Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue

Maria J. Pinto[1]. Tony Belpaeme[1]

***Abstract*— Natural and engaging spoken dialogue systems require seamless turn-taking coordination to avoid awkward interruptions and unnatural pauses. Traditional systems often rely on simplistic silence thresholds, relinquishing the turn after a predetermined period of silence, which invariably leads to a suboptimal interaction experience. This work explores the potential of Large Language Models (LLMs) for improved turn-taking prediction. Building upon research that uses linguistic cues, we investigate how LLMs, with their rich contextual knowledge and semantic encoding of language, can be used for this task. We hypothesize that by analysing dialogue context, syntactic structure, and pragmatic cues within the user's utterance, LLMs can offer more accurate turn-completion predictions. This research evaluates the capabilities of recent LLMs such as Gemini, OpenAI's API, Anthropic's Claude2, and Meta AI's Llama 2 to predict turn-ending points solely based on textual information, and demonstrates how the conversation between elderly users and companion robots can be enhanced by LLM-powered end-of-turn prediction.**

I. INTRODUCTION

One of the cornerstones of a natural and engaging spoken dialogue system lies not just in what it says, but also in **when** it says it. This presents a significant challenge, as systems need to seamlessly coordinate their speaking turns with the user to avoid awkward interruptions, unnatural pauses, or overlapping speech. Traditionally, turn-taking models in these systems have relied on simple indicators like silence thresholds to determine when the turn is complete. However, this approach often leads to interruptions or delayed responses, depending on the specific threshold setting [1], [2], [3].

For a comprehensive understanding of turn-taking dynamics, a broader range of aspects needs to be incorporated into the modelling process. Ideally, the system should be able to discern whether the user's brief pause indicates a desire to continue speaking or if the turn is relinquished [1], [4]. Recognising this need, recent studies have explored more sophisticated approaches that leverage machine learning techniques to capitalise on the nuances of turn-taking cues, including not only prosody but also linguistic cues, such as syntax, semantics, and pragmatics [1], [5], [6], [7]. Although several of these studies have found that linguistic information contributes significantly to turn-taking performance, the results of integrating linguistic cues into turn-taking prediction mechanisms are perhaps not as substantial as what could initially be expected. As Skantze [3] highlights, one possible explanation could be the lack of proper modelling of the dialogue context.

[1] AIRO–IDLab, Ghent University—imec, 9052 Ghent, Belgium MariaJose.PintoBernal@UGent.be

An alternative approach proposed by Ekstedt and Skantze [8] utilises a transformer-based language model for turn-taking prediction. This model, built upon on Open AI's GPT-2 and fine-tuned on various dialog datasets, aims to predict potential turn-completion points in dialog, based on linguistic features (i.e., words) alone.

Building upon this foundation, our research investigates the capabilities of recent large language models (LLMs) such as Gemini, OpenAI's API, Anthropic's Claude2, and Meta AI's Llama 2 to predict or detect the end of a user's turn. Like the model presented by Skantze [8], these models have shown impressive accuracy in various prediction tasks. However, we extend their work by capitalising on the deep contextual knowledge and rich linguistic "understanding" embedded within LLMs to achieve significantly enhanced turn-taking prediction accuracy in conversational systems.

This initial evaluation focuses on the performance of an LLM-based turn-taking prediction model that considers linguistic features such as syntax, semantics, and dialogue context. This allows us to establish the capability of LLMs in understanding turn completion cues solely through language analysis.

We then explore the synergistic effect of combining LLMs with Voice Activity Detection (VAD) in a more realistic setting. By incorporating VAD information which captures the user's speech activity and pauses of varying lengths, we investigate how the model's performance changes. This combined approach aims to achieve highly accurate turn-taking prediction, leading to smoother and more natural conversations in spoken dialogue applications.

II. BACKGROUND

Turn-taking in spoken dialogue systems has been a subject of extensive study, with early seminal work such as Sacks et al. [9] laying the foundation for understanding the dynamics of conversation. Their model emphasises the flexible coordination of dialogue and the role of "turn-constructional units" (TCUs) and "transition-relevant places" (TRPs) where turn-shifts can occur. These observations underscore the importance of recognising cues that distinguish TRPs from non-TRPs.

Traditional spoken dialogue systems often rely on a simplistic turn-taking model, utilising predefined silence duration thresholds detected by Voice Activity Detectors (VADs) to determine the end of a user's turn [10]. This method, prevalent in conversational systems, facilitates turn-taking by delineating the user's turn based on predefined thresholds, typically governed by parameters such as the no-input

timeout and the end-silence timeout [11]. While effective in its simplicity and ease of implementation, this basic model struggles with accurately detecting turn transitions, leading to interruptions or perceived system unresponsiveness. Typically, this approach does well in transactional spoken language interaction, where only simple one-off commands are given, but does not lead to a satisfactory use experience in spoken dialogue.

Researchers have explored integrating turn-taking cues into end-of-turn detection mechanisms to address the challenges. An approach involves detecting the end of inter-pausal units (IPU, the segment in which speech is being produced) using a VAD with shorter silence thresholds, followed by analysing turn-taking cues to identify TRP [4], [12]. Given the end of an IPU, the model has to predict whether the speaker is making a pause and "holding" the turn, or whether the speaker is yielding the turn [3]. Various features and machine learning algorithms have been employed on human-human and human-machine dialogue datasets [5], [1], [13], [14], [15]. However, these methods have limitations. They primarily focus on turn-taking based on silence and assume that turns occur only after a speaker has stopped speaking entirely. As Heldner and Edlund [16] highlight, overlaps are common in human-human dialogue, challenging the assumption that silent pauses indicate the relinquishing of a turn. In addition, IPU-based models are purely reactive, unable to predict upcoming turn transitions like humans do [3]. To achieve this more human-like prediction capability, it is essential that the models can process speech incrementally.

An alternative approach to traditional models is incremental processing, which allows systems to process user input incrementally and make continuous turn-taking decisions. This approach enables a deeper "understanding" of the user's utterance and facilitates advanced turn-taking strategies, such as projecting turn completions and planning responses while the user is speaking [17]. Other models demonstrate the effectiveness of continuous turn-taking models trained on human-human dialogue data, outperforming conventional methods and human judges in end-of-turn detection tasks [6]. In addition, incremental processing enhances interactions that involve task execution, enabling tasks to be initiated while instructions are being given. While offering deeper understanding and flexibility, incremental processing can come at the cost of increased computational complexity compared to simpler models.

In contrast to previous studies that primarily focus on rule-based or data-driven approaches for turn-taking, our research explores the potential of LLMs with VADs to predict or detect the end of a turn. The growing popularity of LLMs and their ability to encode a vast amount of contextual knowledge and linguistic semantics motivates our investigation into how these models can be used for more accurate turn-taking predictions.

III. Data

A. Datasets

To evaluate the ability of LLMs to predict turn-ending points in conversations, we leveraged three dialogue datasets encompassing a spectrum of conversation styles. These datasets provide valuable insights into how LLMs handle both scripted and spontaneous human interactions.

DailyDialog [18]: This dataset provides a rich collection of multi-turn, open-domain dialogues in English. The conversations aim to mimic real-world human interactions, covering diverse topics like relationships, daily life, work, and leisure activities. Its focus on daily communication makes it particularly suitable for evaluating LLMs' ability to understand and respond to natural, everyday conversations.

DialogStudio [19]: This dataset offers a diverse range of data, including open-domain dialogues, natural language understanding prompts, and conversational recommendations. This variety makes it a valuable resource for assessing LLM performance in different conversational tasks that cover various topics, conversational styles, and even cultural nuances. This focus on real-world conversation subtleties like humour, empathy, and complex discussions makes it a valuable resource for evaluating LLM performance in tasks like turn-ending prediction.

Switchboard [20]: This data set consists of spontaneous telephone conversations between humans, with predetermined topics for each call. The inclusion of spontaneous conversations adds a layer of complexity compared to scripted dialogues, as these conversations might have unexpected turns or conversational disfluencies. This dataset can be particularly useful for evaluating how well LLMs handle real-world conversational dynamics

B. Data Collection for LLM-VAD evaluation

To evaluate the effectiveness of a combined system utilising both a Large Language Model (LLM) and Voice Activity Detection (VAD) for predicting turn-taking cues in spoken dialogues; a crucial aspect for natural conversation in Human-Robot Interaction (HRI) applications.

We collected 14 audio recordings from seven Dutch-speaking participants aged 75 and above (77.57 ± 3.82 years old) who engaged in two separate conversations with a social robot, an Aldebaran Robotics' Pepper (see Figure 1 for the setup). All participants provided informed consent according to procedures approved by the Ghent University's Ethics Committee.

The conversations were open-ended, allowing participants to discuss any topics they wished. This approach encouraged a broad range of conversational dynamics, similar to what might occur during real-world interactions with a social robot. Conversations focused on daily life experiences, hobbies, and family matters. The average conversation duration was 9.93 minutes, with a range of 6.27 to 16.31 minutes. Researchers were present to ensure participant well-being, but did not interfere or participate in the conversations.

The audio recordings were captured in a quiet room within the participants' homes to create a comfortable, realistic,

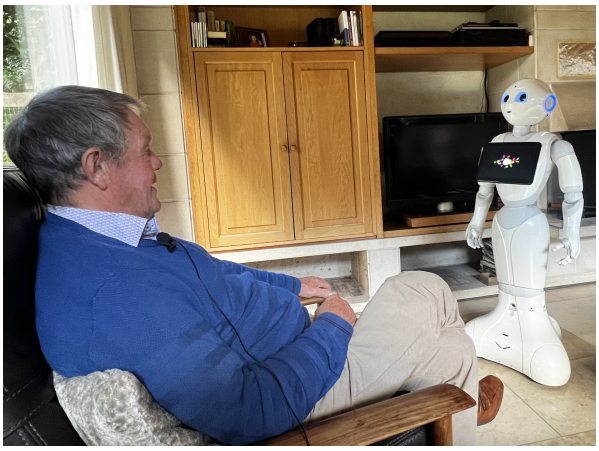

Fig. 1. Participant engaging in natural and autonomous conversation with the social robot Pepper

and familiar setting. This minimised background noise and ensured clear audio quality.

C. Data Extraction

Across all datasets, speaker identification was achieved by inserting labels (”speaker 1” or ”speaker 2”) at the beginning of each turn. Additionally, text normalisation was performed by removing punctuation marks and converting all characters to lowercase. This ensured consistent representation for the LLM regardless of speaker or conversation style. For each turn in the dialogue datasets, the corresponding sentence was extracted. Additionally, the entire conversation history leading up to that turn was compiled.

For the first two datasets, DailyDialog and DialogStudio, the turn was given by considering that the sentence was syntactic or pragmatic completed. Sentence segmentation was achieved using spaCy, a popular open-source Python Natural Language Processing library.

The Switchboard data set, composed of spontaneous spoken conversations, presented a distinct challenge due to its lack of well-defined sentence boundaries and the presence of disfluencies such as backchannels and overlapping speech. The pre-processing of this data was based on the work made by Ekstrand [8]. Backchannels spoken in isolation (e.g., ”mm”) were removed to eliminate non-informative vocalisations. To capture natural pauses, speech segments under 500 ms were defined as Inter-Pausal Units (IPUs). Overlapping speech, if present, could be further addressed using speaker identification techniques. Finally, to prevent merging incomplete utterances, consecutive IPUs from one speaker speaking entirely within another’s IPU were excluded. The final turn sequences were created by merging remaining IPUs with mutual silence, ordered by their occurrence time while acknowledging potential overlap.

To extract turn-taking information for evaluating the LLM-VAD system with elderly users, we segmented the audio recordings into individual speech turns. This process utilised speaker diarization techniques employing the x-vector method developed by MathWorks [21]. Similar to the Switchboard dataset, speech segments shorter than 500 milliseconds were defined as IPUs. An example of this data extraction process is illustrated in Figure 2.

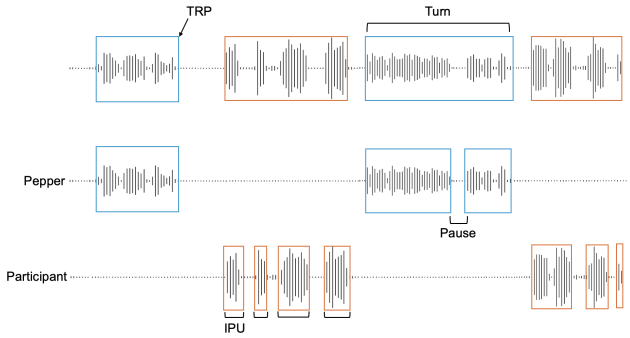

Fig. 2. Example of data extraction for LLM-VAD evaluation

IV. Evaluation

To evaluate a Large Language Model’s (LLM) ability to predict turn-ending points in conversations, we leveraged a prompt-based LLM API. For every turn, a standardised prompt was constructed (see the prompt below). This prompt incorporated the extracted sentence and explicitly instructed the LLM to analyse it for turn-ending cues like questions, invitations for input, or pause indicators. These prompts, encompassing both the sentence and the conversational history are sent to the designated LLM API using the appropriate API call.

> ***prompt:****”Analyse the following conversation snippet and determine whether the current speaker has yielded their turn. Respond with 1 if the turn has been yielded, or 0 if it hasn’t. Use the following criteria to make your decision: (1) Sentence Completion: Analyze whether the speaker’s message represents a complete sentence or thought that suggests closure. (Sentence: {sentence}) (2) Turn-Yielding Cues: Identify if the message includes any indicators such as questions, invitations for input, or pauses that might signal the speaker is yielding their turn. (3) Context Dialogue: Consider the ongoing conversation and the roles of the participants to understand the flow better. (Context: {ongoing_dialog}). Use the provided information to evaluate whether the current speaker has yielded their turn. Examples ...”*

The specific format for sending prompts and receiving responses (”1” for turn ended, ”0” for ongoing turn) will vary depending on the chosen API. However, this core process allows us to assess the LLM’s effectiveness in understanding conversation flow and turn-taking dynamics. By analysing the provided sentence and context for turn-yielding cues, the LLM’s predictions can be compared to ground truth labels in the data, revealing its proficiency in this crucial aspect of conversational understanding.

In Table I, we present the comparative results of five different LLM APIs in predicting turn-ending points in conversations. The table shows each model’s Recall, Precision, F1 score, and response time, providing a comprehensive overview of their performance. Response time refers to the time it took for the LLM’s API to process a speech

segment and return a prediction (lower is better for real-time applications). Highlighted cells (light gray) indicate the best-performing LLM in terms of a specific metric for a particular dataset. It's crucial to consider both accuracy metrics and response time when choosing an LLM for real-time conversational applications.

While our initial evaluation focused on the LLM's ability to predict turn-ending points based on sentence structure and turn-yielding cues, related work often models turn-taking decisions at a finer level. This includes predicting whether a speaker will continue speaking after a brief pause ("HOLD") or relinquish the turn to the other speaker ("SHIFT"). This capability is crucial for spoken dialogue systems such as social robots, allowing them to determine when the turn is relinquished. It could also be applied to predict user turn initiation after a system pause [3].

To investigate the potential of combining LLM with VAD for such fine-grained turn-taking prediction, we selected the two best performing LLMs from the initial evaluation. We then evaluated the full system using the dataset of 14 audio recordings from elderly users. These recordings, featuring open conversations about daily life, hobbies, and family, are characterised by longer pauses (above 700 milliseconds) and an average of 10.4 turns per dialogue, resulting in approximately 3.25 hours of test data. Analysing these recordings with the combined LLM-VAD system will provide insights into its effectiveness at three different time frames: 500 ms, 800 ms, 1200 ms, and 1500 ms. This allowed us to assess if the LLM could detect potential turn shifts even within shorter pauses where some voice activity might still be present.

It is essential to note that while we evaluate the model's performance after brief pauses (considering pauses even at the end of the sentence), the LLM is receiving a continuous stream of audio data. To achieve this, we leveraged speaker recognition from Azure. The audio recordings were uploaded to the Azure API, which segmented the audio by speaker and sent each segment to the LLM for turn-ending prediction. However, during the time frame the LLM is processing one segment (e.g., 800ms), VAD might detect additional voice activity. This triggers the sending of a new, updated segment with the most recent audio information to the LLM, effectively creating a continuous evaluation loop until the specified time frame is over. The results of this evaluation are presented in Figure 3

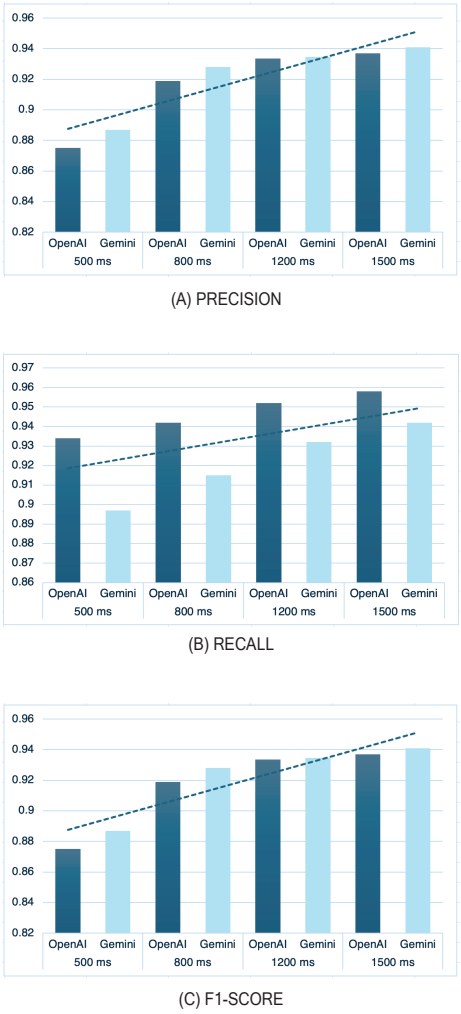

Fig. 3. Performance results of the LLM model and VAD in predicting turn-shifts points, depending on pause length. The dotted line illustrates the trend line, indicating the overall data pattern.

V. DISCUSSION

This study evaluated the capabilities of LLMs for predicting turn-ending points in conversations. We evaluated five LLMs across three dialogue datasets, showing valuable insights into their strengths and weaknesses (refer to Table I). Building upon these findings, we explored the potential of combining LLMs with VAD for fine-grained turn-taking prediction, specifically focusing on short, medium, and long pauses.

Our initial evaluation of five LLMs highlighted several key observations. First, the performance of LLM varied across datasets. We observed that structured dialogue formats, such as DailyDialog, led to predictive accuracy (F1-score) compared to Switchboard, which contained spontaneous conversations with disfluencies. This suggests that current LLMs benefit from well-defined turn structures, and further development might be necessary to handle the complexities of natural speech.

For instance, the LLM from OpenAI (gpt4-turbo) consistently demonstrated high recall across all datasets, with particularly notable performance on the Dialog dataset, achieving the highest F1-Score of 0.903. This indicates gpt4-turbo's robustness in identifying turn endings, a critical attribute for spoken dialogue systems aiming to minimise interruptions and maintain smooth dialogue flow. The high recall suggests that gpt4-turbo is less likely to miss turn endings, making it a reliable choice for applications requiring high engagement and responsiveness.

Gemini Pro presented the highest precision in the DailyDialog dataset, emphasising its strength in minimising false positives during turn-ending predictions. This precision is

					Response time (seg)	
Dataset	LLM	Precision	Recall	F1-Score	Average	STD
Daily Dialog	gpt-3.5-turbo-0125	0.761	0.827	0.792	0.560	2.571
	gpt-4-turbo	0.833	0.915	0.872	0.887	1.610
	gemini-pro	0.880	0.888	0.884	1.302	0.362
	anthropic	0.845	0.907	0.875	3.762	3.867
	llama-2	0.834	0.896	0.864	1.987	3.679
DialogStudio	gpt-3.5-turbo-0125	0.801	0.725	0.761	0.623	2.459
	gpt-4-turbo	0.873	0.934	0.903	0.983	1.293
	gemini-pro	0.866	0.884	0.875	1.291	0.458
	anthropic	0.845	0.900	0.871	4.137	3.641
	llama-2	0.800	0.665	0.726	1.856	3.503
Switchboard	gpt-3.5-turbo-0125	0.819	0.853	0.836	0.589	0.346
	gpt-4-turbo	0.862	0.938	0.898	0.897	2.035
	gemini-pro	0.857	0.903	0.879	1.103	0.399
	anthropic	0.829	0.909	0.867	3.987	3.927
	llama-2	0.813	0.791	0.802	2.034	4.134

TABLE I

COMPARATIVE PERFORMANCE OF DIFFERENT LLM APIS IN PREDICTING TURN-ENDING POINTS IN CONVERSATIONS. LIGHT GRAY INDICATES THE BEST-PERFORMING OF LLM IN THE SPECIFIC METRIC

particularly valuable in scenarios where the cost of interrupting the user erroneously is high, ensuring conversational agents act on more accurate signals of turn completions. In addition, the standard deviation of response time is particularly lower for Gemini Pro across all the datasets, i.e., consistent performance across multiple interactions. Consistency in response time is crucial for user experience, ensuring that users can anticipate the system's responsiveness, which is vital for building trust and reliability in human-robot interactions.

Second, a trade-off emerged between accuracy and response time. While models like gpt4-turbo demonstrated high recall and an impressive F1-Score of 0.903 in the Dialog dataset, emphasising its capability to reduce conversational gaps, others exhibited faster response times at the expense of precision and recall. This highlights the importance of considering application requirements when choosing an LLM for a dialogue system. For tasks prioritising immediate response (e.g., chatbots), faster LLMs might be preferable. For applications where turn-prediction accuracy is paramount (e.g., educational systems, companion robots), high-accuracy LLMs could be a better fit.

These findings underscore the advancements in leveraging LLM technology for conversational AI, demonstrating the potential of these models to enhance the naturalness and responsiveness of dialogue systems. However, the variability in model performance across different datasets also highlights the importance of context and dataset characteristics in model selection. For instance, models performing well on scripted dialogue datasets may not necessarily have the same performance in spontaneous conversation scenarios, and vice versa.

In our second analysis, the incorporation of VAD with LLMs, specifically OpenAI and Gemini, allowed for a refined prediction of turn transitions at various pause lengths. Both models showcased strong prediction capabilities, with their performance converging at longer pause durations. OpenAI slightly outperformed Gemini in F1-Score at longer pauses, suggesting a better handle on turn-taking nuances within extended silences.

Gemini, while performing comparably to OpenAI, offers a slightly lower F1-Score but maintains high precision across all pause durations. This suggests Gemini's consistency in minimising false positives is a strong point, which could be particularly important in user interactions where avoiding interruptions is critical.

These findings reveal that while both models perform well, they may be suited to different types of conversational scenarios. OpenAI might be more suitable in environments where longer pauses are prevalent and the cost of missing a turn transition is higher. In contrast, Gemini could be preferred in more dynamic settings where precision is paramount, and interruptions must be avoided.

It is essential to note that the slight variations in the metrics observed across pause durations can be attributed to the natural variability in conversational styles among users, further emphasised by the data collected from elderly individuals. In interactions with elderly users, longer pauses are often a natural part of the conversation, whether due to cognitive processing times, speech generation, or reflective pauses. These extended pauses necessitate a model that can discern the nuanced difference between a pause that indicates turn yielding and one that is simply a natural lull in the conversation. The above suggest that, while our LLM-VAD systems show promising results, the consideration of user-specific conversational nuances, such as those observed with elderly users, emphasises the need for further research into personalised conversational AI. This could lead to the development of systems that are not only accurate and efficient but also considerate of the user's speech patterns, enhancing the interaction for a more inclusive range of users.

Besides, it is also important to consider the operational context where these models would be deployed. The continuous audio stream processing, facilitated by Azure's speaker recognition, shows that these LLMs are capable of handling real-time data effectively, adjusting to new voice input as it

becomes available. This bodes well for the deployment of such models in real-world HRI systems, where the ability to process incremental information and make turn-taking decisions is essential.

Finally, it's important to acknowledge that evaluating turn-taking prediction is challenging. Since turn-taking behaviour can be subjective, and our evaluation is based on what humans actually did, achieving 100% accuracy might not be realistic. Skantze [6] reported a human F-score of 0.709, which is lower than the performance of our best LLM-VAD models 3.

VI. Conclusions

The combined analysis of LLM and LLM-VAD systems offers a comprehensive perspective on turn-ending prediction augmented by LLMs. While LLMs provide a valuable foundation for identifying turn-ending points based on sentence structure and semantic cues, the LLM-VAD system demonstrates promise for more nuanced turn-taking management, particularly in scenarios involving elderly users or natural, open-ended conversations. This capability is crucial for spoken dialogue systems, such as those implemented on social robots, allowing them to determine when to take the turn themselves and fostering more natural and engaging interactions.

The use of semantic information turns out to be particularly poignant in conversations between robots and elderly people. Too often, turn-taking prediction focuses on speed, using VAD to reduce the silent gap between turns as much as possible, thereby interrupting a turn in which the speaker pauses but without wishing to relinquish the turn. This frequently happens in elderly people, who often pause while collecting their thoughts, but without the wish to relinquish the turn. A robot that is not sensitive to this, would interrupt the conversation and lead to a subpar interaction experience. Our model, which account for the semantics of what is being said in addition to the temporal duration of the pauses, remedies this.

Integrating additional modalities like speaker identification and visual cues, such as gaze fixations, alongside textual content could provide a richer context for LLMs. This richer context could lead to more robust turn-ending prediction across diverse conversation styles.

ACKNOWLEDGMENTS

This research received funding from the Bijzonder Onderzoeksfonds (BOF) of Ghent University and the Flanders AI Research 2 project.

References

[1] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a map task dialogue system," *Computer Speech & Language*, vol. 28, no. 4, pp. 903–922, 2014.

[2] L. Ten Bosch, N. Oostdijk, and L. Boves, "On temporal aspects of turn taking in conversational dialogues," *Speech Communication*, vol. 47, no. 1-2, pp. 80–86, 2005.

[3] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[4] L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody," in *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 2061–2064.

[5] M. Johansson and G. Skantze, "Opportunities and obligations to take turns in collaborative multi-party human-robot interaction," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 305–314.

[6] G. Skantze, "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*, K. Jokinen, M. Stede, D. DeVault, and A. Louis, Eds. Saarbrücken, Germany: Association for Computational Linguistics, Aug. 2017, pp. 220–230. [Online]. Available: https://aclanthology.org/W17-5527

[7] R. Masumura, T. Tanaka, A. Ando, R. Ishii, R. Higashinaka, and Y. Aono, "Neural dialogue context online end-of-turn detection," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, K. Komatani, D. Litman, K. Yu, A. Papangelis, L. Cavedon, and M. Nakano, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 224–228. [Online]. Available: https://aclanthology.org/W18-5024

[8] E. Ekstedt and G. Skantze, "TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2981–2990. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.268

[9] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn taking for conversation," in *Studies in the organization of conversational interaction*. Elsevier, 1978, pp. 7–55.

[10] S. McGlashan, "Voice extensible markup language (voicexml) version 2.0," *World Wide Web Consortium (W3C) Recommendation*, 2004.

[11] S. Witt, "Modeling user response timings in spoken dialog systems," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 231–243, 2015.

[12] A. Raux and M. Eskenazi, "Optimizing endpointing thresholds using dialogue features in a spoken dialogue system," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008, pp. 1–10.

[13] L. Bell, J. Boye, and J. Gustafson, "Real-time handling of fragmented utterances," in *Proc. NAACL workshop on adaptation in dialogue systems*, 2001, pp. 2–8.

[14] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa, "Learning decision trees to determine turn-taking by spoken dialogue systems." in *INTERSPEECH*, 2002, pp. 861–864.

[15] D. Lala, K. Inoue, and T. Kawahara, "Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 226–234.

[16] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[17] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '09. USA: Association for Computational Linguistics, 2009, p. 710–718.

[18] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, G. Kondrak and T. Watanabe, Eds. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995. [Online]. Available: https://aclanthology.org/I17-1099

[19] J. Zhang, K. Qian, Z. Liu, S. Heinecke, R. Meng, Y. Liu, Z. Yu, S. Savarese, and C. Xiong, "Dialogstudio: Towards richest and most diverse unified dataset collection for conversational ai," *arXiv preprint arXiv:2307.10172*, 2023.

[20] J. J. Godfrey and E. Holliman, *Switchboard-1 Release 2 LDC97S62*, Linguistic Data Consortium, Philadelphia, 1993, web Download. [Online]. Available: https://catalog.ldc.upenn.edu/LDC97S62

[21] MathWorks, "Speaker diarization using x-vectors." [Online]. Available: https://www.mathworks.com/help/audio/ug/speaker-diarization-using-x-vectors.html#mw_rtc_SpeakerDiarizationUsingXVectorsExample_DB133825