Investigating the Impact of Incremental Processing and Voice Activity Projection on Spoken Dialogue Systems

Yuya Chiba[1], **Ryuichiro Higashinaka**[2]
[1]NTT Communication Science Laboratories, Japan
[2]Graduate School of Informatics, Nagoya University, Japan
yuya.chiba@ntt.com, higashinaka@i.nagoya-u.ac.jp

Abstract

The naturalness of responses in spoken dialogue systems has been significantly improved by the introduction of large language models (LLMs), although many challenges remain until human-like turn-taking can be achieved. A turn-taking model called Voice Activity Projection (VAP) is gaining attention because it can be trained in an unsupervised manner using the spoken dialogue data between two speakers. For such a turn-taking model to be fully effective, systems must initiate response generation as soon as a turn-shift is detected. This can be achieved by incremental response generation, which reduces the delay before the system responds. Incremental response generation is done using partial speech recognition results while user speech is incrementally processed. Combining incremental response generation with VAP-based turn-taking will enable spoken dialogue systems to achieve faster and more natural turn-taking. However, their effectiveness remains unclear because they have not yet been evaluated in real-world systems. In this study, we developed spoken dialogue systems that incorporate incremental response generation and VAP-based turn-taking and evaluated their impact on task success and dialogue satisfaction through user assessments.

1 Introduction

The advent of large language models (LLMs) has fueled their active incorporation into the response generation of dialogue systems (Shuster et al., 2022; Bubeck et al., 2023; Hudeček and Dušek, 2023). In contrast, many spoken dialogue systems still rely on a timeout method for dialogue control, where user utterances are processed as discrete units. As a result, a significant gap remains in response timing and turn-taking between interactions with dialogue systems and human-to-human conversations.

To develop a spoken dialogue system capable of conversing at a human-like tempo, techniques for modeling turn-taking in human conversation have been actively studied (Skantze, 2021). One such technique, Voice Activity Projection (VAP) (Ekstedt and Skantze, 2022a,b), has gained attention because it can be trained in an unsupervised manner using spoken dialogue data between two speakers. The VAP model takes both speakers' speech as input and predicts future voice activity (VA) rather than a specific turn-taking event itself. This approach enables a single model to handle various turn-taking events. Since the model is trained on human conversations, it is expected to contribute to the development of spoken dialogue systems with more natural turn-taking abilities. However, previous studies have only evaluated models using fixed corpora, and the effect on the overall quality of interactive dialogues remains unclear.

In addition, even if systems are equipped with a natural turn-taking model, such a model will be ineffective if response generation cannot begin immediately once a turn-shift is detected. Incremental response generation is an approach that addresses this issue. In this approach, a user's speech is recognized through streaming speech recognition, and candidate responses are generated as partial speech recognition results are obtained. When a turn-shift is identified, the most suitable candidate response is selected as the system's final utterance. Several dialogue systems have been developed that incorporate incremental response generation (Nakano et al., 2000; Michael, 2020; Chiba et al., 2024). Although the effects of incremental processing in dialogue systems have been evaluated for various tasks (Skantze and Schlangen, 2009), the impact of recent LLM-based response generation remains unclear.

In this study, we evaluated the impact of VAP-based turn-taking combined with incremental response generation on task success and dialogue

Proceedings of the 31st International Conference on Computational Linguistics, pages 3687–3696
January 19–24, 2025.

satisfaction. To the best of our knowledge, this is the first study to investigate the effectiveness of a turn-taking algorithm in interactive settings using state-of-the-art dialogue processing techniques. We developed both task-oriented and non-task-oriented dialogue systems and conducted user evaluations through module ablation. This paper analyzes our experimental results and provides design guidelines for future spoken dialogue systems employing VAP-based turn-taking.

2 Related Work

2.1 Turn-taking models

Many studies have modeled human-to-human turn-taking for integration into dialogue systems (Skantze, 2021). Researchers have explored approaches that incrementally observe user speech to determine when the system should respond. Among deep learning models, Long Short-Term Memory (LSTM), a type of sequential model, has been widely used in various studies (Maier et al., 2017; Hough and Schlangen, 2017; Skantze, 2017). An extended LSTM model that effectively integrates acoustic and linguistic information has also been proposed (Roddy et al., 2018; Roddy and Harte, 2020).

In recent years, such Transformer-based approaches as TurnGPT (Ekstedt and Skantze, 2020) and Voice Activity Projection (VAP) (Ekstedt and Skantze, 2022a,b) have been introduced in this area. The VAP model, in particular, has gained attention because it can be trained in an unsupervised manner using only the spoken dialogue data between two speakers. Various extensions are being explored for VAP, including the integration of multimodal information (Onishi et al., 2023) and linguistic data (Liermann et al., 2023). Additionally, Inoue et al. (2024) adapted the VAP model to Japanese conversations. By employing the VAP model, dialogue systems are expected to enable more natural turn-taking. However, since previous studies have only demonstrated its effectiveness using fixed corpora, it remains unclear whether it can effectively improve task success and dialogue satisfaction in interactive settings.

2.2 Implementation of dialogue systems based on incremental processing

Various systems have been implemented to perform incremental dialogue processing. Miyazaki et al. (2005) developed an early dialogue system that generates responses using partial speech recognition and language understanding results with the WIT toolkit (Nakano et al., 2000). This system facilitated the development of incremental spoken dialogue systems, but cannot be extended with modern components such as LLMs. Schlangen and Skantze (2011) proposed, for dialogue systems based on incremental processing, an architecture that handles messages exchanged between modules as small pieces of information called incremental units (IUs). The modules are developed as incremental modules (IMs) that process information each time IUs are received. Various systems have been built using this architecture (Schlangen et al., 2010; Skantze and Schlangen, 2009). RETICO (Michael, 2020), a toolkit for spoken dialogue systems based on this architecture, has also been recently developed.

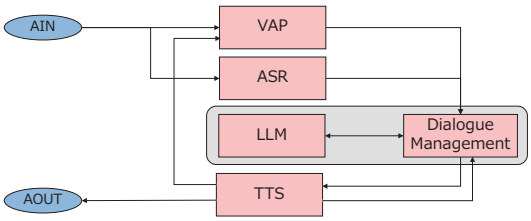

Figure 1: Module structure of a spoken dialogue system on Remdis: Pink rectangles represent incremental modules (IMs). Blue ovals indicate input and output.

Remdis (Realtime Multimodal Dialogue System Toolkit[1]) (Chiba et al., 2024) is another toolkit for building modular spoken dialogue systems based on incremental processing. Similar to RETICO, Remdis employs an architecture proposed by Schlangen et al. (2010), but it manages turn-taking using a VAP model and utilizes an LLM for response generation. We chose Remdis for constructing experimental systems because it supports advanced response generation with LLMs, and its modules for incremental processing and VAP can be toggled on and off, making it ideal for our purposes. We provide a more detailed description of Remdis in the next section.

3 Remdis: Realtime Multimodal Dialogue System Toolkit

3.1 Overview

This paper constructs systems for comparison using Remdis (Chiba et al., 2024). Figure 1 illustrates a modular structure of the spoken dialogue

[1] https://github.com/remdis/remdis

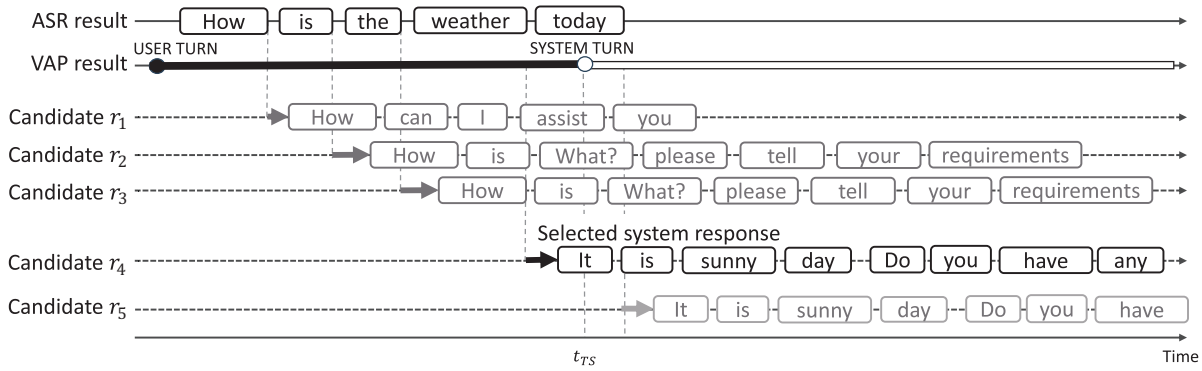

Figure 2: Incremental response generation using Voice Activity Projection (VAP): In Remdis, as partial speech recognition results are received, streaming response generation is performed in parallel by LLM. When VAP decision shifts from user's turn to the system's turn, the LLM that began generation with the most recent speech recognition result is selected to generate system's utterance for this turn. In the figure, VAP predicts a turn-shift at t_{TS}, and candidate r_4 is selected as system's utterance.

system built on Remdis. Each module is designed as an incremental module (IM) that processes and transmits results whenever it receives an IU.

3.2 Incremental processing

Remdis performs pseudo-incremental response generation through parallel LLM execution and streaming response generation. The system initiates streaming response generation with a new LLM instance each time it receives partial speech recognition results. Streaming speech recognition is implemented using the Google Cloud Speech-to-Text API[2]. When a turn-shift is detected, the system finalizes its utterance using the LLM instance that generated a response based on the most recent speech recognition results.

Figure 2 shows an example of Remdis's response generation. Each time word-level speech recognition results are received, a new LLM instance is initiated to generate a response based on the user's utterance at that moment. In this example, partial speech recognition results were received five times, resulting in five LLM instances that generated responses. The turn-shift timing from the user to the system is denoted as t_{TS}. At this point, among the LLM instances generating responses, the one using the most recent speech recognition result is selected for the final system response. In the figure, candidate r_4 is chosen as the system's response.

The generated response is converted into a waveform through speech synthesis. We used `ttslearn`[3] for the speech synthesis. Since current speech synthesis systems assume that the entire utterance text is provided at once, the quality of synthesized speech significantly degrades when executed in short units, such as tokens. To mitigate this issue, Remdis buffers the text generated by the response and performs speech synthesis incrementally at each punctuation unit.

3.3 Turn-shift prediction

Remdis uses the VAP model's output to determine the turn-shift (t_{TS}). A VAP module in Remdis utilizes a model trained with a package distributed at `https://github.com/ErikEkstedt/VAP`. The VAP model provides two estimates: p_{now} and p_{future}. They represent the aggregated values of estimated future VA, where p_{now} covers the interval from 0.0 to 0.6 seconds, and p_{future} covers the interval from 0.6 to 2.0 seconds, indicating the probability that one speaker will take a turn. In this study, when both p_{now} and p_{future} exceed a certain threshold, it becomes the system's turn; when they are below the threshold, it becomes the user's turn. When the user is speaking, if p_{now} exceeds the threshold and p_{future} is below it, the system provides backchannel responses. If the VAP module is disabled, the timing of the confirmed speech recognition is used for t_{TS}.

4 Experiments

We developed both task-oriented and non-task-oriented dialogue systems with varying turn-taking conditions. In this study, all the interac-

[2] `https://cloud.google.com/speech-to-text`

[3] `https://github.com/r9y9/ttslearn`

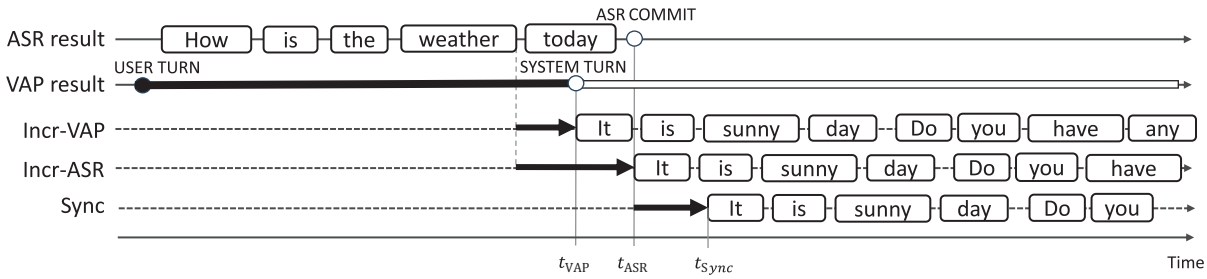

Figure 3: Difference of response timing among comparison systems: In this example, Incr-VAP, Incr-ASR, and Sync generate responses at t_{VAP}, t_{ASR}, and t_{Sync}.

tions with the dialogue system were conducted in Japanese. The experiments were approved by the ethical review committee for the graduate school of informatics, Nagoya University.

4.1 Methods for comparison

This study compares spoken dialogue systems between incremental and conventional synchronous processing, as well as systems with and without VAP-based turn-taking. With Remdis, these systems can be easily implemented by enabling or disabling specific modules. Participants interacted with multiple spoken dialogue systems that differed only in their response generation and turn-taking methods and evaluated the interactive dialogue quality of each system.

The methods for comparison are summarized below.

Incr-VAP: performs incremental response generation, and the turn-shifts are based on the VAP prediction results. It uses all the modules shown in Fig. 1.

Incr-ASR: performs incremental response generation, but the turn-shifts are based on the confirmation of the speech recognition rather than VAP. It uses modules other than VAP.

Sync: simulates a conventional synchronous dialogue system. It starts the response generation using the entire user utterance after confirming the speech recognition's end. Although this decision allows the system to use the entire user utterance for response generation, the response timing is delayed.

Figure 3 shows the differences in response timings among the comparison methods. t_{VAP}, t_{ASR}, and t_{Sync} represent the response generation timings for each one. In many cases, the VAP model predicts the turn-shifts faster than the speech recognition's confirmation, meaning that Incr-VAP has quicker response timing than Incr-ASR. Unfortunately, in Incr-VAP, the number of ignored tokens increases, which may reduce the accuracy of the response content.

4.2 Task-oriented dialogue system

We used JMultiWOZ (Ohashi et al., 2024) to implement a task-oriented dialogue system. JMultiWOZ, which is a Japanese multi-domain task-oriented corpus inspired by MultiWOZ (Budzianowski et al., 2018), provides conversations spanning six travel-related domains: "tourist attractions," "accommodations," "restaurants," "shopping facilities," "taxis," and "weather."

The response generation process in MultiWOZ is broadly divided into two steps: Dialogue State Tracking (DST) and Response Generation (RG). In the former step, the dialogue history is input into the LLM to estimate the dialogue state, which includes the user's intent and search conditions. The dialogue state is then used as a query to search the database. In the RG step, the search results and dialogue state from the DST step are integrated with the dialogue history to generate responses. An LLM is employed in both the DST and RG steps. In the incremental systems, DST is conducted in parallel each time partial speech recognition results are received. The most recent LLM instance that is ready to generate responses is selected, and response generation is performed in a streaming manner. Ohashi et al. (2024) compared an LLM-Pipeline using ChatGPT (`gpt-3.5-turbo` and `gpt4`) (Hudeček and Dušek, 2023) and a T5-based Pipeline[4] from a previous study (Bang et al., 2023). Since they reported superior performance with the T5-based Pipeline, we employed it for our experiments.

Since DST's inference speed is particularly crit-

[4] https://github.com/nu-dialogue/jmultiwoz

You are a chat assistant engaging in a casual conversation. For the following user utterance, create and output a clever reaction or question that is separated by punctuation marks. Summarize the output in about one sentence; a lengthy explanation is not necessary.

Table 1: Prompt for non-task-oriented dialogue systems (translated from Japanese)

ical for real-time interaction, we used the `t5-base` model. We also quantized the DST model to 8-bit integers using `ctranslate2`[5] to further improve its inference speed. For the RG step, we used the `t5-large` model to generate as natural responses as possible. The beam width was set to 1, and the maximum number of output tokens was set to 256. The temperature was set to 1.0. As the input length increases, the response generation speed significantly decreases, and so the dialogue history was limited to five utterances. To reduce the computational load, response generation was carried out every five input words.

4.3 Non-task-oriented dialogue system

In non-task-oriented dialogue systems, we used ChatGPT for response generation. For this study, we used `gpt-3.5-turbo` because at the time of our experiments, it has the fastest response times and best performances. The prompts for the non-task-oriented dialogue systems are shown in Table 1. The prompts request that utterances be split at punctuation to increase the segmentation points and keep them as brief as possible to ensure a smooth conversation flow. Note that we did not specify a conversational topic in the prompts. The experiment's participants introduced the topic themselves, followed by the system's utterance; conversations were primarily user-driven.

For the input, the user's current utterance was used with the three preceding utterances as a dialogue history. Response generation was performed each time partial speech recognition results were received. The maximum token length was set to 128, and the temperature was set to 1.0.

4.4 VAP conditions

The duration of the user and system utterances input into the model was set to 20 seconds, following previous studies (Ekstedt and Skantze, 2022a,b). The threshold for turn-taking determination was set at 0.5 based on our preliminary experiments. Since VAP implementation does not support sequential processing, the speech data incrementally stored in the buffer were input into the model at every timestep to predict future VA.

We used the Japanese VAP model for our experiments (Sato et al., 2024). This model was pretrained on the Switchboard Corpus and fine-tuned using various Japanese dialogue corpora, such as the CALLHOME Japanese Speech (Wheatley et al., 1996) and the Travel Agency Dialogue Corpus (Inaba et al., 2024). Evaluation of this model on the Japanese dataset yielded an F-score of 74.0% for Shift/Hold and 71.4% for S-pred[6]. These results are comparable to the model's performance on English datasets in previous studies (Ekstedt and Skantze, 2022a,b).

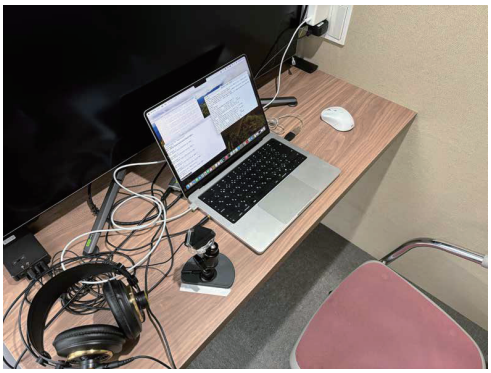

Figure 4: Experimental setup

4.5 Experimental conditions

Twenty-three people (12 males and 11 females) were engaged through a recruiting agency to join our experiments. Their ages ranged from their 20s to their 60s. Each participant engaged in dialogues with six systems (i.e., task-oriented and non-task-oriented systems using three different turn-taking methods). They visited a booth (Fig. 4) and conversed with the systems. We randomized whether the task- or non-task-oriented dialogue systems were presented first. The presentation order of the methods was also randomized. The goals of the task-oriented dialogues were randomly generated for the "weather," "shopping," and "restaurant" domains. The topics for the non-task-oriented dialogues were selected from typical everyday con-

[5] https://github.com/OpenNMT/CTranslate2

[6] Shift/Hold and S-pred both represent the accuracy of the turn-shift predictions. For a more precise definition, refer to (Ekstedt and Skantze, 2022a).

System	Method	Gap↓
Task	Sync	1.120 (± 0.077)
Task	Incr-ASR	1.063 (± 0.061)
Task	Incr-VAP	**0.748** (± **0.056**)
Chat	Sync	0.959 (± 0.054)
Chat	Incr-ASR	0.763 (± 0.038)
Chat	Incr-VAP	**0.610** (± **0.036**)

Table 2: Average and standard error of gaps for each system (sec): Task and Chat represent task- and non-task-oriented dialogue systems. Bold fonts indicate shortest gap in each system.

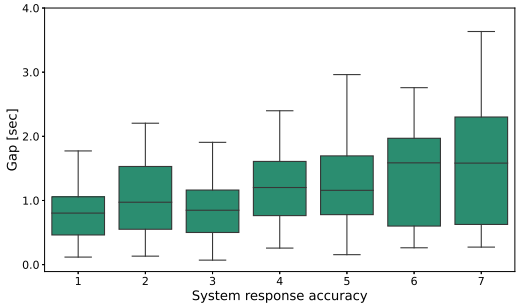

Figure 5: Distribution of gaps for *system response accuracy* in both systems: Gaps are averaged for each dialogue.

versation subjects (e.g., "work and study," "family," and "movies"). There were 16 conversation topics in total.

Before each dialogue, an experimenter explained the goals of the task-oriented dialogues and provided conversation topics for the non-task-oriented dialogues. For the task-oriented dialogues, the experiment ended when the participants believed they had achieved their goal or when they determined that the task could not be completed. The dialogues were automatically terminated after three minutes. After each interaction with a system, participants completed a survey for subjective evaluation. A post-experiment survey was also conducted after all the dialogues were completed.

4.6 Objective and subjective evaluations

For the objective evaluations, we calculated the gap between the end of the user's utterance and the beginning of the system's utterance. This gap was measured based on the Voice Activity Detection (VAD) results. The systems continuously monitored the user and system utterances and performed VAD. We used the WebRTC VAD system[7] with an interval of 0.001 seconds and a frame length of 0.010 seconds. The aggressiveness mode was set to its default setting.

For subjective evaluations, we employed four commonly used scales for dialogue system assessments: *satisfaction*, *naturalness*, *engagement*, and *consistency*. Additionally, we used six scales from the SASSI questionnaire (Hone and Graham, 2000): *system response accuracy*, *likeability*, *cognitive demand*, *annoyance*, *habitability*, and *speed*. Participants rated these scales on a seven-point scale (1: lowest, 7: highest).

[7]https://github.com/wiseman/py-webrtcvad

5 Experimental Results

5.1 Comparison of gaps between systems

First, we compared the gaps between the systems. Table 2 shows the mean and standard error of the gaps for each one. The first system utterance was excluded, as response generation tends to be slower immediately after model-loading. The introduction of incremental processing and turn-taking with VAP reduced the gap for both systems. Factoring in turn-taking conditions, we conducted a one-way ANOVA, and found a significant difference ($p < 0.001$). Additionally, a multiple comparisons t-test with Bonferroni correction revealed significant differences among all the conditions for the task-oriented systems and between Sync and Incr-VAP for the non-task-oriented systems ($p < 0.001$). This suggests that incremental processing and VAP reduced the gap duration. However, the average gap in the human-human dialogues is typically around 0.2 seconds (Skantze, 2021), indicating room for further improvement.

5.2 Comparison of user evaluations

Next we present the subjective evaluation results in Table 3, which shows the average scores for ten items for each system. The scores for Sync, which simulates conventional dialogue systems, are above average for items such as *satisfaction* and *system response accuracy*. Most user ratings decreased as the response generation process progressed in Sync, Incr-ASR, and Incr-VAP. One possible reason for this decline is that faster turn-taking decisions resulted in selecting LLMs that used more incomplete user utterances. Fig. 5 shows the distribution of the gaps for the ratings for the *system response accuracy* across all the systems. The gaps were averaged for each dia-

System	Method	Sat↑	Nat↑	Eng↑	Con↑	SRA↑	Like↑	CD↓	Annoy↓	Habit↑	Sp↑
Task	Sync	**3.652**	**3.913**	**3.739**	**4.478**	**4.261**	**3.957**	**4.435**	**3.522**	**3.652**	5.174
Task	Incr-ASR	3.261	3.696	3.261	3.957	4.000	3.609	5.217	3.913	3.391	**5.217**
Task	Incr-VAP	2.478	3.043	2.957	4.304	3.043	3.130	4.826	4.217	2.783	**5.217**
Chat	Sync	**4.217**	**4.043**	**4.043**	**4.609**	**4.435**	**3.957**	4.739	**3.261**	**3.478**	**5.304**
Chat	Incr-ASR	3.478	3.652	3.565	4.087	3.696	3.522	**4.696**	3.478	3.043	5.130
Chat	Incr-VAP	3.000	3.304	2.913	3.565	3.261	3.391	5.261	4.000	3.000	5.217

Table 3: Average subjective evaluation score for each system: Task and Chat represent task- and non-task-oriented dialogue systems. Sat: satisfaction, Nat: naturalness, Eng: engagement, Con: consistency, SRA: system response accuracy, Like: likeability, CD: cognitive demand, Annoy: annoyance, Habit: habitability, Sp: speed. Bold fonts indicate best scores in each system.

Method	Success	Fail	Out of time
Sync	23.8	4.8	71.4
Incr-ASR	22.7	4.5	72.7
Incr-VAP	9.1	13.6	77.0

Table 4: Task success and fail rate for task-oriented dialogue systems [%]

logue. Dialogues with high scores were more frequently distributed among those with larger average gaps. The correlation coefficient between gaps and ratings was $r = 0.265$, $p < 0.01$, suggesting that dialogues with shorter response times tended to generate more inaccurate responses to user input. In terms of *speed*, all the systems were rated highly. However, contrary to expectations, the ratings were nearly equal across all the systems. Participants did not perceive the gap differences achieved through the incremental response generation and the VAP models. We conducted a one-way ANOVA factoring the systems for *speed* and found no significant differences.

Table 4 displays the task success and failure rates for the task-oriented dialogue systems. The task success rate for Sync was 23.8%. In a previous study, Ohashi et al. (2024) reported task success rates exceeding 65% using the same model. One possible reason for this discrepancy is that the response generation models used in their experiments were trained on text data, which may not align well with user utterances in spoken dialogues. For example, while users tend to provide longer inputs with more slot information to achieve their goals in text dialogues, they generally provide shorter, more segmented inputs in spoken dialogues. These findings correspond with experiments using SpokenWOZ (Si et al., 2023), which collected spoken dialogues based on the MultiWOZ framework. Si et al. (2023) reported a larger decrease in task success in spoken dialogues compared to text-based ones.

From these results, we conclude that the simple introduction of incremental response generation with VAP did not improve the user evaluations, despite reducing the gap durations. In the current framework, as response timing quickened, the user utterances for the response generation were more incomplete, perhaps leading to an increase in responses that did not align with the user input.

5.3 Error analysis

A portion of a dialogue with Incr-VAP is shown in Fig. 6. This example contains several turn-shift prediction errors. Around 1.0 second, the VAP output leans toward the system's turn, even though the user is trying to continue speaking. The VAP model used in the experiment tended to incorrectly determine turn-taking at the inter-pausal unit (IPU) termination. This result suggests that the model is ineffectively capturing prosodic information near the end of an IPU. Additionally, since the two system utterances from 1.0 to 5.5 seconds were originally part of one utterance, the system misappraised the turn-shift during the pause between these segments. Furthermore, the system utterance ("What kind of food do you like?") was generated due to this misjudgment and uttered in succession. This likely occurred because the VAP model was trained on human conversations and thus mismatches with synthesized speech. The pauses in the synthesized speech are longer than those in human speech, leading the model to incorrectly judge turn-shifts.

6 Discussions

We next identify issues with current systems and discuss future challenges for achieving real-time

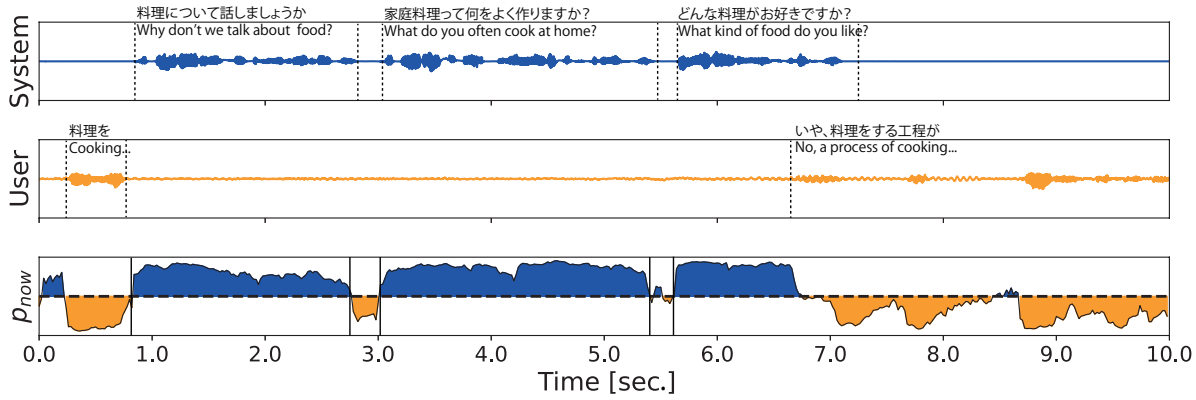

Figure 6: Typical examples of turn-shift prediction error: Texts are transcriptions of utterances translated from Japanese. First and second panels display waveforms of system and user speech. Third panel shows the transition of p_{now}. In this figure, segments where p_{now} is greater than the center represent system turns, and segments where p_{now} is smaller than the center represent user turns.

spoken dialogue systems with high user evaluations.

Incremental response generation: Experiments showed that the response accuracy decreased as the response time shortened, resulting in Sync achieving the highest subjective evaluation scores. This suggests that, to improve the user evaluations, we must not only reduce the response times but also simultaneously ensure response accuracy. Current studies on dialogue systems assume that an entire user utterance is always provided, which does not align with the fast-paced nature of spoken dialogues. Therefore, a framework is needed for incremental response generation that can deliver appropriate responses aligned with the context and the tempo of spoken dialogues, even when only partial user utterances are available. The key to achieving this lies in interpreting fragmented user utterances and constructing a shared understanding with users based on such interpretations. For response generation using LLMs, considering past utterances would be effective while also predicting user utterances that have not yet been observed (e.g., (Ohagi et al., 2024)). Fine-tuning the model using the transcriptions of spoken dialogues could also prove useful.

Another issue involves handling user disfluencies. Since the speech recognition system and the LLM used in this study seemingly failed to fully account for such disfluencies, instances of incorrect responses could have been generated in reaction to user disfluencies or intermediate results from the speech recognition process. Accurately recognizing disfluencies and enhancing LLMs to better manage them is essential, as discussed in Baumann et al. (2017).

Robust turn-shift prediction: There is room for improving the robustness of the VAP model. In particular, the model used in our experiments was trained on human conversations and often behaves unexpectedly when synthesized speech is input. For example, incorrect turn-shifts frequently occur during system turns because pauses in synthesized speech are longer than those in human utterances. Additionally, the prediction results are significantly influenced by microphone characteristics and settings during the experiment. Therefore, the model must be trained with a wider range of speech data, including synthesized speech.

Moreover, the VAP model often assumed it was the system's turn as soon as the user's speech stopped. From post-experiment surveys, many participants felt frustrated that the system did not wait for them to fully finish their utterances. A turn-shift decision method that considers pauses and other conversational cues is needed to better accommodate natural dialogue flow.

Conversational speech synthesis: Current speech synthesis systems are designed to receive and output entire utterances, and the quality of the synthesized speech is not guaranteed when partial utterance texts are provided as input. This limitation necessitates buffering a certain length of speech for synthesis, which introduces a delay before the system's speech is uttered. Streaming speech synthesis, which incrementally generates speech from partial text, is necessary to address this issue. To maximize the effectiveness of cur-

rent VAP models, the generated speech must also be made more conversational. Training speech synthesis models using spoken dialogue data, as done in previous studies (Rubenstein et al., 2023; Nguyen et al., 2023; Iizuka and Mori, 2022), will be useful for achieving this goal.

7 Conclusions

We investigated the effectiveness of incremental processing with VAP-based turn-taking in spoken dialogue systems. For the first time, this paper empirically verified the effectiveness of such turn-taking algorithms in interactive systems using state-of-the-art dialogue processing. We developed both task-oriented and non-task-oriented dialogue systems, conducted experiments, and evaluated the quality of the interactions. The results showed that although incremental processing and VAP-based turn-taking successfully reduced gap duration, they did not notably improve the dialogue quality, because shorter response times caused more inaccurate system responses. Our findings suggest a potential path toward achieving real-time dialogue systems with high-quality interactive performance.

8 Limitations

Effect of model performance: For the non-task-oriented dialogue systems, we used `gpt-3.5-turbo`, which was deemed the best model at the time of our experiments. Similarly, for the task-oriented dialogue systems, we used the T5-based model, which we believed to be the best system for the JMultiWOZ task. However, these dialogue models may not necessarily be the best choice for real-time spoken dialogue. More recent models, such as `gpt-4o-mini`, may offer better user evaluations. The impact of model performance needs to be investigated further in future studies.

Cultural differences in turn-taking: Cultural differences exist in acceptable pause lengths and the frequency of backchannels in conversations (Jokinen et al., 2013; Cutrone, 2005). These studies suggest that longer pauses are acceptable in Japanese conversations, and that Japanese speakers tend to use more backchannels. Our participants reported that frequent interruptions by the system were annoying, although perhaps the results may differ in other cultures. We believe that more general conclusions can be drawn by investigating the effects of the examined methods across different languages and cultural contexts.

Acknowledgments

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011.

References

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. In *Findings of ACL*, pages 7355–7369.

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there. *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, pages 421–432.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, pages 1–155.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ–a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proc. EMNLP*, pages 5016–5026.

Yuya Chiba, Koh Mitsuda, Akinobu Lee, and Ryuichiro Higashinaka. 2024. The Remdis toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models. In *Proc. IWSDS*, pages 1–6.

Pino Cutrone. 2005. A case study examining backchannels in conversations between Japanese–British dyads. *Journal of Cross-Cultural and Interlanguage Communication*, 24(3):237–274.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. *arXiv preprint arXiv:2010.10874*, pages 1–10.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? Investigations using voice activity projection models. In *Proc. SIGDIAL*, pages 541–551.

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

Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6(3-4):287–303.

Julian Hough and David Schlangen. 2017. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proc. EACL*, pages 326–336.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are large language models all you need for task-oriented dialogue? In *Proc. SIGDIAL*, pages 216–228.

Takahisa Iizuka and Hiroki Mori. 2022. How does a spontaneously speaking conversational agent affect user behavior? *IEEE Access*, 10:111042–111051.

Michimasa Inaba, Yuya Chiba, Zhiyang Qi, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2024. Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(9).

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using voice activity projection. *arXiv preprint arXiv:2401.04868*, pages 1–7.

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems*, 3(2):1–30.

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

Angelika Maier, Julian Hough, David Schlangen, et al. 2017. Towards deep end-of-turn prediction for situated spoken dialogue systems. In *Proc. INTERSPEECH*, pages 1676–1680.

Thilo Michael. 2020. RETICO: An incremental framework for spoken dialogue systems. In *Proc. SIGDIAL*, pages 49–52.

Noboru Miyazaki, Mikio Nakano, and Kiyoaki Aikawa. 2005. Spoken dialogue understanding using an incremental speech understanding method. *SYSTEMS and COMPUTERS in Japan*, 36(12):75–84.

Mikio Nakano, Noboru Miyazaki, Norihito Yasuda, Akira Sugiyama, Jun-ichi Hirasawa, Kohji Dohsaka, and Kiyoaki Aikawa. 2000. WIT: A toolkit for building robust and real-time spoken dialogu systems. In *Proc. SIGDIAL*, pages 150–159.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Masaya Ohagi, Tomoya Mizumoto, and Katsumasa Yoshikawa. 2024. Investigation of look-ahead techniques to improve response time in spoken dialogue system. In *Proc. INTERSPEECH*, pages 3580–3584.

Atsumoto Ohashi, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. 2024. JMultiWOZ: A large-scale Japanese multi-domain task-oriented dialogue dataset. In *Proc. LREC-COLING*, pages 9554–9567.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *Proc. HAI*, pages 13–21.

Matthew Roddy and Naomi Harte. 2020. Neural generation of dialogue response timings. In *Proc. ACL*, pages 2442–2452.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale RNNs. In *Proc. ICMI*, pages 186–190.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, pages 1–27.

Yuki Sato, Yuya Chiba, and Ryuichiro Higashinaka. 2024. Effects of multiple Japanese datasets for training voice activity projection models. In *Proc. O-COCOSDA*, pages 313–318.

David Schlangen, Timo Baumann, Hendrik Buschmeier, Okko Buß, Stefan Kopp, Gabriel Skantze, and Ramin Yaghoubzadeh. 2010. Middleware for incremental processing in conversational agents. In *Proc. SIGDIAL*, pages 51–54.

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

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, pages 1–38.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. SpokenWOZ: a large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Proc. NeurIPS*, pages 39088–39118.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proc. SIGDIAL*, pages 220–230.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:1–26.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proc. EACL*, pages 745–753.

Barbara Wheatley, Masayo Kaneko, and Megumi Kobayashi. 1996. CALLHOME Japanese Speech, LDC96S37, Linguistic Data Consortium.