Predicting Turn-Taking in Child–Adult Conversations Using Voice Activity Projection

Youcef Brahimi, Cesar Blanc, and Abdellah Fourtassi
Aix Marseille Univ, CNRS, LIS, Marseille, France
Correspondence: firstname.lastname@univ-amu.fr

Abstract

Turn-taking is a hallmark of human conversation, yet its developmental trajectory remains poorly understood. Adults typically respond within a few hundred milliseconds, suggesting reliance on predictive cues rather than simply waiting for silence. In contrast, children's longer gaps raise the question of whether they depend on simpler, reactive strategies. This study provides the first large-scale test of competing hypotheses about children's turn-taking, using corpora of child–adult and adult–adult dialogues. In Study 1, we compared a simple silence-based threshold model with the Voice Activity Projection (VAP) model, which predicts upcoming speech activity from acoustic features. Results showed that silence alone could not account for children's behavior, whereas predictive acoustic models performed well, indicating that even early turn-taking relies on anticipatory mechanisms. In Study 2, we asked what cues support these predictions by comparing models based on acoustic features alone with models combining acoustic and lexical information. For adult conversations, lexical cues improved prediction, but for child–adult dialogues, acoustic information was sufficient to solve the task. Together, these findings suggest that children's turn-taking is predictive but primarily grounded in acoustic patterns, revealing both continuity with adult mechanisms and developmental differences in how linguistic cues are integrated.

1 Introduction

Turn-taking is the ability to coordinate turns in a conversation, avoiding both excessively long pauses and inappropriate interruptions. It is argued to be one of the defining features of human social behavior and a driver for knowledge transmission and learning more generally (Levinson, 2025; Clark, 2022).

Understanding the way turn-taking develops in childhood is, therefore, of utmost importance for theories of socio-cognitive development—with implications ranging from heath (e.g., better understanding and mitigating atypical social behavior in autism) to education (e.g., interactive curricula that accommodate children's age-appropriate conversational skills), through child-oriented dialog systems (e.g., design of theory-informed e-tutoring).

Turn-taking is one of the most challenging abilities to learn in childhood; having a protracted developmental trajectory, beyond early childhood (Casillas et al., 2016). This may come as a surprise, especially in light of its very simple implementation in traditional spoken dialog systems: Turn shifts were simply cued by *silence*—signaling the end of a turn and yielding the floor for the interlocutor. We refer to this as the silence-based account (Skantze, 2021).

While a silence-based model can be good enough in some human-computer applications, it is inadequate as a *scientific* account for the human-human natural turn-taking (e.g., Sacks et al., 1974). In particular, analysis of dialog across many cultures shows that the duration of silence in turn shifts (hereafter *gaps*) is too small; the median is generally below 200ms and drops to near 0ms in some cultures (Stivers et al., 2009). When compared to the time it takes humans to plan and produce a response—around 600ms for a simple one-word utterance (Levelt, 1993)—it is unlikely that adults wait for silence to start planning a response; Instead, they most likely rely on turn-taking (vs. yielding) cues to *predict* the end of the turn and start planning the response *before* the silence (Levinson, 2016)—hereafter the *predictive* account.[1]

Thus, adults turn-taking cannot be captured by

[1] In using the contrast silence-based vs. predictive, we take inspiration from dialog system literature (Skantze, 2021). However, we do not claim our use of this contrast fully reflects the nuances made in that line of work nor that it contributes to it. Here, we define and use the terms in a way that is more directly relevant to our goal: the characterization of children's development.

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 338–347
February 26–March 1, 2026.

a silence-based account and requires a more sophisticated predictive one. What about children? Some studies have shown that preschoolers can predict end of turns in simple cases (Casillas and Frank, 2017; Lindsay et al., 2019). However, this is typically observed in controlled settings and/or where children watch a conversation they are not part of. In contrast, when children are observed in naturalistic, participatory settings—especially in child–adult conversations—their response latencies are substantially longer than those of adults. Typical estimates are around 1 second, persisting up to 5 or 6 years of age and continuing to be refined throughout middle childhood. (Casillas et al., 2016; Nguyen et al., 2022; Levinson, 2025).

Larger response latencies in children can, in principle, be reconciled with both the silence-based and the predictive accounts. Under the former, a gap of around 1 second does not contradict the above-mentioned 600 ms minimum required for speech production (Levelt, 1993). This would suggest that children may not need to rely on prediction, but could instead wait for the interlocutor's turn to end before initiating response planning. Under the latter account, however, although children's gaps are longer than those of adults, their speech production processes are also slower (e.g., due to immature articulatory control and memory retrieval mechanisms; (Clark and Lindsey, 2015)). As a result, longer gaps do not necessarily imply the absence of predictive mechanisms, and children may still need to anticipate turn endings in order to respond in a timely–albeit slower—manner.

The current study

We address two main questions:

Question 1: Silence-based vs. Predictive accounts While the silence-based model is clearly inadequate for characterizing adults conversations; this is not *a priori* the case for child-adult conversation. This presents us with two alternative developmental hypotheses:

1. **Silence-based account** Children initially rely on a silence-based model and only later switch to a predictive model—e.g., as the requirement for fast responses becomes more socially pressing.

2. **Predictive account** Children initially rely on a predictive model—albeit a slower one than in adults. The prediction grows to match adults' speed as language processing and planning skills mature.

Question 2: Acoustic-only vs. Multimodal cues in the predictive account While the silence-based model can be directly tested using standard, non-parametric measures in signal detection theory, the predictive account is trickier: testing it requires specifying what predictive cues we are considering. Research on adult-adult conversions have proposed many such cues, including in acoustic modality such as rising or falling pitch (e.g., Bögels and Torreira, 2015), verbal modality such as lexical, syntactic or semantic cues (e.g., De Ruiter et al., 2006), and visual modality, such as gaze aversion (e.g., Kendon, 1967).

In the case of children, data is scarce, but recent modeling studies have pointed to the primary role of acoustic cues compared to other modalities (e.g., Agrawal et al., 2023; Liu et al., 2022) Following this, and to the extent to children rely on a predictive model, we have the following developmental hypotheses:

1. **Acoustic-only account** Children initially rely primarily on acoustic cues to anticipate turn endings. Only at later developmental stages do they progressively integrate cues from other modalities.

2. **Multimodal account** From early on, children draw on a combination of acoustic and non-acoustic (e.g., lexical, visual) cues to coordinate turn-taking.

Research strategy and predictions We investigate these two questions and test their predictions on a large corpus of child-adult conversations (Ohio Child Speech Corpus) and adult-adult conversations (Switchboard corpus). We capitalize on advances in recent, self-supervised techniques in turn-taking modeling, namely the Voice Activity Projection Modeling framework (Ekstedt and Skantze, 2022b)—going beyond many pioneering modeling studies of early conversational skills (Liu et al., 2022; Park et al., 2017), which —though insightful—were limited by their small scale and coverage, due to methods requiring laborious cue extraction and manual annotation.

We use a task that is common in this line of modeling—and which is specifically diagnostic in our case. Broadly speaking (see details in Methods below), we test if, during a moment of mutual

silence, the end of turn can be successfully determined, indicating either a turn shift (hereafter SHIFT) or a mere pause within the same turn, indicating the speaker is intending to hold the floor (hereafter HOLD).

For **question 1**, the silence-based account predicts there to be threshold in the silence duration that successfully separates SHIFT form HOLD. This should be found in child-adult conversations but not in adult-adult conversions. In the predictive account, such threshold does not exist; but predictive cues *preceding* the silence can successfully separate SHIFT form HOLD. This should be the case in both child-adult conversation and in adult-adult conversation.

As for **question 2**—and to the extent that a predictive model proves necessary in Question 1—the Acoustic-only account predicts that acoustic cues in the speaker's speech are sufficient to separate SHIFT form HOLD in child-adult conversations, but not in adult-adult conversations. In the multimodal account, information from the acoustic modality are insufficient; cues from other modalities are necessary to successfully solve the task in both child-adult conversations and adult-adult conversations

The paper is organized in two studies, addressing our two research questions. We end with a discussion of the results, their impact, and their limitations.

2 Study 1: Silence-based vs. Predictive accounts

First we present the datasets and their properties. Then we explain the Methods. Next, we present the results of the main analyses: 1) Testing the silence-based account using methods in signal detection theory, 2) Testing the predictive account using the Voice Activity Projection model.

2.1 Datasets

We contrast two English-language spoken conversational datasets of child-adult conversations (Ohio Child Speech Corpus) and of adult-adult conversations (Switchboard).

2.1.1 Ohio Child Speech Corpus (OCSC)

The OCSC corpus (Wagner et al., 2025) is a collection of dialogues involving N=303 children, all of whom participated in a seven-task elicitation protocol conducted in by an adult researcher in a science museum lab. The corpus is made up of 303 dialogues with children ranging in age from 4 to 9 years old. The length of the corpus is approximately 148 hours of audio. Children are mostly monolingual English speakers (91%), come from a highly educated background (79% had at least one parent who had earned a Bachelor's degree), and approximately half are female (54%). Descriptive statistics by age groups are shown in Table 1. The corpus is publicly available on TalkBank.[2]

2.1.2 Switchboard

The Switchboard dataset is a compilation of English telephone conversations. It consists of 2,438 different dyadic dialogues, totaling around 260 hours of data. These dialogues were produced by 543 unique speakers (302 male and 241 female) from across the United States, each taking part in multiple calls. Each conversation lasts on average 6–10 minutes and covers a wide range of everyday topics prompted by pre-assigned themes.

2.2 Methods

2.2.1 Speech segments

We define a speech segment as a stretch of audio from one speaker. A turn can be made of one segment —followed by a turn shift, or it can be made of several segments, separated with pause. We use Voice Activity Detection (VAD) for automation, and in particular Silero.[3]

2.2.2 SHIFT-HOLD Task

We use a binary task that classifies moment of silence as either a *gap* between two turns (i.e., indicating a turn SHIFT) or a *pause* within the same turn (i.e., indicating a turn HOLD). Following previous work (Inoue et al., 2024), the task is performed on events that have the following criteria:[4] a) There is a mutual silence of at least 250ms, b) this silence should be preceded by a speech segment of at least one second, and c) it should also be followed by another speech segment of at least one second, whether from the same speaker (HOLD) or the interlocutor (SHIFT).

In the case of child-adult conversations in the OCSC corpus, and in addition to the overall task, we also report outcome for two special cases, corresponding to who is speaking before the silence,

[2] https://talkbank.org/childes/access/Eng-NA/OCSC.html

[3] https://github.com/snakers4/silero-vad

[4] These are supposed to remove noise and filter out more granular events that are not typically considered a turn switch, such as short backchannels.

Age group	# Children	Average Session length (min)	Child-initiated event		Adult-initiated event	
			Hold	Shift	Hold	Shift
4 years old	26	25.1	1164	838	1905	828
5 years old	54	28.8	5115	2327	3531	2331
6 years old	60	30.6	7237	2522	3411	2466
7 years old	63	32.9	9560	2189	2901	2197
8 years old	57	32.4	8090	1861	2046	1902
9 years old	43	32.6	7187	1246	1530	1236
All children	303	30.9	38353	10983	15324	10960

Table 1: Descriptive statistics per age group in OCSC.

i.e., adult-initiated vs. child-initiated events.[5] After applying these criteria, we end up, in the case of OCSC, with the numbers shown in Table 1 (including a breakdown by age group), and, in the case of switchboard, with a total of 7490 shifts and 71096 holds (we will comment on these numbers in the Results).

2.2.3 Voice Activity Projection (VAP)

While the silence-based account can be evaluated using simple, non-parametric measures from signal detection theory (e.g., the Area Under the ROC Curve), properly testing the predictive account requires more sophisticated machine-learning methods.

More specifically, we use the Voice Activity Projection (VAP) modeling framework; introduced by (Ekstedt and Skantze, 2022b) and commonly used in recent research on turn-taking in the speech technology literature (e.g., Ekstedt and Skantze, 2022a; Inoue et al., 2024; Russell and Harte, 2025). We use, as a starting point for our modeling the instance—as well as the implementation—by Inoue et al. (2024). For specific details regarding the VAP model, we refer the readers to the above-mentioned papers. In the following, we provide only a high-level description.

In broad terms, the model uses a self-supervised predictive objective; it takes as input the previous voice activity in the conversation and outputs predictions regarding the next pattern of (binary) voice activity (i.e., who will be speaking when). The intermediate pipeline is as follows: For each speaker/channel, the input Voice Activity **(VA)** is encoded using a pre-trained speech model. In the instance we use, it is a pre-trained Contrastive Predictive Coding model **(CPC)**. This encoding is then fed to a Self-Attention **(SA)** layer, meant to capture longer temporal cues. The outputs from the two channels are fed to a Cross-Attention **(CA)** mechanism, meant to capture potential interactive information between speakers. Finally, the output of cross-attention is passed to linear layers to predict the next Voice Activity. Note that the final output is a binary classification into (future) speech vs. silence, not a generation of actual speech—unlike spoken generative models (e.g., dGSLM, (Nguyen et al., 2023).

2.2.4 Zero-shot testing of VAP

The model is not explicitly taught to distinguish cases of pauses vs. gaps from the task events described in subsection 2.2.2. The model is trained in a fully self-supervised fashion (as described above), learning to predict the binary patterns of speech and silence. The SHIFT-HOLD task is given to the model after training convergence, in a *zero-shot* style. The SHIFT-HOLD task is built from a subset of the data that the model has not seen during training (more details on this in the Results).

Crucially, the model's prediction for the task are derived from cues preceding the silence—i.e., silence duration information is not used. This testing approach is identical to original studies (Ekstedt and Skantze, 2022a; Inoue et al., 2024) and we refer the readers to these papers for further details.

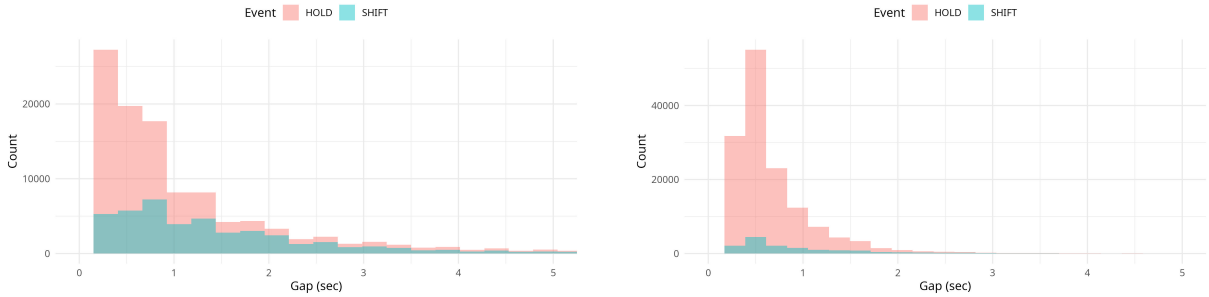
2.3 Results

2.3.1 The silence-based account

Remember that the silence-based account predicts silence information alone to successfully solve the SHIFT-HOLD task in child-adult conversation—but not in adult-adult conversations. Using the conversational events described in Subsection 2.2.2, we test if silence duration can classify them into gaps vs. pauses.

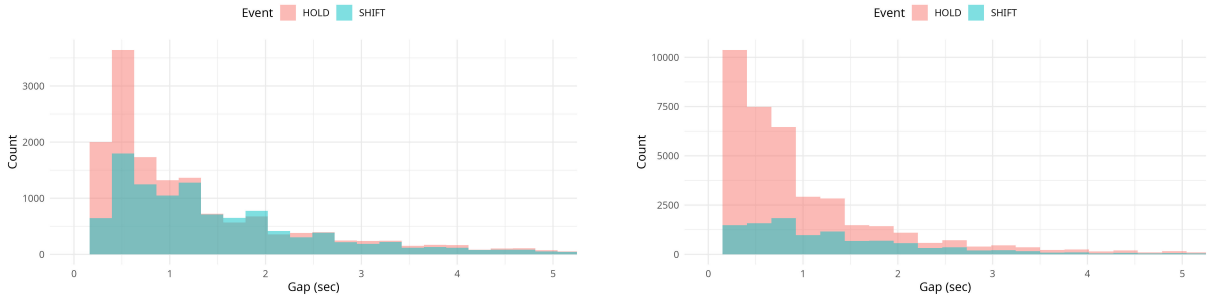
Figure 1 shows the frequency distribution of gap durations vs. pause durations. First, we note that

[5] Though this distinction does not mean we can test children and adults in isolation—see Limitations in the Discussion.



(a) OCSC (b) Switchboard

Figure 1: Distribution of turn–HOLD & turn-SHIFT durations.



(a) Adult (b) Child

Figure 2: Distribution of turn–HOLD & turn-SHIFT duration by role in OCSC (Adult-initiated vs Child-initiated events).

the numbers of holds is generally larger than the number of shifts. This fact is more pronounced in Switchboard (9.5% of shift, compared to 29.0% in OCSC). This imbalance, especially in Switchboard, is unlikely to be an artifact of our speech segmentation, since it was also noted in previous studies using different data pre-processing methods (e.g., 9.9% in Inoue et al. (2024)).

The same Figure 1 shows extensive overlaps between the distributions of HOLD and SHIFT. This fact suggests the duration of silence is unlikely to separate these two categories: pauses can be longer than gaps (and vice versa), not only in adult-adult conversations (which we already knew), but *also* in child-adult conversation, which is what wanted to test. This observation was confirmed quantitatively using the Area Under the Roc Curve (AUC), summarizing the classification performance under various duration thresholds: we obtain AUC = 0.62 for OCSC and AUC = 0.63 for Switchboard, indicating poor performance. The performance was equally low in both the case of child- and adult-initiated events within OCSC, with values of AUC= 0.62 and AUC = 0.59, respectively (see Also Figure 2).

2.3.2 The predictive account

Remember that, according to the predictive account, silence information cannot solve the SHIFT-HOLD task in child-adult conversation. Rather, it is the information that *precedes* the silence that solve the task in *both* child-adult and adult-adult conversations. As mentioned above, we tested this hypothesis with the VAP model on both corpora, using the zero-shot approach on the SHIFT-HOLD task, as described in Subsection 2.2.4.

The results are reported for SHIFT-HOLD examples that were built from a testing set in both corpora. In switchboard, and, to replicate Inoue et al. (2024), we used the exact test set made of 10% of the data (the rest was used for VAP training and validation). As for OCSC–a smaller corpus– we used 20% of the data to reach a comparable sample size (the rest was used for VAP training and validation). The test set of OCSC was balanced for age group, but—to avoid data contamination—it contained **different** participants than the ones seen by the model in training.

We report results averaged across predictions from 3 models trained with different random seeds.

For switchboard, we obtained an F1 score of 69.71 and a balanced accuracy of 80.62; replicating results from Inoue et al. (2024): F1 score 70.11 and balanced accuracy 81.02. For OCSC, we obtained high overall score (higher than in Switchboard), with an F-score of 90.28 and balanced accuracy of 93.96. When breaking down events in the test set by speaker, child-initiated events led to F1 score and balanced accuracy of 88.95 and 94.14, respectively and adult–initiated events led to 91.77 and 93.1 . When further breaking down the test data into different age group, we found no noticeable developmental patterns whether in child-initiated or adult–initiated events (numbers not shown).

3 Study 2: Acoustic-only vs. Multimodal accounts

Results of Study 1 reject the silence-based account and strongly corroborate the predictive accounts. The VAP model, which makes prediction based on cues occurring before the silence, provides a much better model of *both* child-adult and adult-adult conversations.

The goal of Study 2 is to follow-up on Study 1, zooming in on the nature of the cues that help in making accurate turn-taking prediction in child-adult vs. adult-adult conversations. As we mentioned in the Introduction, we specifically test an Acoustic-only account vs. the Multimodal account of development. This will be addressed with two analyses.

In the first, we use an ablation analysis to test if the acoustic cues are *necessary* for explaining the performance of the VAP model in Study 1. In the second analysis, we test if the acoustic cues are *sufficient*: we integrate information from the verbal modality and test if it provides additional, non-redundant information.

3.1 Ablation analysis

To test if the acoustic cues (represented by the CPC encoder) are necessary, we perform an ablation analysis on the VAP model. We test both variants that keep the CPC and remove other modules and variants that do the opposite (see the VAP's modules in Subsection 2.2.3). All ablated variants were trained and evaluated under the same conditions as the original model in Study 1.

The results of the ablation analysis are shown in Table 2. When removing cross-attention (CA) and self-attention (SA) we see no noticeable changes with OCSC and only small drops in Switchboard. However, in both corpora, the removal of CPC encoder had the largest impact. In fact, performance on both corpora drops to chance level in the absence of CPC—i.e., when the model only sees binary input data from VA. The model does not recover when the attention mechanism were added back (VA + SA).

Model Variant	Switchboard		OCSC	
	F1	BAcc	F1	BAcc
Original	69.71	80.62	90.28	93.96
VA + CPC + SA	66.55	78.29	90.08	93.76
VA + CPC	56.96	76.23	89.09	93.21
VA + SA	17.00	45.42	36.90	53.35
VA only	17.00	49.39	36.90	46.65

Table 2: Ablation study results on the Shift/Hold prediction task. Metrics are F1-score and balanced accuracy (BAcc).

This analysis highlights the essential role of fine-grained acoustic cues—as captured by the CPC representation—in turn-taking prediction. It also rules out an alternative explanation according to which turn-taking could be accounted for solely by coarse temporal patterns of speech and silence (i.e., VA + SA). Instead, the results show that explicit access to acoustic information is required, including in child–adult conversations.

3.2 Multimodal integration

In addition to acoustic cues in the CPC encoder, we integrate lexical information from the verbal modality. This required expanding the original VAP architecture (Inoue et al., 2024; Ekstedt and Skantze, 2022b). We test if adding lexical information provides additional, non-redundant information.

3.2.1 Methods

To represent cues from the verbal modality (i.e., text), we use pre-trained BERT (Devlin et al., 2019), providing contextual embeddings from dialogue transcripts. For Switchboard, manual transcripts with word-level timestamps were already available. As for OCSC, we used WhisperX (Bain et al., 2023) for automatic speech recognition and alignment, generating word-level transcripts with timestamps.

Integrating verbal and acoustic cues is not straightforward because they operate on different

time scales. Acoustic cues unfold over relatively short intervals, and the VAP model samples this information at a frame rate of 50 Hz (i.e., every 20 ms). In contrast, verbal cues at the word level unfold over longer time scales. To integrate the two modalities, we adopted a simple alignment strategy: all frames falling within the temporal span of a given word were assigned the same BERT embedding for that word.

Next, we used an early fusion approach where BERT embeddings were concatenated with the CPC audio representations, frame by frame, before being passed on to the self-attention layer.

3.2.2 Results

The results of the multimodal integration are shown in Table 3. There was a substantial increase in performance on Switchboard. However, there were no noticeable improvements on OCSC.

Model	Switchboard		OCSC	
	F1	BAcc	F1	BAcc
VAP (Audio only)	70.20	80.99	**90.38**	**94.03**
VAP + Text	**87.10**	**92.14**	90.25	93.78

Table 3: Impact of adding textual modality on VAP's performance for the Shift/Hold prediction task.

Data size vs. Modality Could the benefit of additional cues interact with the amount of training data available? In other words, might multimodal integration prove more beneficial when data are scarce, allowing the model to compensate for limited input by leveraging both modalities more effectively? Figure 3 show the F-scores, when the models are trained on different percentages of the original datasets.

The results indicate that model performance was not strongly affected by data scarcity. In OCSC, no multimodal benefit was observed. In Switchboard, the multimodal effect remained robust, but it did not confer a greater advantage in the low-data regime.

4 Discussion

Our starting point was the developmental question of how children manage conversational turn-taking. Adults are known to take turns with extremely short gaps, suggesting that they anticipate upcoming endings rather than simply reacting to silence (Levinson, 2016). In contrast, children's gaps are longer

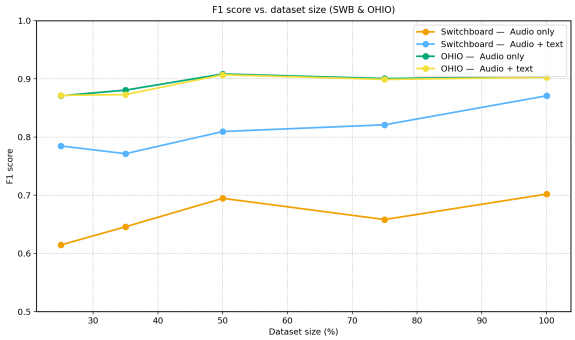


Figure 3: F1-score vs. dataset size on Switchboard and OCSC (audio-only vs. audio+text).

(Casillas et al., 2016; Nguyen et al., 2022), raising the possibility that they rely on a simpler, silence-based strategy in which pauses serve as the primary signal.

From this contrast, two competing hypotheses emerge. If children mainly depend on silence, then models that predict turn transitions from pause duration alone should perform relatively well on child–adult dialogues, and predictive models should offer little advantage. If, however, children—like adults—make use of anticipatory cues, then predictive models should outperform silence-based accounts even in child data.

Study 1 tested this by comparing a simple silence-based threshold model with the Voice Activity Projection (VAP) model, which predicts future speech activity from acoustic features. The results clearly rejected the silence-based model and supported the predictive account: the threshold model struggled to tell shifts from holds, whereas VAP, using acoustic cues occurring *before* the silence, achieved much higher accuracy. This suggests that, although children's turn gaps are longer than adults', they are unlikely to rely on silence to determine when a turn has ended. Silence is an unreliable cue, as it can equally signal a within-turn pause or a turn completion. Instead, children likely draw on anticipatory information in the speech signal (and potentially other multimodal cues; see Limitations), in a manner broadly comparable to adults.

Study 2 was a natural follow-up to the first, moving beyond the question of whether turn-taking is predictive to ask what kinds of cues support this prediction. For adults, prior work suggests that both acoustic and lexical information contribute to anticipating turns (De Ruiter et al., 2006; Bögels and Torreira, 2015), raising the question of whether

child-adult conversations require the same multimodal resources. To test this, we compared models that used only acoustic cues with models that also incorporated text embeddings—using time-aligned dialog transcriptions. The results confirmed the expected multimodal benefit in adult–adult dialogues, where adding lexical cues improved predictions. In contrast, in child–adult conversations, verbal information contributed little beyond the acoustic channel; indeed, performance was already near ceiling with acoustic cues alone, leaving minimal room for improvement with additional signals.

Taken together, the two studies suggest that children's turn-taking patterns are neither purely reactive, wait-for-silence strategies nor fully comparable to adults' use of multimodal cues. Rather, they reflect an intermediate developmental stage.

On the one hand, there appears to be continuity with adults in the use of *predictive* mechanisms for identifying turn endings. In this respect, the present study provides naturalistic corroboration of previous experimental findings, showing that preschoolers begin planning their responses as early as possible; rather than waiting for the interlocutor's turn to fully end (Lindsay et al., 2019).

On the other hand, the findings point to a developmental change in the *composition* of this anticipatory process, specifically in the cues required for accurate prediction. Whereas adult–adult conversations benefit from the integration of signals across multiple modalities to optimally anticipate turn endings, child–adult conversations appear to be less ambiguous, with acoustic cues alone largely sufficient to identify turn ending in most cases. These results provide large-scale corroboration of earlier, small-scale studies, which also highlighted the central role of acoustic information in predicting turns and backchannels in child–caregiver multimodal interactions.(Agrawal et al., 2023; Liu et al., 2022).

Limitations

While this study offers novel insights into children's turn-taking, it represents only an initial step. There are several limitations to consider, some of which point to important directions for further research.

First, our analyses are based on observational, correlational data rather than direct testing of children's processing. What we capture are the surface patterns of turn-taking as they unfold in interaction, from which we infer which mechanisms are more or less plausible. This allows us to constrain theories: for instance, if silence alone cannot account for the observed coordination, then a purely reactive account is unlikely. At the same time, such analyses cannot tell us definitively what children do or do not represent internally or how they actually process and plan their turns.

Furthermore, our modeling approach cannot fully disentangle the child's contribution from the adult's. Turn-taking is inherently dyadic: it depends both on the speaker providing clear turn-yielding cues *and* on the listener being able to pick them up. Our models therefore capture reliable *coordination* patterns. At the same time, this also gives us confidence that the child is an active participant in the coordination—which cannot be orchestrated by the adult alone.

One important limitation of the current study is that our child–adult and adult–adult data differ not only in developmental stage but also in conversational setting: the OCSC corpus is based on structured child-adult tasks, whereas Switchboard captures more spontaneous telephone conversations. This raises the possibility that some of the observed contrasts reflect task and context differences in addition to age. A next step would be to test whether the same patterns hold in corpora that better align in terms of conversational context. That said, developmental conversational resources that span the relevant age range and provide sufficient data for modeling remain very scarce (see Goumri et al., 2024).

Finally, our analyses were limited to auditory and verbal modalities because this is what the available child corpora provide. In natural face-to-face interaction, visual signals such as gaze, gesture, and posture can be helpful for regulating turns in adults (Holler and Levinson, 2019; Kendrick et al., 2023; Russell and Harte, 2025), and it is possible that children also draw on these cues. By focusing only on speech and text, our study captures an essential part of the coordination process but not its *full* multimodal basis.

To conclude, while the study has limitations that call for further work, it is worth emphasizing the broader impact. To our knowledge, this is one of the first studies to test competing hypotheses about children's turn-taking at scale, using large conversational corpora and state-of-the-art predictive models. This was made possible by an interdisciplinary approach that builds on the sustained efforts of the spoken dialogue systems community, applied here to research questions in developmental research.

5 Acknowledgment

This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix Marseille University (A*MIDEX). Furthermore, this study was also supported by the ANR MACOMIC (ANR-21-CE28-0005-01) grant.

References

Abhishek Agrawal, Jing Liu, Kübra Bodur, Benoit Favre, and Abdellah Fourtassi. 2023. Development of Multimodal Turn Coordination in Conversations: Evidence for Adult-like behavior in Middle Childhood. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45).

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Sara Bögels and Francisco Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Marisa Casillas, Susan C. Bobb, and Eve V. Clark. 2016. Turn-taking, timing, and planning in early language acquisition. *Journal of Child Language*, 43(6):1310–1337.

Marisa Casillas and Michael C. Frank. 2017. The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language*, 92:234–253. Place: Netherlands Publisher: Elsevier Science.

Eve V. Clark. 2022. Language is Acquired in Interaction. In *Algebraic Structures in Natural Language*. CRC Press. Num Pages: 18.

Eve V Clark and Kate L Lindsey. 2015. Turn-taking: A case study of early gesture and word use in answering where and which questions. *Frontiers in psychology*, 6:890.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Erik Ekstedt and Gabriel Skantze. 2022a. How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551, Edinburgh, UK. Association for Computational Linguistics.

Erik Ekstedt and Gabriel Skantze. 2022b. Voice activity projection: Self-supervised learning of turn-taking events. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*, pages 5190–5194. : International Speech Communication Association.

Dhia Elhak Goumri, Abhishek Agrawal, Mitja Nikolaus, Hong Duc Thang Vu, Kübra Bodur, Elias Emmar, Cassandre Armand, Chiara Mazzocconi, Shreejata Gupta, Laurent Prévot, Benoit Favre, Leonor Becerra-Bonache, and Abdellah Fourtassi. 2024. CHICA: A Developmental Corpus of Child-Caregiver's Face-to-face vs. Video Call Conversations in Middle Childhood. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3153–3164, Torino, Italia. ELRA and ICCL.

Judith Holler and Stephen C. Levinson. 2019. Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of LREC-COLING*.

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

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

Willem JM Levelt. 1993. *Speaking: From intention to articulation*. MIT press.

Stephen C. Levinson. 2016. Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, 20(1):6–14.

Stephen C. Levinson. 2025. *The Interaction Engine: Language in Social Life and Human Evolution*. Cambridge University Press, Cambridge.

Laura Lindsay, Chiara Gambi, and Hugh Rabagliati. 2019. Preschoolers optimize the timing of their conversational turns through flexible coordination of language comprehension and production. *Psychological science*, 30(4):504–515.

Jing Liu, Mitja Nikolaus, Kübra Bodur, and Abdellah Fourtassi. 2022. Predicting Backchannel Signaling in Child-Caregiver Multimodal Conversations. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ICMI '22 Companion, pages 196–200, New York, NY, USA. Association for Computing Machinery.

Phu-An Nguyen, Jing Zhang, Jaehong Hyun, Yoshua Bengio, and Mathieu Riviere. 2023. dgslm: A generative spoken language model with discrete latent representations. In *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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

Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 100–108, New York, NY, USA. Association for Computing Machinery.

Sam O'Connor Russell and Naomi Harte. 2025. Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 209–221, Vienna, Austria. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735. Publisher: Linguistic Society of America.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

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

Laura Wagner, Sharifa Alghowinhem, Abeer Alwan, Kristina Bowdrie, Cynthia Breazeal, Cynthia G. Clopper, Eric Fosler-Lussier, Izabela A. Jamsek, Devan Lander, Rajiv Ramnath, and Jory Ross. 2025. The ohio child speech corpus. *Speech Communication*, 170:103206.