[121]	Multi-Look-Ahead ASR with Estimates of Syntactic Completeness (ESC)	Audio	Extremely low latency ASR (Multi-Look-Ahead ASR) enables near-full utterance use for real-time response timing estimation, improving accuracy.	Simulated Japanese dialogue database	Latency in streaming ASR limits the available linguistic context for timing estimation.
Multimodal Turn-taking	[51]	Voice Activity Projection (VAP) with Non-Verbal Feature Integration	Audio, Visual	Incorporating nonverbal features such as action units and gaze direction enhance the accuracy of turn-taking event prediction in expert novice conversations. Additionally, including gaze and gesture features increase the accuracy of voice activity prediction.	NoXi database (Expert-Novice Conversation Dataset)	There is limited exploration of the specific action units that are most effective for prediction. Additionally, accurate predictions are highly dependent on the quality and completeness of multimodal inputs.

5. Deep Learning (Network Architectures) for Turn-Taking Prediction

When considering deep learning architectures for turn-taking in spoken dialogue, researchers must navigate the inherent temporal and contextual complexities of human conversation. The coordination of speaking turns depends on subtle verbal and non-verbal cues that evolve over time. A primary focus of this review article is to promote the development of new methods and tools for analysing these models. This will enable us to uncover the complex cues underlying predictions about turn-taking. These tools will not only enhance the practical application of these models but also provide new insights into the fundamental mechanisms of inter-speaker coordination. This section of the review focuses on the computational modelling of turn-taking using deep learning. Here, we provide a brief overview of the deep learning architectures used in this research field, including Recurrent Neural Networks (RNNs), Contrastive Predictive Coding (CPC), and transformer networks.

Early deep learning efforts in this domain leveraged Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to capture the sequential dynamics of voice activity and model the continuous flow of dialogue [70]. For instance, ref. [70] pioneered continuous turn-taking models using LSTMs to predict future speech activity, moving away from reactive, classification-based approaches on Inter-pausal Units (IPUs).

5.1. Contrastive Predictive Coding

Furthermore, in the Contrastive Predictive Coding (CPC) architecture for turn-taking, researchers focus on its ability to learn high-quality representations from raw audio data by predicting future latent representations. The core idea of CPC, introduced by [177], is to capture shared information across different parts of a signal while discarding local noise, making it particularly suitable for modelling the temporal dynamics of spoken dialogue. The typical CPC architecture for speech consists of an acoustic encoder g_{enc}, often a multi-layered convolutional neural network (CNN) [178,179], to process the raw waveform and extract lower-level features, followed by an autoregressive model g_{ar}, such as a Long Short-Term Memory (LSTM) network [180], to aggregate these features over time and predict future encoded representations, as depicted in Figure 9. In the research work by [123], the Voice Activity Projection (VAP) model leverages a CPC-based speech encoder to learn useful speech representations directly from the raw waveform for predicting future voice activity, which is central to modelling turn-taking dynamics. Rivière et al. [181] further refined CPC for the acoustic domain, demonstrating its utility across languages. Their English model was used in the VAP models presented in the thesis. This approach is essential because it allows the model to learn task-relevant acoustic features without relying on extensive manual feature engineering, enabling it to capture nuances in prosody and timing that are crucial for turn-taking. Based on Noise-Contrastive Estimation (NCE), the contrastive loss function used in CPC trains the model to discriminate the accurate future representations from a set of negative samples, encouraging it to learn representations that are predictive of the future while being robust to irrelevant variations. Discussing CPC in the context of turn-taking is relevant because it represents a significant step towards self-supervised learning of meaningful acoustic features, paving the way for more sophisticated models that can understand the intricate relationship between speech signals and conversational dynamics, including the prediction of turn shifts and backchannels, as demonstrated by the zero-shot tasks performed by the VAP model [123]. While subsequent work has explored transformer-based architectures, such as TurnGPT [71], for leveraging textual context, CPC remains a foundational technique for effectively learning from the acoustic

domain, highlighting the importance of directly modelling the speech signal for understanding turn-taking.

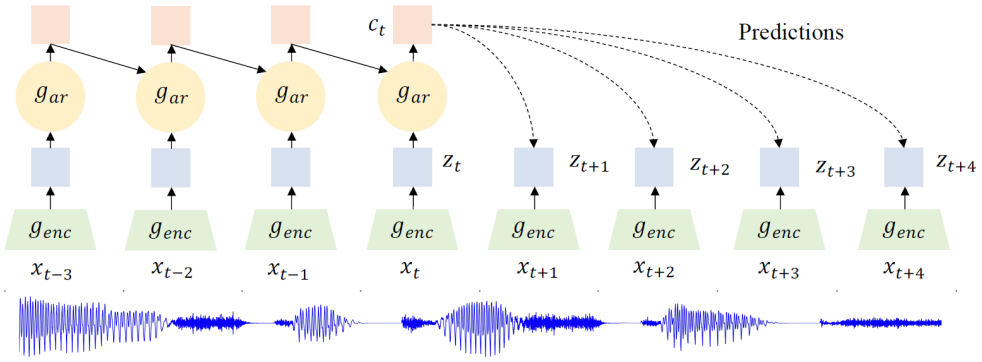


Figure 9. The CPC architecture predicts future latent representations, providing a comprehensive framework for extracting meaningful speech representations. Source [177].

5.2. Transformers

The crucial role of context in turn-taking prediction further motivated the adoption of transformer-based architectures, which excel at modelling long-range dependencies through their self-attention mechanisms. The attention mechanism, the core of the transformer architecture introduced by [182] (as depicted in Figure 10), allows the model to weigh the importance of different parts of the input sequence when processing each element, enabling it to capture intricate relationships that span across multiple turns. This is particularly relevant for turn-taking, as cues for turn completion can be subtle and dependent on the broader pragmatic context of the dialogue rather than just immediate syntactic or prosodic features. The transformer architecture, typically comprising multi-head self-attention layers, positional encoding, and feed-forward networks within encoder and decoder blocks (or decoder-only variants), marks a significant departure from recurrent and convolutional networks that previously dominated sequence transduction tasks. In the domain of turn-taking, TurnGPT [71] stands out as an early and influential example of leveraging transformers. Based on the GPT-2 architecture, TurnGPT is a unidirectional, decoder-only transformer that frames turn-taking prediction as a language modelling task. By augmenting the vocabulary with special turn-shift tokens, the model learns the probability distribution of these tokens alongside regular words, effectively representing the likelihood of a Transition-Relevance Place (TRP). The input to TurnGPT typically includes word embeddings and positional embeddings to account for the sequential nature of language, as well as speaker embeddings, allowing the model to differentiate between interlocutors. The success of TurnGPT demonstrated the effectiveness of transformers at capturing linguistic context relevant to turn-taking, outperforming previous RNN-based approaches. A key nuance of transformer techniques, such as the one employed in TurnGPT, is their ability to project future turn completions by generating continuation sequences and observing when a speaker token is sampled. Positional encoding techniques, such as ALIBI (attention with linear biases), further enhance the model's ability to extrapolate to longer sequences, potentially improving its ability to model extended turns. The relevance of discussing transformers in turn-taking research lies in their state-of-the-art performance in natural language processing, their ability to model long-range dependencies critical for understanding dialogue context, and their adaptability to various predictive tasks related to turn-taking, such as TRP prediction and turn completion projection. While primarily operating on sequential data, ongoing research explores methods to integrate the strengths of transformers with acoustic models (like VAP, which also utilises a transformer-based architecture) and potentially handle the overlapping nature of spoken dialogue more effectively.

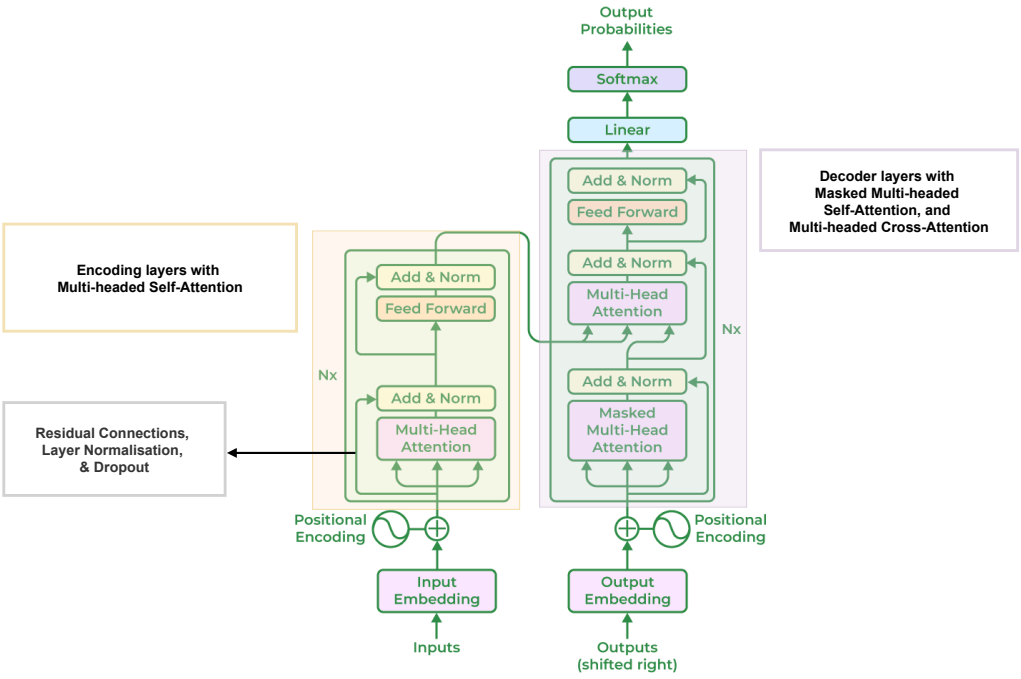


Figure 10. The full transformer architecture features an encoder module on the left and a decoder module on the right. Source [182].

The transition towards large language models (LLMs) underscores the growing emphasis on leveraging their strong contextual understanding and generative capabilities to build more natural, human-like conversational agents. However, effectively modelling the real-time, overlapping nature of the spoken dialogue with predominantly sequential models remains a significant challenge. The relevance of discussing deep learning in this context lies in its ability to learn intricate patterns from vast amounts of data, enabling spoken dialogue systems to move beyond simplistic rule-based or silence-detection policies towards more nuanced, predictive, and ultimately more natural turn-taking behaviours, which are crucial for seamless human–computer interaction. Recognising the multimodal nature of human communication, ref. [140] explicitly explored the fusion of acoustic models (HuBERT) with linguistic large language models (GPT-2 and RedPajama) to improve predictive accuracy in turn-taking and backchannel tasks. They demonstrated the effectiveness of integrating these modalities through multitask instruction fine-tuning, improving model performance, and contextual understanding.

5.3. Input Features and Data Flow

Understanding how spoken dialogue data is processed and fed into neural network architectures is fundamental to developing effective turn-taking models across the verbal and acoustic domains. In the verbal domain, the initial step typically involves transcribing spoken interactions into textual form. Given that turn-taking is inherently a dynamic process with overlaps and backchannels, a crucial part of the extraction process is segmenting the continuous dialogue into individual turns. This can involve identifying Inter-Pausal Units (IPUs) defined by short silence durations. Handling overlaps and backchannels during this segmentation is a nuanced process, with some approaches excluding isolated backchannels or specific types of overlaps to create a more linear turn sequence for language models. Once segmented, the textual data undergoes tokenisation, where words are converted into discrete units (tokens), often at the word or subword level. For model training, these token sequences are augmented with speaker identifiers or embeddings to distinguish interlocutors and with positional embeddings to encode the sequential order of

words within and across turns. This sequence of token embeddings, often representing the dialogue history, serves as input to sequence-based models, such as transformer networks, which aim to predict turn-taking-related events based on linguistic context. Labelling in the verbal domain can be implicit, where language models are trained to predict the next token, and the occurrence of a unique turn-shift token serves as the target for learning turn-taking probabilities. Alternatively, explicit labelling might involve annotating turn boundaries within the token sequences.

In the acoustic domain, the processing pipeline starts with the raw audio waveform of spoken dialogue. Unlike text, which is discrete, audio is a continuous signal that requires framing or segmentation for effective processing. A common initial step is to extract frame-level acoustic features at a specific rate (for example, 20 ms frames at 50 Hz). Traditional methods often relied on hand-engineered features such as Mel-frequency cepstral coefficients (MFCCs). However, modern approaches increasingly use models like Contrastive Predictive Coding (CPC) to learn speech representations directly from the raw waveform, deriving frame-level speech features without relying on prior feature extraction. Additionally, Voice Activity Detection (VAD) is often used to segment speech and silence for each speaker, producing binary voice activity features. Models also utilise the history of voice activity, measuring the activity ratio between speakers over different time windows. This provides longer-range contextual information. The frame-level acoustic features, potentially combined with voice activity data, serve as input to various neural network architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer networks. This enables the model to learn temporal patterns and prosodic cues relevant to turn-taking directly from the audio signal. For instance, the Voice Activity Projection (VAP) model uses a self-supervised objective to predict the joint voice activity of both speakers in future time frames. Labelling in this domain involves defining target VAP states that represent different combinations of future speaker activity. Temporal batching and windowing are crucial in acoustic modelling, as continuous audio is processed into fixed-length segments, or windows, during training.

While this review primarily focuses on the verbal and acoustic domains in isolation, multimodal approaches incorporating visual cues, such as gaze and gestures, are recognised as necessary for human turn-taking. These would involve extracting features from video data and aligning them with audio and text. Training and evaluation strategies differ across domains. Verbal models are often trained using the language model loss (cross-entropy) and evaluated on their ability to predict turn shifts. Acoustic models, such as VAP, are trained with prediction-based losses and can be evaluated in a zero-shot manner on various turn-taking tasks. Offline evaluation metrics are commonly used to assess model performance by comparing predictions against recorded human turn-taking behaviour. However, the correlation between offline metrics and real-time interactive performance remains an area of investigation. Despite progress, significant challenges remain. Many datasets lack adequate representation of diverse demographics or interaction contexts, potentially leading to biased or contextually limited models. Additionally, large, resource-intensive neural architectures pose challenges for real-time applications, as they often exceed the computational constraints of practical deployment scenarios. Robust handling of overlapping speech, ambient noise, and ambiguous speaker cues remains a substantial obstacle to reliable turn-taking prediction. Including this detailed discussion in this review is crucial, as it provides a foundational understanding of how turn-taking models learn from data, highlights distinct approaches across domains, and underscores existing challenges, enabling readers to critically evaluate and contribute to this complex and evolving field.

6. Conclusions and Future Directions

Across the contributions reviewed in the literature, several compelling insights align with the objectives of this comprehensive review. One of the key points emphasised by [9] is that all social activities require a turn-taking protocol that defines the sequence and participants involved in each action. Similarly, ref. [123] identifies turn-taking as a fundamental mechanism in human interaction, enabling participants to coordinate who speaks at any given moment, thus preventing simultaneous speech and listening. The study further emphasises that each turn represents a controlled segment of interaction, where the current speaker directs the flow of dialogue, suggesting that modelling these dynamics could greatly benefit conversational systems. While existing studies on turn-taking within spoken dialogue systems generally treat the concept as static, the broader literature highlights its cultural variability [30] and sensitivity to contextual factors [48]. Traditional computational approaches, which often rely on silence or IPU-based heuristics, struggle to accommodate these complexities in real-time interactions, particularly when managing overlaps, hesitations, and backchannels (as discussed in Sections 2.1 and 3.2). Therefore, developing effective conversational systems demands adaptability and continuous adjustment to navigate these dynamic turn-taking behaviours.

This review has identified several gaps in turn-taking research, some of which have been addressed by recent works, as outlined in (Section 4). A key challenge in building naturalistic spoken dialogue systems is enabling seamless turn management, ensuring systems can predict and coordinate turns effectively to avoid interruptions and long pauses. Achieving this requires access to high-quality data, yet publicly available datasets in this field are limited in size and accessibility. Consortia control many relevant datasets and make them available only at a high cost. In addition to data limitations, this study highlights the constraints of current frameworks that attempt to align more closely with natural human behaviour. In contrast, recent advancements demonstrate that Large Language Models (LLMs) provide a richer contextual understanding of dialogue by incorporating complex linguistic and pragmatic cues, surpassing simpler models in turn-ending predictions. These models are not without shortcomings. Studies such as [142,183] show that LLMs can analyse deeper language structures and recognise subtle speaker intent behind pauses, which is critical for seamless turn management. However, assuming their strong response generation capabilities automatically translate into effective conversational performance would be naive. Modern incremental processing models, such as Voice Activity Projection (VAP), have significantly improved predictive accuracy by leveraging prosodic and acoustic features through self-supervised learning, as discussed in (Section 4.1.2). Yet, despite these advances, LLMs still struggle with precise turn-taking predictions due to their limited incremental processing and insufficient sensitivity to real-time acoustic and multimodal signals, which are essential factors in spontaneous conversational contexts.

The preference for LLMs over simpler pre-trained models stems from their advantages in handling turn-taking and conversational dynamics. Unlike traditional systems that rely on fixed silence thresholds or shallow linguistic cues, often resulting in unnatural interactions, LLMs analyse deeper structural and semantic aspects of language. This enables them to predict turn completions and yield points with a more holistic understanding of dialogue context and speaker intent. Moreover, LLMs are particularly adept at managing spontaneous, less-structured conversations, distinguishing between pauses that indicate turn relinquishment and those signalling a speaker's intent to continue. This capability is especially critical in interactions with elderly users, who may exhibit longer pauses. While current LLMs demonstrate considerable progress in these areas, Moshi [184], a promising audio foundation model, seeks to bridge the remaining gaps with its full-duplex, end-to-end architecture. Unlike conventional systems that integrate speech recognition, natural

language understanding, dialogue management, and text-to-speech as separate components, Moshi operates holistically, listening and generating audio almost simultaneously. This approach offers a glimpse into a future of genuinely interactive spoken dialogue. However, the path to seamless human–agent conversation remains fraught with challenges, and even a sophisticated model like Moshi exposes the limitations of current LLM-driven approaches to turn-taking. Despite its architectural strengths, ref. [185] reveals several critical drawbacks, which are mitigated by advanced evaluation protocols. Moshi occasionally hesitates to take the floor, remaining silent even when the user has finished speaking. This reluctance can lead to unnatural pauses, disrupting the conversational rhythm. Conversely, when Moshi does engage, it can be overly assertive, interrupting at inopportune moments in a manner that feels disruptive rather than cooperative, contrasting sharply with the subtle, supportive overlaps characteristic of human dialogue.

Additionally, Moshi demonstrates a notable reluctance to provide brief backchannel verbal acknowledgements, such as "uh-huh" and "mm-hmm," that signal active listening and encourage the speaker to continue. The absence of these subtle cues can leave users feeling as though they are speaking into a void, uncertain of the conversational agent's engagement. Another key limitation lies in Moshi's ability to convey its conversational intentions. Users are often left guessing whether Moshi has completed its turn or intends to continue speaking, leading to awkward overlaps or premature user responses. Moreover, when users attempt to interject, Moshi frequently disregards the interruption and continues as if the user had not spoken. This failure to yield the floor when appropriate undermines the fluidity of interaction and can create frustration. While these limitations are specific to Moshi, they emphasise broader challenges that LLMs face in mastering natural turn-taking. Although LLMs excel at generating fluent language, they often struggle with the precise timing and contextual sensitivity required for truly dynamic exchanges. Their shortcomings stem from a limited sensitivity to prosodic cues, pragmatic implications, and the real-time feedback loops that regulate human conversation. The evaluation of Moshi highlights the pressing need to move beyond simple input–output paradigms and toward more sophisticated models capable not only of understanding and generating speech but also of orchestrating their participation in a fluid, responsive manner. Achieving this requires integrating LLMs with turn-taking predictive models. While turn-taking models excel at identifying patterns and timing for speaker transitions, they often lack the depth to interpret conversational semantics and intent. A complementary dynamic is achieved by coupling LLMs with these models: LLMs deliver rich conversational insights. In contrast, turn-taking models refine the temporal aspects, ensuring transitions are seamless and contextually appropriate.

Future research should address key limitations highlighted in this review, particularly the biases stemming from datasets lacking demographic diversity and the limited availability of corpora maintained by specific research groups. Creating accessible, contextually rich, and diverse datasets is essential to ensuring robust, generalizable research on turn-taking. Furthermore, there should be a stronger emphasis on the sophisticated integration of multimodal signals, such as gaze, gestures, and prosodic elements, to enhance both predictive accuracy and the naturalness of conversations. Additionally, exploring advanced predictive models that closely mimic human turn-taking latency could significantly improve interactive systems. Ultimately, interdisciplinary research that combines insights from linguistics, cognitive neuroscience, and computational modelling will be crucial to achieving more intuitive, human-like conversational interactions. Moreover, developing frameworks for identifying Backchannel Relevant Places (BRPs) will significantly enhance conversational agents' ability to respond appropriately in dynamic dialogues. Future research should also aim to establish new evaluation metrics that capture real-time conversational nuances be-

yond traditional measures of accuracy and inference speed. This approach will ensure that conversational systems are responsive and contextually adaptive in natural human–robot interactions. In summary, current evidence indicates that robust turn-taking behaviour arises not from any single modelling paradigm but from the strategic combination of complementary strengths: acoustic precision, multimodal awareness, and contextual reasoning. For researchers selecting an approach, lightweight acoustic predictors remain the most reliable foundation, while hybrid audio-linguistic or multimodal designs offer the most significant potential for future progress.

Recent discussions in the conversational AI community have emphasised the importance of recognising the ontological and ethical dimensions of turn-taking automation. As systems become increasingly predictive and socially responsive, it is essential to question what kinds of agency and interactional roles are being assigned to machines. Barbierato et al. [186] argue that machine learning represents a distinct ontological paradigm within AI, one that necessitates transparency about how algorithms interpret, reproduce, or alter human communicative norms [186]. In the context of turn-taking, this raises questions about fairness, interpretability, and the social appropriateness of automated timing decisions, particularly in sensitive domains such as healthcare, education, and assistive technologies. Addressing these concerns will be crucial for developing conversational systems that are not only technically efficient but also ethically aligned with human expectations of dialogue and agency.

Author Contributions: All authors contributed equally to the conceptualisation, methodology, investigation, and writing of this manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive grants from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable. This article is a review study that synthesises and analyses previously published literature. As such, it did not involve any primary data collection from human or animal subjects and therefore did not require ethical review or approval by an Institutional Review Board.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study is a review article/meta-analysis and did not involve the collection or analysis of original primary data. All information used was obtained from publicly available sources, as cited in the manuscript.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Moore, R.K. A comparison of the data requirements of automatic speech recognition systems and human listeners. In Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, 1–4 September 2003; pp. 2581–2584. [CrossRef]
2. Mehl, M.R.; Vazire, S.; Ramírez-Esparza, N.; Slatcher, R.B.; Pennebaker, J.W. Are Women Really More Talkative Than Men? *Science* **2007**, *317*, 82. [CrossRef] [PubMed]
3. Abreu, F.; Pika, S. Turn-taking skills in mammals: A systematic review into development and acquisition. *Front. Ecol. Evol.* **2022**, *10*, 987253. [CrossRef]
4. Cartmill, E.A. Overcoming bias in the comparison of human language and animal communication. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2218799120. [CrossRef]
5. Nguyen, T.; Zimmer, L.; Hoehl, S. Your turn, my turn. Neural synchrony in mother-infant proto-conversation. *Philos. Trans. R. Soc. Biol. Sci.* **2022**, *378*, 20210488. [CrossRef]

6. Endevelt-Shapira, Y.; Bosseler, A.N.; Mizrahi, J.C.; Meltzoff, A.N.; Kuhl, P.K. Mother-infant social and language interactions at 3 months are associated with infants' productive language development in the third year of life. *Infant Behav. Dev.* **2024**, *75*, 101929. [CrossRef]
7. Heldner, M.; Edlund, J. Pauses, gaps and overlaps in conversations. *J. Phon.* **2010**, *38*, 555–568. [CrossRef]
8. Sacks, H.; Schegloff, E.A.; Jefferson, G.D. A simplest systematics for the organization of turn-taking for conversation. *Language* **1974**, *50*, 696–735. [CrossRef]
9. Skantze, G. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Comput. Speech Lang.* **2021**, *67*, 101178. [CrossRef]
10. Levinson, S.C.; Torreira, F. Timing in turn-taking and its implications for processing models of language. *Front. Psychol.* **2015**, *6*, 731. [CrossRef] [PubMed]
11. Garrod, S.; Pickering, M.J. The use of content and timing to predict turn transitions. *Front. Psychol.* **2015**, *6*, 751. [CrossRef] [PubMed]
12. Achiam, O.J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 Technical Report . *arXiv* **2024**, arXiv:2303.08774.
13. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
14. Holler, J.; Kendrick, K.H.; Casillas, M.; Levinson, S.C. *Turn-Taking in Human Communicative Interaction*; Frontiers Media SA: Lausanne, Switzerland, 2016.
15. Levinson, S.C. Turn-taking in Human Communication—Origins and Implications for Language Processing. *Trends Cogn. Sci.* **2016**, *20*, 6–14. [CrossRef]
16. Pika, S.; Wilkinson, R.; Kendrick, K.H.; Vernes, S.C. Taking turns: Bridging the gap between human and animal communication. *Proc. R. Soc. Biol. Sci.* **2018**, *285*, 20180598. [CrossRef]
17. Ford, C.E.; Thompson, S.A. Interactional Units in Conversation: Syntactic, Intonational, and Pragmatic Resources for the Management of Turns . In *Interaction And Grammar*; Cambridge University Press: Cambridge, UK, 1996; pp. 134–184.
18. Local, J.; Walker, G. How phonetic features project more talk. *J. Int. Phon. Assoc.* **2012**, *42*, 255–280. [CrossRef]
19. Bögels, S.; Torreira, F. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *J. Phon.* **2015**, *52*, 46–57. [CrossRef]
20. Bögels, S.; Torreira, F. Turn-end Estimation in Conversational Turn-taking: The Roles of Context and Prosody. *Discourse Processes* **2021**, *58*, 903–924. [CrossRef]
21. Magyari, L.; de Ruiter, J.P. Prediction of Turn-Ends Based on Anticipation of Upcoming Words. *Front. Psychol.* **2012**, *3*, 376. [CrossRef]
22. Riest, C.; Jorschick, A.B.; de Ruiter, J.P. Anticipation in turn-taking: Mechanisms and information sources. *Front. Psychol.* **2015**, *6*, 89. [CrossRef] [PubMed]
23. Selting, M. The construction of units in conversational talk. *Lang. Soc.* **2000**, *29*, 477–517. [CrossRef]
24. de Ruiter, J.P.; Mitterer, H.; Enfield, N.J. Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language* **2006**, *82*, 515–535. [CrossRef]
25. Goodwin, C.; Heritage, J. Conversation analysis. *Annu. Rev. Anthropol.* **1990**, *19*, 283–307. [CrossRef]
26. Edmondson, W. *Spoken Discourse: A Model for Analysis*; A Longman paperback; Longman: London, UK, 1981.
27. Brady, P.T. A technique for investigating on-off patterns of speech. *Bell Syst. Tech. J.* **1965**, *44*, 1–22. [CrossRef]
28. Welford, W.; Welford, A.; Brebner, J.; Brebner, J.; Kirby, N. *Reaction Times*; Academic Press: Cambridge, MA, USA, 1980.
29. Wilson, M.; Wilson, T.P. An oscillator model of the timing of turn-taking. *Psychon. Bull. Rev.* **2005**, *12*, 957–968. [CrossRef]
30. Stivers, T.; Enfield, N.J.; Brown, P.; Englert, C.; Hayashi, M.; Heinemann, T.; Hoymann, G.; Rossano, F.; Ruiter, J.P.D.; Yoon, K.E.; et al. Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 10587–10592. [CrossRef] [PubMed]
31. Pouw, W.; Holler, J. Timing in conversation is dynamically adjusted turn by turn in dyadic telephone conversations. *Cognition* **2022**, *222*, 105015. [CrossRef] [PubMed]
32. Ferrer, L.; Shriberg, E.; Stolcke, A. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In Proceedings of the Interspeech, Denver, CO, USA, 16–20 September 2002.
33. Ward, N.G. *Prosodic Patterns in English Conversation*; Cambridge University Press: Cambridge, UK, 2019.
34. Barthel, M.; Meyer, A.S.; Levinson, S.C. Next Speakers Plan Their Turn Early and Speak after Turn-Final "Go-Signals". *Front. Psychol.* **2017**, *8*, 393. [CrossRef]
35. Pickering, M.J.; Garrod, S. An integrated theory of language production and comprehension. *Behav. Brain Sci.* **2013**, *36*, 329–347. [CrossRef]
36. Holler, J.; Kendrick, K.H. Unaddressed participants' gaze in multi-person interaction: Optimizing recipiency. *Front. Psychol.* **2015**, *6*, 98. [CrossRef]

37. Yngve, V.H. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting Chicago Linguistic Society, April 16–18. 1970*; Chicago Linguistic Society; Department of Linguistics, University of Chicago: Chicago, IL, USA, 1970; pp. 567–578.
38. Clark, H.H. *Using Language*; "Using" Linguistic Books; Cambridge University Press: Cambridge, UK, 1996.
39. Knudsen, B.; Creemers, A.; Meyer, A.S. Forgotten Little Words: How Backchannels and Particles May Facilitate Speech Planning in Conversation? *Front. Psychol.* **2020**, *11*, 593671. [CrossRef] [PubMed]
40. Coates, J. No gap, lots of overlap: Turn-taking patterns in the talk of women friends. In *Researching Language and Literacy in Social Context: A Reader; Multilingual Matters*; Multilingual Matters Ltd.: Bristol, UK, 1994; pp. 177–192.
41. Schegloff, E.A. Overlapping talk and the organization of turn-taking for conversation. *Lang. Soc.* **2000**, *29*, 1–63. [CrossRef]
42. Poesio, M.; Rieser, H. Completions, Coordination, and Alignment in Dialogue. *Dialogue Discourse* **2010**, *1*, 1–89. [CrossRef]
43. French, P.; Local, J. Turn-competitive incomings. *J. Pragmat.* **1983**, *7*, 17–38. [CrossRef]
44. Bennett, A. Interruptions and the interpretation of conversation. In Proceedings of the Annual Meeting of the Berkeley Linguistics Society, Berkeley, CA, USA, 18–20 February 1978; pp. 557–575.
45. Gravano, A.; Hirschberg, J. A Corpus-Based Study of Interruptions in Spoken Dialogue. In Proceedings of the Interspeech, Portland, OR, USA, 9–13 September 2012.
46. Heldner, M.; Hjalmarsson, A.; Edlund, J. Backchannel relevance spaces. In Proceedings of the Nordic Prosody XI, Tartu, Estonia, 15–17 August 2012; Peter Lang Publishing Group: Lausanne, Switzerland, 2013; pp. 137–146.
47. Ward, N.G. Using prosodic clues to decide when to produce back-channel utterances. In Proceedings of the Fourth International Conference on Spoken Language Processing, ICSLP '96, Philadelphia, PA, USA, 3–6 October 1996; Volume 3, pp. 1728–1731.
48. Gravano, A.; Hirschberg, J. Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* **2011**, *25*, 601–634. [CrossRef]
49. Kurata, F.; Saeki, M.; Fujie, S.; Matsuyama, Y. Multimodal Turn-Taking Model Using Visual Cues for End-of-Utterance Prediction in Spoken Dialogue Systems. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023.
50. Kendrick, K.H.; Holler, J.; Levinson, S.C. Turn-taking in human face-to-face interaction is multimodal: Gaze direction and manual gestures aid the coordination of turn transitions. *Philos. Trans. R. Soc.* **2023**, *378*, 20210473. [CrossRef]
51. Onishi, K.; Tanaka, H.; Nakamura, S. Multimodal Voice Activity Prediction: Turn-taking Events Detection in Expert-Novice Conversation. In Proceedings of the 11th International Conference on Human-Agent Interaction, Gothenburg, Sweden, 4–7 December 2023.
52. Lai, C. What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue. In Proceedings of the Interspeech, Chiba, Japan, 26–30 September 2010.
53. Gravano, A.; Benus, S.; Chavez, H.; Hirschberg, J.; Wilcox, L. On the role of context and prosody in the interpretation of 'okay'. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 23–30 June 2007.
54. Bavelas, J.B.; Coates, L.; Johnson, T. Listener Responses as a Collaborative Process: The Role of Gaze. *J. Commun.* **2002**, *52*, 566–580. [CrossRef]
55. Wang, K.; Cheung, M.M.; Zhang, Y.; Yang, C.; Chen, P.Q.; Fu, E.Y.; Ngai, G. Unveiling Subtle Cues: Backchannel Detection Using Temporal Multimodal Attention Networks. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023.
56. Johansson, M.; Skantze, G. Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction. In Proceedings of the SIGDIAL Conference, Prague, Czech Republic, 2–4 September 2015.
57. Skantze, G.; Johansson, M.; Beskow, J. Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 67–74. [CrossRef]
58. Traum, D.R.; Rickel, J. Embodied agents for multi-party dialogue in immersive virtual worlds. In Proceedings of the Adaptive Agents and Multi-Agent Systems, Bologna, Italy, 15–19 July 2002.
59. Vertegaal, R.; Slagter, R.J.; van der Veer, G.C.; Nijholt, A. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA, 31 March–5 April 2001.
60. Katzenmaier, M.; Stiefelhagen, R.; Schultz, T. Identifying the addressee in human-human-robot interactions based on head pose and speech. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA, 13–15 October 2004; pp. 144–151.
61. Ba, S.O.; Odobez, J.M. Recognizing Visual Focus of Attention From Head Pose in Natural Meetings. *IEEE Trans. Syst. Man, Cybern. Part (Cybern.)* **2009**, *39*, 16–33. [CrossRef]
62. Stiefelhagen, R.; Zhu, J. Head orientation and gaze direction in meetings. In Proceedings of the CHI '02 Extended Abstracts on Human Factors in Computing Systems, Minneapolis, MN, USA, 20–25 April 2002.
63. Anderson, A.H.; Bader, M.; Bard, E.G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.D.; Kowtko, J.C.; McAllister, J.; Miller, J.; et al. The Hcrc Map Task Corpus. *Lang. Speech* **1991**, *34*, 351–366. [CrossRef]

64. Godfrey, J.; Holliman, E.; McDaniel, J. SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of the [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, USA, 23–26 March 1992; Volume 1, pp. 517–520. [CrossRef]
65. Cieri, C.; Miller, D.; Walker, K. The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text. In Proceedings of the International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004.
66. Reece, A.; Cooney, G.; Bull, P.; Chung, C.; Dawson, B.; Fitzpatrick, C.A.; Glazer, T.; Knox, D.; Liebscher, A.; Marin, S. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Sci. Adv.* **2023**, *9*, eadf3197. [CrossRef]
67. Cafaro, A.; Wagner, J.; Baur, T.; Dermouche, S.; Torres, M.T.; Pelachaud, C.; André, E.; Valstar, M.F. The NoXi database: Multimodal recordings of mediated novice-expert interactions. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017.
68. Duncan, S.; Niederehe, G. On signalling that it's your turn to speak. *J. Exp. Soc. Psychol.* **1974**, *10*, 234–247. [CrossRef]
69. Razavi, S.Z.; Kane, B.; Schubert, L.K. Investigating Linguistic and Semantic Features for Turn-Taking Prediction in Open-Domain Human-Computer Conversation. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019.
70. Skantze, G. 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; Jokinen, K., Stede, M., DeVault, D., Louis, A., Eds.; Association for Computational Linguistics: Saarbrücken, Germany, 2017; pp. 220–230. [CrossRef]
71. Ekstedt, E.; Skantze, G. TurnGPT: A Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; Cohn, T., He, Y., Liu, Y., Eds.; pp. 2981–2990. [CrossRef]
72. da Silva Morais, E.; Damasceno, M.; Aronowitz, H.; Satt, A.; Hoory, R. Modeling Turn-Taking in Human-To-Human Spoken Dialogue Datasets Using Self-Supervised Features. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
73. Jung, D.; Cho, Y.S. Automatic Conversation Turn-Taking Segmentation in Semantic Facet. In Proceedings of the 2023 International Conference on Electronics, Information, and Communication (ICEIC), Singapore, 5–8 February 2023; pp. 1–4.
74. Tree, J.E.F. Interpreting Pauses and Ums at Turn Exchanges. *Discourse Process.* **2002**, *34*, 37–55. [CrossRef]
75. Rose, R.L. Um and uh as differential delay markers: The role of contextual factors. In Proceedings of the Disfluency in Spontaneous Speech (DiSS), the 7th Workshop on Disfluency in Spontaneous Speech, Edinburgh, UK, 8–9 August 2015; pp. 73–76.
76. Rose, R.L. Filled Pauses in Language Teaching: Why and How. Ph.D. Thesis, Waseda University, Shinjuku, Japan, 2008.
77. Clark, H.H.; Tree, J.E.F. Using uh and um in spontaneous speaking. *Cognition* **2002**, *84*, 73–111. [CrossRef] [PubMed]
78. O'connell, D.C.; Kowal, S.H. Uh and Um Revisited: Are They Interjections for Signaling Delay? *J. Psycholinguist. Res.* **2005**, *34*, 555–576. [CrossRef]
79. Kirjavainen, M.; Crible, L.; Beeching, K. Can filled pauses be represented as linguistic items? Investigating the effect of exposure on the perception and production of um. *Lang. Speech* **2021**, *65*, 263–289. [CrossRef]
80. Corley, M.; Stewart, O.W. Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Lang. Linguist. Compass* **2008**, *2*, 589–602. [CrossRef]
81. Greenwood, D.D. The Mel Scale's disqualifying bias and a consistency of pitch-difference equisections in 1956 with equal cochlear distances and equal frequency ratios. *Hear. Res.* **1997**, *103*, 199–224. [CrossRef]
82. Local, J.; Kelly, J.D.; Wells, W.H. Towards a phonology of conversation: Turn-taking in Tyneside English. *J. Linguist.* **1986**, *22*, 411–437. [CrossRef]
83. Koiso, H.; Horiuchi, Y.; Tutiya, S.; Ichikawa, A.; Den, Y. An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs. *Lang. Speech* **1998**, *41*, 295–321. [CrossRef]
84. Duncan, S. Some Signals and Rules for Taking Speaking Turns in Conversations. *J. Personal. Soc. Psychol.* **1972**, *23*, 283–292. [CrossRef]
85. Edlund, J.; Heldner, M. Exploring Prosody in Interaction Control. *Phonetica* **2005**, *62*, 215–226. [CrossRef]
86. Hjalmarsson, A. The additive effect of turn-taking cues in human and synthetic voice. *Speech Commun.* **2011**, *53*, 23–35. [CrossRef]
87. Selting, M. On the Interplay of Syntax and Prosody in the Constitution of Turn-Constructional Units and Turns in Conversation. *Pragmatics* **1996**, *6*, 371–388. [CrossRef]
88. Boersma, P. Praat: Doing Phonetics by Computer [Computer Program]. 2011. Available online: http://www.praat.org/ (accessed on 17 March 2025).
89. Kendon, A. Some functions of gaze-direction in social interaction. *Acta Psychol.* **1967**, *26 1*, 22–63. [CrossRef]
90. Jokinen, K.; Nishida, M.; Yamamoto, S. On eye-gaze and turn-taking. In Proceedings of the EGIHMI '10, Hong Kong, China, 7 February 2010.
91. Oertel, C.; Wlodarczak, M.; Edlund, J.; Wagner, P.; Gustafson, J. Gaze Patterns in Turn-Taking. In Proceedings of the Interspeech, Portland, OR, USA, 9–13 September 2012.

92. Vertegaal, R.; Weevers, I.; Sohn, C.; Cheung, C. GAZE-2: Conveying eye contact in group video conferencing using eye-controlled camera direction. In Proceedings of the International Conference on Human Factors in Computing Systems, Ft. Lauderdale, FL, USA, 5–10 April 2003.
93. Nakano, Y.I.; Nishida, T. Attentional Behaviors as Nonverbal Communicative Signals in Situated Interactions with Conversational Agents. In *Conversational Informatics: An Engineering Approach;* Wiley Online Library: Hoboken, NJ, USA, 2007; pp. 85–102.
94. Lee, J.; Marsella, S.; Traum, D.R.; Gratch, J.; Lance, B. The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In Proceedings of the International Conference on Intelligent Virtual Agents, Paris, France, 17–19 September 2007; Springer: Berlin/Heidelberg, Germany, 2007.
95. Zellers, M.; House, D.; Alexanderson, S. Prosody and hand gesture at turn boundaries in Swedish. *Speech Prosody* **2016**, *2016*, 831–835.
96. Holler, J.; Kendrick, K.H.; Levinson, S.C. Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychon. Bull. Rev.* **2017**, *25*, 1900–1908. [CrossRef] [PubMed]
97. Ter Bekke, M.; Drijvers, L.; Holler, J. The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. In *Proceedings of the Gesture and Speech in Interaction Conference*; KTH Royal Institute of Technology: Stockholm, Sweden, 2020.
98. Guntakandla, N.; Nielsen, R.D. Modelling Turn-Taking in Human Conversations. In Proceedings of the AAAI Spring Symposia, Palo Alto, CA, USA, 23–25 March 2015.
99. Johansson, M.; Hori, T.; Skantze, G.; Höthker, A.; Gustafson, J. Making Turn-Taking Decisions for an Active Listening Robot for Memory Training. In Proceedings of the International Conference on Software Reuse, Limassol, Cyprus, 5–7 June 2016.
100. Maier, A.; Hough, J.; Schlangen, D. Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017.
101. Masumura, R.; Tanaka, T.; Ando, A.; Ishii, R.; Higashinaka, R.; Aono, Y. Neural Dialogue Context Online End-of-Turn Detection. In Proceedings of the SIGDIAL Conference, Melbourne, Australia, 12–14 July 2018.
102. Lala, D.; Inoue, K.; Kawahara, T. Smooth Turn-taking by a Robot Using an Online Continuous Model to Generate Turn-taking Cues. In Proceedings of the International Conference on Multimodal Interaction, ICMI 2019, Suzhou, China, 14–18 October 2019; pp. 226–234. [CrossRef]
103. Meena, R.; Skantze, G.; Gustafson, J. Data-driven models for timing feedback responses in a Map Task dialogue system. *Comput. Speech Lang.* **2014**, *28*, 903–922. [CrossRef]
104. yiin Chang, S.; Li, B.; Sainath, T.N.; Zhang, C.; Strohman, T.; Liang, Q.; He, Y. Turn-Taking Prediction for Natural Conversational Speech. *arXiv* **2022**, arXiv:2208.13321. [CrossRef]
105. Liu, C.; Ishi, C.T.; Ishiguro, H. A Neural Turn-Taking Model without RNN. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019.
106. Aylett, M.P.; Carmantini, A.; Pidcock, C.; Nichols, E.; Gomez, R.; Siskind, S.R. Haru He's Here to Help!: A Demonstration of Implementing Comedic Rapid Turn-taking for a Social Robot. In Proceedings of the Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, 13–16 March 2023.
107. DeVault, D.; Sagae, K.; Traum, D.R. Can I Finish? Learning When to Respond to Incremental Interpretation Results in Interactive Dialogue. In Proceedings of the SIGDIAL Conference, London, UK, 11–12 September 2009.
108. Schlangen, D. From Reaction To Prediction Experiments with Computational Models of Turn-Taking. In Proceedings of the INTERSPEECH 2006—ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
109. Ward, N.G.; DeVault, D. Ten Challenges in Highly-Interactive Dialog System. In Proceedings of the AAAI Spring Symposia, Palo Alto, CA, USA, 23–25 March 2015.
110. Hariharan, R.; Häkkinen, J.; Laurila, K. Robust end-of-utterance detection for real-time speech recognition applications. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 1, pp. 249–252.
111. Atterer, M.; Baumann, T.; Schlangen, D. Towards Incremental End-of-Utterance Detection in Dialogue Systems. In *COLING 2008, 22nd International Conference On Computational Linguistics, Posters Proceedings, 18–22 August 2008*; Scott, D., Uszkoreit, H., Eds.; Coling 2008 Organizing Committee: Manchester, UK, 2008; pp. 11–14. https://aclanthology.org/C08-2003/
112. Uro, R.; Tahon, M.; Wottawa, J.; Doukhan, D.; Rilliard, A.; Laurent, A. Annotation of Transition-Relevance Places and Interruptions for the Description of Turn-Taking in Conversations in French Media Content. In Proceedings of the International Conference on Language Resources and Evaluation, Torino, Italy, 20–25 May 2024.
113. Uro, R.; Tahon, M.; Doukhan, D.; Laurent, A.; Rilliard, A. Detecting the terminality of speech-turn boundary for spoken interactions in French TV and Radio content. *arXiv* **2024**, arXiv:2406.10073. [CrossRef]

114. Threlkeld, C.; de Ruiter, J. The Duration of a Turn Cannot be Used to Predict When It Ends. In Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Edinburgh, UK, 7–9 September 2022; Lemon, O., Hakkani-Tur, D., Li, J.J., Ashrafzadeh, A., Garcia, D.H., Alikhani, M., Vandyke, D., Dušek, O., Eds.; 2022; pp. 361–367. [CrossRef]
115. Fujie, S.; Katayama, H.; Sakuma, J.; Kobayashi, T. Timing Generating Networks: Neural Network Based Precise Turn-Taking Timing Prediction in Multiparty Conversation. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 3226–3230. [CrossRef]
116. Torreira, F.; Bögels, S. Vocal reaction times to speech offsets: Implications for processing models of conversational turn-taking. *J. Phon.* **2022**, *94*, 101175. [CrossRef]
117. Threlkeld, C.; Umair, M.; de Ruiter, J. Using Transition Duration to Improve Turn-taking in Conversational Agents. In Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Edinburgh, UK, 7–9 September 2022; Lemon, O., Hakkani-Tur, D., Li, J.J., Ashrafzadeh, A., Garcia, D.H., Alikhani, M., Vandyke, D., Dušek, O., Eds.; pp. 193–203. [CrossRef]
118. Schlangen, D.; Skantze, G. A General, Abstract Model of Incremental Dialogue Processing. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, 1 March 2009; Lascarides, A., Gardent, C., Nivre, J., Eds.; pp. 710–718.
119. Skantze, G.; Hjalmarsson, A. Towards incremental speech generation in conversational systems. *Comput. Speech Lang.* **2013**, *27*, 243–262. [CrossRef]
120. Ward, N.G.; Fuentes, O.; Vega, A. Dialog prediction for a general model of turn-taking. In Proceedings of the Interspeech 2010, Chiba, Japan, 26–30 September 2010; pp. 2662–2665. [CrossRef]
121. Sakuma, J.; Fujie, S.; Zhao, H.; Kobayashi, T. Improving the response timing estimation for spoken dialogue systems by reducing the effect of speech recognition delay. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 2668–2672. [CrossRef]
122. Inoue, K.; Jiang, B.; Ekstedt, E.; Kawahara, T.; Skantze, G. 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), Torino, Italy, 20–25 May 2024; Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N., Eds.; pp. 11873–11883.
123. Ekstedt, E.; Skantze, G. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 5190–5194. [CrossRef]
124. Park, C.; Lim, Y.; suk Choi, J.; Sung, J.E. Changes in linguistic behaviors based on smart speaker task performance and pragmatic skills in multiple turn-taking interactions. *Intell. Serv. Robot.* **2021**, *14*, 357–372. [CrossRef]
125. Chen, K.; Li, Z.; Dai, S.; Zhou, W.; Chen, H. Human-to-Human Conversation Dataset for Learning Fine-Grained Turn-Taking Action. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 3231–3235. [CrossRef]
126. Inaishi, T.; Enoki, M.; Noguchi, H. A Voice Dialog System without Interfering with Human Speech Based on Turn-taking Detection. In Proceedings of the 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), Vancouver, BC, Canada, 8–12 August 2021; pp. 820–825.
127. Ekstedt, E.; Skantze, G. 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, Singapore and Online, 29–31 July 2021; Li, H., Levow, G.A., Yu, Z., Gupta, C., Sisman, B., Cai, S., Vandyke, D., Dethlefs, N., Wu, Y., Li, J.J., Eds.; pp. 431–437. [CrossRef]
128. Bîrlădeanu, A.; Minnis, H.; Vinciarelli, A. Automatic Detection of Reactive Attachment Disorder Through Turn-Taking Analysis in Clinical Child-Caregiver Sessions. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea 18–22 September 2022; pp. 1407–1410. [CrossRef]
129. O'Bryan, L.; Segarra, S.; Paoletti, J.; Zajac, S.; Beier, M.E.; Sabharwal, A.; Wettergreen, M.A.; Salas, E. Conversational Turn-taking as a Stochastic Process on Networks. In Proceedings of the 2022 56th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 31 October–2 November 2022; pp. 1243–1247.
130. Feng, S.; Xu, W.; Yao, B.; Liu, Z.; Ji, Z. Early prediction of turn-taking based on spiking neuron network to facilitate human-robot collaborative assembly. In Proceedings of the 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE), Mexico City, Mexico, 20–24 August 2022; pp. 123–129.
131. Shahverdi, P.; Tyshka, A.; Trombly, M.; Louie, W.Y.G. Learning Turn-Taking Behavior from Human Demonstrations for Social Human-Robot Interactions. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 7643–7649.
132. Ekstedt, E.; Skantze, G. How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In Proceedings of the SIGDIAL Conferences, Edinburgh, UK, 7–9 September 2022.
133. Ekstedt, E.; Skantze, G. Show & Tell: Voice Activity Projection and Turn-taking. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023.

134. Ekstedt, E.; Wang, S.; Székely, É.; Gustafson, J.; Skantze, G. Automatic Evaluation of Turn-taking Cues in Conversational Speech Synthesis. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023.
135. Sato, Y.; Chiba, Y.; Higashinaka, R. Effects of Multiple Japanese Datasets for Training Voice Activity Projection Models. In Proceedings of the 2024 27th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Hsinchu, Taiwan, 17–19 October 2024; pp. 1–6. [CrossRef]
136. Sato, Y.; Chiba, Y.; Higashinaka, R. Investigating the Language Independence of Voice Activity Projection Models through Standardization of Speech Segmentation Labels. In Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macau, China, 3–6 December 2024; pp. 1–6. [CrossRef]
137. Chen, T.; Wang, Q.; Wu, B.; Itani, M.; Eskimez, S.E.; Yoshioka, T.; Gollakota, S. Target conversation extraction: Source separation using turn-taking dynamics. *arXiv* **2024**, arXiv:2407.11277. [CrossRef]
138. Kanai, T.; Wakabayashi, Y.; Nishimura, R.; Kitaoka, N. Predicting Utterance-final Timing Considering Linguistic Features Using Wav2vec 2.0. In Proceedings of the 2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), Singapore, 28–30 September 2024; pp. 1–5. [CrossRef]
139. Umair, M.; Sarathy, V.; Ruiter, J. Large Language Models Know What To Say But Not When To Speak. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; Al-Onaizan, Y., Bansal, M., Chen, Y.N., Eds.; pp. 15503–15514. [CrossRef]
140. Wang, J.; Chen, L.; Khare, A.; Raju, A.; Dheram, P.; He, D.; Wu, M.; Stolcke, A.; Ravichandran, V. Turn-Taking and Backchannel Prediction with Acoustic and Large Language Model Fusion. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 12121–12125.
141. Jeon, H.; Guintu, F.; Sahni, R. Lla-VAP: LSTM Ensemble of Llama and VAP for Turn-Taking Prediction. *arXiv* **2024**, arXiv:2412.18061.
142. Pinto, M.J.; Belpaeme, T. Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue. In Proceedings of the 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN), Pasadena, CA, USA, 26–30 August 2024; pp. 1733–1738. [CrossRef]
143. Lucarini, V.; Grice, M.; Wehrle, S.; Cangemi, F.; Giustozzi, F.; Amorosi, S.; Rasmi, F.; Fascendini, N.; Magnani, F.; Marchesi, C.; et al. Language in interaction: Turn-taking patterns in conversations involving individuals with schizophrenia. *Psychiatry Res.* **2024**, *339*, 116102. [CrossRef] [PubMed]
144. Amer, A.Y.A.; Bhuvaneshwara, C.; Addluri, G.K.; Shaik, M.M.; Bonde, V.; Muller, P. Backchannel Detection and Agreement Estimation from Video with Transformer Networks. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; pp. 1–8.. [CrossRef]
145. Jain, V.; Leekha, M.; Shah, R.R.; Shukla, J. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Virtual, 8–13 May 2021.
146. Ortega, D.; Meyer, S.; Schweitzer, A.; Vu, N.T. Modeling Speaker-Listener Interaction for Backchannel Prediction. *arXiv* **2023**, arXiv:2304.04472. [CrossRef]
147. Inoue, K.; Lala, D.; Skantze, G.; Kawahara, T. Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection. *arXiv* **2024**, arXiv:2410.15929.
148. Sharma, G.; Stefanov, K.; Dhall, A.; Cai, J. Graph-based Group Modelling for Backchannel Detection. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022.
149. Morikawa, A.; Ishii, R.; Noto, H.; Fukayama, A.; Nakamura, T. Determining most suitable listener backchannel type for speaker's utterance. In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, 6–9 September 2022.
150. Ishii, R.; Ren, X.; Muszynski, M.; Morency, L.P. Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling? In Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, Virtual, 14–17 September 2021.
151. Onishi, T.; Azuma, N.; Kinoshita, S.; Ishii, R.; Fukayama, A.; Nakamura, T.; Miyata, A. Prediction of Various Backchannel Utterances Based on Multimodal Information. In Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents, Würzburg, Germany, 19–22 September 2023.
152. Lala, D.; Inoue, K.; Kawahara, T.; Sawada, K. Backchannel Generation Model for a Third Party Listener Agent. In Proceedings of the International Conference On Human-Agent Interaction, HAI 2022, Christchurch, New Zealand, 5–8 December 2022; pp. 114–122. [CrossRef]
153. Kim, S.; Seok, S.; Choi, J.; Lim, Y.; Kwak, S.S. Effects of Conversational Contexts and Forms of Non-lexical Backchannel on User Perception of Robots. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021, Prague, Czech Republic, 27 September–1 October 2021; pp. 3042–3047. [CrossRef]

154. Shahverdi, P.; Rousso, K.; Klotz, J.; Bakhoda, I.; Zribi, M.; Louie, W.Y.G. Emotionally Specific Backchanneling in Social Human-Robot Interaction and Human-Human Interaction. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 4059–4064. [CrossRef]
155. Seering, J.; Khadka, M.; Haghighi, N.; Yang, T.; Xi, Z.; Bernstein, M.S. Chillbot: Content Moderation in the Backchannel. *Proc. ACM Hum. Comput. Interact.* **2024**, *8*, 1–26. [CrossRef]
156. Lala, D.; Inoue, K.; Kawahara, T. Evaluation of Real-time Deep Learning Turn-taking Models for Multiple Dialogue Scenarios. In Proceedings of the 2018 On International Conference On Multimodal Interaction, ICMI 2018, Boulder, CO, USA, 16–20 October 2018; pp. 78–86. [CrossRef]
157. Bae, Y.H.; Bennett, C.C. Real-Time Multimodal Turn-taking Prediction to Enhance Cooperative Dialogue during Human-Agent Interaction. In Proceedings of the 32nd IEEE International Conference On Robot And Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, 28–31 August 2023; pp. 2037–2044. [CrossRef]
158. Inoue, K.; Jiang, B.; Ekstedt, E.; Kawahara, T.; Skantze, G. Real-time and Continuous Turn-taking Prediction Using Voice Activity Projection. *arXiv* **2024**, arXiv:2401.04868.
159. Hosseini, S.; Deng, X.; Miyake, Y.; Nozawa, T. Encouragement of Turn-Taking by Real-Time Feedback Impacts Creative Idea Generation in Dyads. *IEEE Access* **2021**, *9*, 57976–57988. [CrossRef]
160. Tian, L. Improved Gazing Transition Patterns for Predicting Turn-Taking in Multiparty Conversation. In Proceedings of the 2021 5th International Conference on Video and Image Processing, Kumamoto, Japan, July 23–25 July 2021.
161. Hadley, L.V.; Culling, J.F. Timing of head turns to upcoming talkers in triadic conversation: Evidence for prediction of turn ends and interruptions. *Front. Psychol.* **2022**, *13*, 1061582. [CrossRef] [PubMed]
162. Paetzel-Prüsmann, M.; Kennedy, J. Improving a Robot's Turn-Taking Behavior in Dynamic Multiparty Interactions. In Proceedings of the Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, 13–16 March 2023.
163. Moujahid, M.; Hastie, H.F.; Lemon, O. Multi-party Interaction with a Robot Receptionist. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Sapporo, Japan, 7–10 March 2022; pp. 927–931.
164. Iitsuka, R.; Kawaguchi, I.; Shizuki, B.; Takahashi, S. Multi-party Video Conferencing System with Gaze Cues Representation for Turn-Taking. In Proceedings of the International Conference on Collaboration Technologies and Social Computing, Virtual, 31 August–3 September 2021.
165. Wang, P.; Han, E.; Queiroz, A.C.M.; DeVeaux, C.; Bailenson, J.N. Predicting and Understanding Turn-Taking Behavior in Open-Ended Group Activities in Virtual Reality. *arXiv* **2024**, arXiv:2407.02896. [CrossRef]
166. Deadman, J.; Barker, J.L. Modelling Turn-taking in Multispeaker Parties for Realistic Data Simulation. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022.
167. Yang, J.; Wang, P.H.; Zhu, Y.; Feng, M.; Chen, M.; He, X. Gated Multimodal Fusion with Contrastive Learning for Turn-Taking Prediction in Human-Robot Dialogue. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 7–13 May 2022; pp. 7747–7751.
168. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
169. Fatan, M.; Mincato, E.; Pintzou, D.; Dimiccoli, M. 3M-Transformer: A Multi-Stage Multi-Stream Multimodal Transformer for Embodied Turn-Taking Prediction. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2023; pp. 8050–8054.
170. Lee, M.C.; Deng, Z. Online Multimodal End-of-Turn Prediction for Three-party Conversations. In Proceedings of the International Conference on Multimodal Interaction, San Jose, CA, USA, 4–8 November 2024.
171. Fauviaux, T.; Marin, L.; Parisi, M.; Schmidt, R.; Mostafaoui, G. From unimodal to multimodal dynamics of verbal and nonverbal cues during unstructured conversation. *PLoS ONE* **2024**, *19*, e0309831. [CrossRef]
172. Jiang, B.; Ekstedt, E.; Skantze, G. What makes a good pause? Investigating the turn-holding effects of fillers. *arXiv* **2023**, arXiv:2305.02101. [CrossRef]
173. Umair, M.; Mertens, J.B.; Warnke, L.; de Ruiter, J.P. Can Language Models Trained on Written Monologue Learn to Predict Spoken Dialogue? *Cogn. Sci.* **2024**, *48*, e70013. [CrossRef]
174. Yoshikawa, S. Timing Sensitive Turn-Taking in Spoken Dialogue Systems Based on User Satisfaction. In Proceedings of the 20th Workshop of Young Researchers' Roundtable on Spoken Dialogue Systems, Kyoto, Japan, 16–17 September 2024; Inoue, K., Fu, Y., Axelsson, A., Ohashi, A., Madureira, B., Zenimoto, Y., Mohapatra, B., Stricker, A., Khosla, S., Eds.; pp. 32–34.
175. Liermann, W.; Park, Y.H.; Choi, Y.S.; Lee, K. Dialogue Act-Aided Backchannel Prediction Using Multi-Task Learning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; pp. 15073–15079. [CrossRef]

176. Bilalpur, M.; Inan, M.; Zeinali, D.; Cohn, J.F.; Alikhani, M. Learning to generate context-sensitive backchannel smiles for embodied ai agents with applications in mental health dialogues. In Proceedings of the CEUR Workshop Proceedings, Enschede, The Netherlands, 15–19 July 2024; Volume 3649, p. 12.
177. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
178. LeCun, Y.; Boser, B.E.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.E.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
179. Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech, Signal Process.* **1989**, *37*, 328–339. [CrossRef]
180. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
181. Rivière, M.; Joulin, A.; Mazaré, P.E.; Dupoux, E. Unsupervised pretraining transfers well across languages. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 7414–7418. [CrossRef]
182. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
183. D'Costa, P.R.; Rowbotham, E.; Hu, X.E. What you say or how you say it? Predicting Conflict Outcomes in Real and LLM-Generated Conversations. *arXiv* **2024**, arXiv:2409.09338. [CrossRef]
184. D'efossez, A.; Mazar'e, L.; Orsini, M.; Royer, A.; P'erez, P.; J'egou, H.; Grave, E.; Zeghidour, N. Moshi: A speech-text foundation model for real-time dialogue. *arXiv* **2024**, arXiv:2410.00037.
185. Arora, S.; Lu, Z.; Chiu, C.C.; Pang, R.; Watanabe, S. Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics. *arXiv* **2025**, arXiv:2503.01174. [CrossRef]
186. Barbierato, E.; Gatti, A.; Incremona, A.; Pozzi, A.; Toti, D. Breaking Away From AI: The Ontological and Ethical Evolution of Machine Learning. *IEEE Access* **2025**, *13*, 55627–55647. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.