Analysis of Voice Activity Detection Errors in API-based Streaming ASR for Human-Robot Dialogue

Kenta Yamamoto, Ryu Takeda and Kazunori Komatani
SANKEN, Osaka University, Japan
{kentayamamoto,rtakeda,komatani}@sanken.osaka-u.ac.jp

Abstract

In human-robot dialogue systems, streaming automatic speech recognition (ASR) services (e.g., Google ASR) are often utilized, with the microphone positioned close to the robot's loudspeaker. Under these conditions, both the robot's and the user's utterances are captured, resulting in frequent failures to detect user speech. This study analyzes voice activity detection (VAD) errors by comparing results from such streaming ASR to those from standalone VAD models. Experiments conducted on three distinct dialogue datasets showed that streaming ASR tends to ignore user utterances immediately following system utterances. We discuss the underlying causes of these VAD errors and provide recommendations for improving VAD performance in human-robot dialogue.

1 Introduction

Several embodied robots capable of speech interaction have been developed (Minato et al., 2024; Inoue et al., 2016). In this situation, since such robots are embodied, automatic speech recognition (ASR) is performed without the user wearing a headset microphone, meaning the microphone is not close to the user's mouth. Instead, the robot usually uses its built-in loudspeaker and a microphone placed nearby.

With advancements in ASR technology, even researchers who do not specialize in ASR can easily use it. Several previous studies have compared the performance of off-the-shelf ASR, providing valuable information for researchers and developers outside the speech recognition community, e.g., (Georgila and Traum, 2024). One of the simplest ways to use ASR is through streaming-based ASRs accessed via APIs. By using these, ASRs do not need to be downloaded or installed in advance, making them very convenient. In this case, because the ASR is streaming-based, both voice activity detection (VAD) and speech recognition are performed simultaneously on the server side.

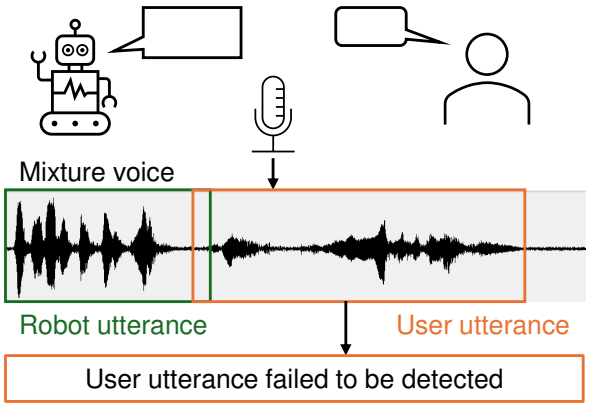

Figure 1: Robot fails to detect user utterance on dialogue

VAD (Atal and Rabiner, 1976) is a crucial front-tend technology for spoken dialogue systems. Its role is to detect the active speech segments from the input signals captured by microphones. VAD is primarily used to determine the boundaries of a user utterance, facilitating turn-taking in dialogue (Brady, 1965; Medennikov et al., 2020; Skantze, 2021). However, if the system misses the user utterance, causing the dialogue to break down. Thus, accurately detecting user speech segments is also important for ASR (Kingsbury et al., 2002; Novitasari et al., 2022). Errors in failing to detect user utterances are significant problems for both turn-taking and ASR and therefore must be avoided.

In this paper, we demonstrate situations where system fails to detect user utterances occur in robot dialogue. Specifically, we focus on scenarios where (1) a streaming-based ASR is used, and (2) the robot's loudspeaker and microphone are positioned in close proximity. In such cases, the robot's voice is also picked up by the microphone along with the user's voice. This leads to frequent failures in VAD when using streaming-based ASR trained on single-speaker data, as depicted in Figure 1. VAD

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 245–253
May 27–30, 2025.

error result in the robot ignoring the user utterance. This causes the robot to remain silent without taking its turn, making it difficult for general users to understand the robot's status, leading to significant frustration. Therefore, our research aims to provide insights for human-robot interaction (HRI) researchers and practitioners to achieve robust VAD.

This study addresses the following two research questions:

(RQ1) To what extent do commonly used tools, such as Google APIs employed in constructing spoken dialogue systems, ignore user utterances?

(RQ2) Under what circumstances are user utterances more likely to be ignored?

To answer these research questions, we conducted multiple analyses. As experimental validation, we demonstrate VAD error (i.e., failures of VAD to detect user utterances) in a streaming-based ASR system using three datasets with different microphone configurations. We also compare the performance to a case where VAD is performed separately. On the basis of these results, we discuss the appropriate system configurations for conveniently building conversational robots.

2 Related works

Several studies have analyzed the performance and errors of publicly available ASRs. As ASR can be increasingly used by non-experts, it is useful to analyze the performance and issues of various ASRs. Pasandi and Pasandi (2016) showed that the Google API's ASR tends to ignore disfluencies, i.e., non-fluent words. Addlesee et al. (2020) compared the ASRs of Microsoft, IBM, and Google, investigating their robustness against disfluencies and overlaps. Georgila et al. (2020) analyzed the characteristics of several ASRs, demonstrating that performance degrades in domains requiring specialized vocabulary, as well as in noisy environments. Georgila and Traum (2024) investigated the impact of accents, i.e., various native and non-tative accents in English, on the performance of off-the-shelf ASRs. The Whisper model was found to be particularly sensitive to variations in accents.

These studies focus on analyzing ASR performance. However, it is also helpful to analyze the causes, especially in situations involving interaction with embodied robots. Error analysis in a convenient streaming-based ASR, rather than in segmented speech files, is also helpful. Particularly, a failure in VAD causes more damage to the dialogue than simple ASR errors, as it means the system does not recognize that the user has spoken. Therefore, we focus on VAD and examine situations when its errors occur in dialogues with embodied robots, where a robot utterance can be picked up by the microphone along with the user utterance.

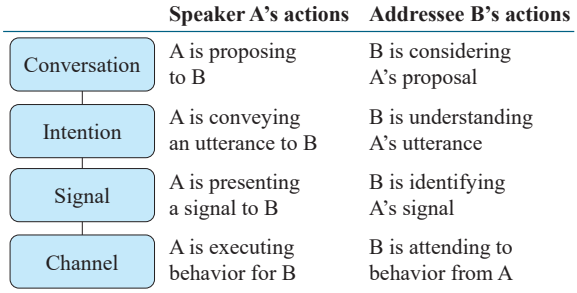

Figure 2: Action ladders (Clark, 1996)

A Model that predicts future user utterance segments, rather than detecting the end of the user utterance segment, was proposed. Voice activity projection (VAP) (Ekstedt and Skantze, 2022) is a model that predicts the future voice activity of two speakers, on the basis of raw audio input. The model requires two channels of recorded speech as input. Since our study targets a situation in which a dialogue system is used in a simple recording environment, and thus targets speech segment detection on the basis of one-channel recorded speech, VAP is not included in the comparison in this study.

3 Target of Analysis

VAD is a crucial component in spoken dialogue systems. According to the action ladder (Clark, 1996) illustrated in Fig. 2, joint actions between interlocutors at the signal level are established when the addressee successfully identifies the signal emitted by the speaker. On the basis of this, joint actions at higher levels, such as the Intention level and Conversation level, can be established.

Errors in VAD lead to two main issues in spoken dialogue systems:

1. Joint actions at the Intention Level (or higher levels) cannot be established due to missed voice activity segments.

2. Joint actions fail to form even at the Signal level.

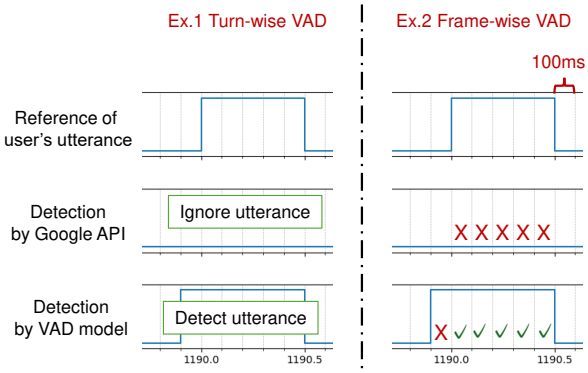

Figure 3: Calculation method of VAD accuracy in this study

The former issue refers to speech recognition errors caused by incorrect speech segmentation, thus hindering joint actions. This is an issue known in speech recognition.

The latter issue is more serious in spoken dialogue systems, especially for robot dialogue systems that interact with a wide range of users. If the system fails to identify that the user has spoken, it cannot take any action in response, leading to the user utterance being ignored. In speech communication, failing to respond can greatly reduce the user's motivation to continue the dialogue, making this problem more severe than simply having incorrect recognition results.

Corresponding to these two issues, we will analyze VAD performance both frame-wise and turn-wise, as illustrated in Fig. 3.

(Ex.1) Turn-wise VAD error rate: We calculated the percentage of missed detection across all of the user's turns.

(Ex.2) Frame-wise VAD scores: We calculated detection accuracy every 100 milliseconds.

Furthermore, we examined the extent to which speech recognition accuracy improves due to speech segment detection.

4 Data and Models

This section describes the dataset and VAD models used for the analysis in this study.

4.1 Dataset

In this experiment, we utilize three datasets (Table 1). In our analysis, we evaluate the accuracy of speech segment detection for user utterances at both the turn and frame levels. Therefore, as a reference for the amount of data used in the evaluation, we present the number of dialogues, the number of utterances, and the total utterance duration in the dataset. Both datasets have annotations for system speech segments and user speech segments.

Hazumi1911 [1] This is open Japanese dialogue data. The dataset consists of conversational data recorded in a laboratory setting with a microphone placed in front of a robot. In this dataset, the system and the user had an about 15-minute chit-chat about diverse topics (Komatani and Okada, 2021). Hazumi1911 contains face-to-face conversations. The appearance of the dialogue system is that of a 2D Computer graphics (CG) agent.

Hazumi2010 [2] Like Hazumi1911, this dataset is publicly available and was recorded under nearly identical conditions. The main difference is that Hazumi2010 comprises remote dialogues conducted online.

Avatar Fes. This dataset includes dialogue data recorded using a dialogue robot in a real-world environment. The dialogue system was implemented on a small robot that engaged in 3-minute attentive listening dialogues (Inoue et al., 2020). Recordings were made of participants interacting with the dialogue system at a trial event, the Avatar Festival. Since the recordings took place in an event hall, there is a significant amount of background noise. Furthermore, participants were not always fully engaged with the dialogue system, and there were instances where third parties spoke to them. Such data reflects dialogues under conditions representative of actual usage scenarios of the dialogue system. Dialogues in which participants terminated the interaction prematurely were excluded from the dataset.

The datasets used in this study encompass a variety of environments. Hazumi1911 and Hazumi2010 contain dialogue datasets collected in laboratory environments, whereas the Avatar Fes. dataset contains data from real-world usage scenarios. One key difference among the dialogues in these datasets is the length of turn-taking; therefore,

[1] https://github.com/ouktlab/Hazumi1911
[2] https://github.com/ouktlab/Hazumi2010

Table 1: Dataset details

Dataset	Number of dialogues	Average number of turns	Average of user utterance duration [second]
Hazumi1911	23	83.5	866.9
Hazumi2010	33	74.1	704.9
Avatar Fes.	138	30.4	81.2

we present basic statistics on turn-taking durations. This is significant because the difficulty in distinguishing between system and user speech during VAD depends on the length of turn-taking. The distribution of lengths of silence between speaker turn for the Hazumi1911 and Hazumi2010 datasets is shown in Figure 4. We calculated the distribution of the intervals from the end of the system utterance to the beginning of the user utterance for all utterances except the backchannel. In Hazumi1911, the distribution ranges from negative values (indicating overlap) to instances where the user takes a long time to respond. In contrast, in Hazumi2010, user speech is concentrated after the end of system utterance (0s). This may be due to the fact that Hazumi1911 involves face-to-face interactions, while Hazumi2010 consists of online interactions. In online dialogues, participants tend to wait until the system has finished speaking before they respond, which may reduce the likelihood of recognition errors occurring in speech immediately following the system utterance. In Avatar Fes, the length of turn-taking varies greatly. This dataset consists of real-world dialogues with various background noises and is not necessarily limited to one-to-one conversations, as there may be interruptions from other speakers.

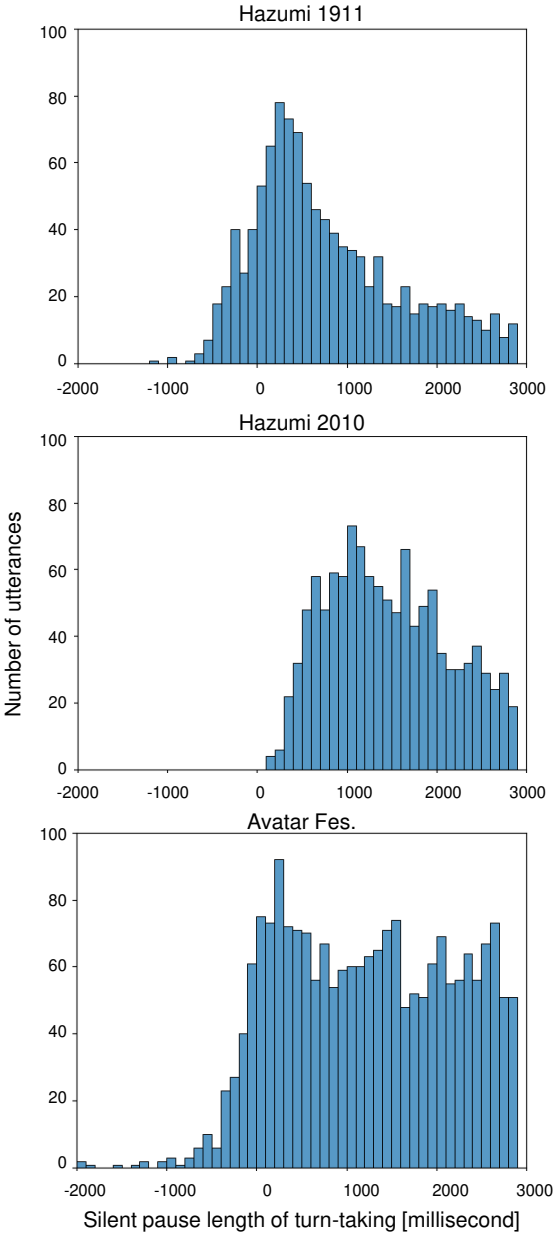

Figure 4: Histogram of silent pause length (System to User) in each dataset

4.2 Compared models

This section describes the models used in this experiment. We used the Google Speech Recognition API and, for comparison purposes, two publicly available models specialized in VAD. To perform analyses that assume typical usage scenarios, we selected models that are user-friendly and readily accessible.

Google ASR API The first model is a streaming-based ASR, **Google Speech-to-Text**[3]. We used the default model, accessing it via an API from Python for ASR. This model provides the start and end times for each recognized word. Other ASR APIs, such as Whisper, are also available. However, since Whisper requires a separate VAD setup, which might be difficult for practitioners to use easily, we chose to use Google ASR for this study.

Pyadin (VAD) The second model is a pre-trained

[3]https://cloud.google.com/speech-to-text?hl=en

VAD model based on DNN-HMM[4] (Takeda and Komatani, 2024). This model is a hybrid model of a hidden Markov model (HMM) and deep neural network (DNN) using a transformer-encoder. This model was trained on diverse datasets encompassing various environmental conditions. It remains robust against variations in signal amplitudes and speech distortions.

Silero VAD The third model is the VAD model based on long short-term memory (LSTM)[5]. Despite its low computational complexity and suitability for real-time processing, this model achieves higher accuracy than power-based VAD models. Furthermore, since the model is publicly available, it can be easily tested. In this study, we used it as a baseline model for model-based VAD.

5 Results of Analyses

The analysis results for Ex.1 and Ex.2 are summarized in Section 5.1. More detailed frame-level analyses are conducted in Sections 5.2 and 5.3. In Section 5.2, we analyze the impact of speech segment detection errors on speech recognition. In Section 5.3, we investigate at which points during user utterances the detection of speech segments fails.

5.1 VAD accuracy by each model

Dialogue systems can identify when they are speaking. Therefore, in this experiment, we exclude the detection results for the system's voice segments to calculate the outcomes.

We analyzed cases where the user utterance was entirely ignored. Table 2 shows the number of exchanges in which no user utterance was detected.

Chi-square tests conducted on 2×3 contingency tables for each dataset (Hazumi1911, Hazumi2010, Avatar Fes.) showed significant differences in user speech ignoring rates among the three models (Google, Pyadin, Silero; $p < 0.01$). Subsequent pairwise comparisons using Fisher's exact test with Bonferroni correction indicated that, for Hazumi1911 and Avatar Fes., all model pairs differed significantly. In Hazumi1911, significant differences were found between Google and Pyadin ($p < 0.01$) as well as between Pyadin and Silero ($p < 0.01$), but not between Google and Silero. Overall, these results suggest that under Hazumi1911 and Avatar Fes. conditions, the Google model's ignoring rate was notably higher than those of the other two models. Chi-square tests conducted on 2×3 contingency tables for each dataset (Hazumi1911, Hazumi2010, Avatar Fes.) showed significant differences in user speech ignoring rates among the three models (Google, Pyadin, Silero; $p < 0.01$). Subsequent pairwise comparisons using Fisher's exact test with Bonferroni correction indicated that, for Hazumi1911 and Avatar Fes., all model pairs differed significantly. In Hazumi1911, significant differences were found between Google and Pyadin ($p < 0.01$) as well as between Pyadin and Silero ($p < 0.01$), but not between Google and Silero. Overall, these results suggest that under Hazumi1911 and Avatar Fes. conditions, the Google model's ignoring rate was notably higher than those of the other two models.

The accuracy of speech segment detection at 100 ms intervals is shown in Table 3. The experimental results show that using VAD yields higher detection accuracy across all datasets.

Table 2: Turn-wise VAD scores: The number of ignored user turns (**Ex.1**)

Data	Model	Ignored / Total
Hazumi1911	Google	319/1920 (17%)
	Pyadin	39/1920 (2%)
	Silero	99/1920 (5%)
Hazumi2010	Google	69/2446 (3%)
	Pyadin	5/2446 (0.2%)
	Silero	69/2446 (3%)
Avatar Fes.	Google	1582/4449 (36%)
	Pyadin	48/4449 (1%)
	Silero	322/4449 (7%)

Table 3: Frame-wise VAD scores (**Ex.2**)

Data	Model	Pre	Rec	F-value
Hazumi1911	Google	0.73	0.73	0.73
	Pyadin	0.87	0.89	0.88
	Silero	0.88	0.89	0.88
Hazumi2010	Google	0.53	0.88	0.64
	Pyadin	0.79	0.97	0.86
	Silero	0.84	0.92	0.87
Avatar Fes.	Google	0.61	0.56	0.47
	Pyadin	0.58	0.95	0.71
	Silero	0.67	0.85	0.74

[4]https://github.com/ouktlab/pyadintool
[5]https://github.com/snakers4/silero-vad

Table 4: ASR results for each dataset using different segmentation methods for user utterances: ASR is performed using Google ASR in all conditions.

Data	Segmentation	CER
Hazumi1911	Only ASR	0.55
	VAD (Pyadin) ⇒ ASR	0.50
	Manual ⇒ ASR	**0.42**
Hazumi2010	Only ASR	0.54
	VAD (Pyadin) ⇒ ASR	0.51
	Manual ⇒ ASR	**0.40**
Avatar Fes.	Only ASR	0.65
	VAD (Pyadin) ⇒ ASR	0.43
	Manual ⇒ ASR	**0.42**

5.2 Impacts of VAD errors on ASR

The following experiment analyzes the impact of speech segment detection on speech recognition accuracy. We compared speech recognition accuracy under three conditions, using Google Speech-to-Text for all conditions:

Only ASR We input the entire dialogue audio into Google ASR.

VAD (Pyadin) ⇒ ASR We split into audio files for each user utterance using VAD. We used Google ASR on those audio files.

Manual ⇒ ASR We split into audio files for each user utterance on the basis of manually annotated boundaries. We used Google ASR on those audio files.

Table 4 shows the differences in speech recognition accuracy in each condition. The character error rate (CER) in the Manual condition is the smallest in each dataset. The results show that using VAD can reduce ASR errors due to streaming-based ASR. Also, results in the Google condition have more errors than those in the VAD condition. These results make it clear that errors in VAD within Google ASR significantly affect ASR accuracy.

5.3 VAD error trends in Google ASR

We examined the patterns of VAD errors in Google ASR. We investigated the relationship between the time to user utterance after system utterance and VAD error rates. The results of this analysis are shown in Figure 5. For each start time (in 100 ms increments) after the end of system utterance, we assessed whether user utterance at that timing was detected. The results indicate the percentage of user utterance that was not detected as utterance for each frame.

In all data, there is a high frequency of detection errors immediately following system utterances. User utterance immediately after system speech may be misrecognized as noise and not detected as part of the speech segment. For Google ASR, only one of the voices may be identified as a voice when multiple voices are entered in a certain interval. In this case, the system speech is prioritized and the following user utterance is ignored. This phenomenon may occur in models trained with the assumption of a single speaker.

6 Discussion and Recommendations

6.1 Discussion

From the results in Section 5.1, we found that Google ASR API often ignores the user's turn, which is a very serious problem for spoken dialogue systems. We also found that this issue can be mitigated by using a separately trained VAD model. However, when examining the frame-level analysis results, we observed that the accuracy varies significantly depending on the environment, confirming that VAD tends to fail more easily in noisy environments. As shown in Table 3, even in low-noise environments, the precision of Hazumi2010 is lower than that of Hazumi1911, while its recall is higher. This is likely due to the fact, as illustrated in Figure 4, that Hazumi2010 exhibits delayed turn-taking by users, resulting in fewer overlaps with system utterances. Consequently, the likelihood of missed detections is reduced, affecting observed recall and precision metrics.

From the results presented in Section 5.2, we found that ASR accuracy decreases when there are many errors in VAD. This trend is consistent with those reported in previous studies and has been reconfirmed in the present research.

In Section 5.3, we analyzed the timing in which utterances are ignored and found a strong tendency for user utterances immediately following system utterances to be ignored. This tendency is particularly pronounced when using the Google ASR API. Specific examples of errors are shown in Figure 6. These examples were sampled from dialogues in the Avatar Fes. dataset, where detection errors were frequently observed. In Example 1, Google ASR fails to detect the user utterance immediately following the system's question, making it impos-

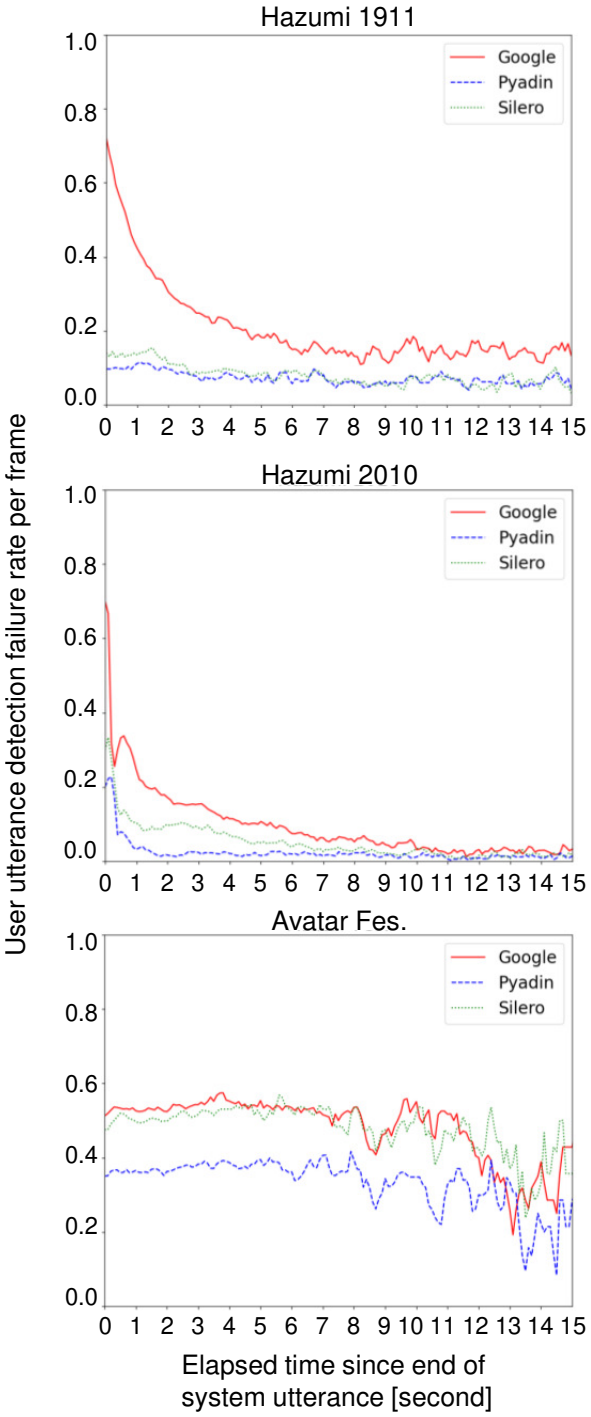

Figure 5: The relationship between user utterance timing and the detection failure rate per frame (100 ms)

sible to determine whether the user responded to the system's question. However, standalone VAD models successfully detected this utterance. This is thought to be due to confusion between user and system utterances, possibly stemming from the architecture of the speech recognition model. The Transformer model has a fixed input window and may not be able to recognize speaker differences within that window. Therefore, it may not be able to distinguish user speech immediately after system speech from the system speech itself.

One method to address this problem is to control turn-taking so that user utterances do not overlap with system utterances. Using the robot's movements, we can adjust aspects such as the timing of the user utterance. For instance, before transferring the turn to the user, the robot can look at the user's face; or during the system utterance, it can make large gestures to prevent the user from speaking simultaneously. In this way, by adjusting the user's speaking timing, we can avoid the user utterance overlapping with the system utterances.

Other factors, such as background noise and robot operation sounds, can also cause VAD errors. In Example 2, we present a case where all three VAD models failed to detect the user utterance. This failure occurred due to loud background noise, such as footsteps and laughter, being captured by the microphone in the middle of the user utterance, preventing proper detection. In noisy environments like this, the system frequently misses user utterances. In the case of robots that control gestures, the sounds generated by the robot's movements may also cause errors in VAD and ASR (Nishimura et al., 2006; Ince et al., 2011). Therefore, it is necessary to implement background noise suppression and minimize the robot's operational sounds during user utterances.

6.2 Recommendations

On the basis of the findings of this study, we offer the following recommendations to HRI researchers and practitioners. These suggestions aim to enhance the accurate recognition of user utterances, especially in situations where implementing an advanced speech processing environment is not feasible. Specifically, we propose methods to reduce omissions in the VAD of user utterances when developing dialogue systems.

Employing VAD Model The VAD feature included with Google ASR tends to miss user utterances. To address this issue, we recommend using an independent VAD model. As demonstrated by our experimental results in Section 5, errors can be significantly reduced by employing individually trained VAD models.

Using Separate Microphones Our experiments have shown that the system utterances and user utterances need to be properly separated

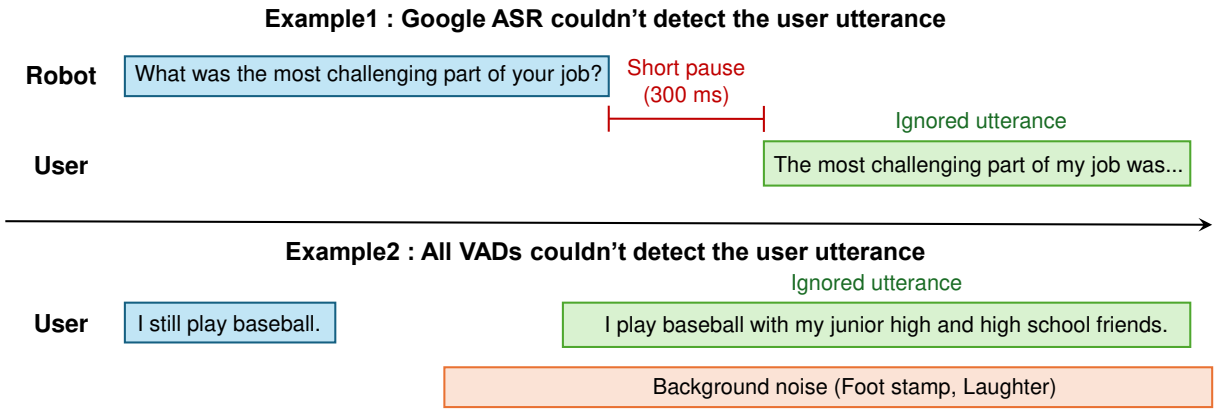

Figure 6: Example of VAD failures in detecting user utterances in Avatar Fes. dataset

to suppress error. Therefore, we advise configuring the microphone placement to ensure that system speech and user speech are not confused in the audio inputs. The most effective method is to provide the user with a handheld microphone. If this is not possible, the microphone should be positioned away from the system's speakers to minimize interference.

Understanding Error Trends in VAD VAD tends to fail in detecting the beginning of user utterances. This suggests that the initial part of the user utterance may not be recognized. Robust dialogue processing is therefore crucial, taking into account the possibility of missed detections of user responses, especially those immediately following system utterances. Such robust handling can ensure more reliable dialogue system performance even when some user utterances are not initially detected.

7 Conclusion

In this study, we analyzed the patterns of speech segment detection errors in typical speech recognition scenarios involving dialogue robots. When utilizing streaming-based automatic speech recognition (ASR) systems, such as Google API, in environments where both system and user speech are input, we observed instances where user utterance was not detected. Through an error analysis of voice activity detection (VAD) on dialogue data between a dialogue system and users, we clarified the tendencies of missed detections.

On the basis of our analysis, we made the following two contributions:

Answer to RQ1 We observed that the accuracy of Google ASR declines as the operational environment approaches real-world conditions. However, by integrating a dedicated VAD model, we can effectively prevent the system from disregarding the user's turn.

Answer to RQ2 Our findings indicate that user utterances occurring immediately after system utterances are prone to being overlooked. Therefore, it is important to ensure that user and system utterances do not overlap to prevent missing user inputs.

To mitigate such errors, a speech segment detection model needs to be used preprocess ASR. Alternatively, ensuring that system speech is not captured by the microphone during ASR in robot dialogues is crucial. Additionally, encouraging users to wait briefly after the system finishes speaking before responding may also be effective.

To achieve more robust turn-taking, errors in turn-taking, such as those involving voice activity detection (VAP), neecd to be examined in future studies. We hope that our findings will contribute to improving the performance of dialogue robots and enhancing the user experience in human-robot interactions.

Acknowledgement

This work was supported by JST Moonshot R&D Grant Number JPMJPS2011.

References

Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020. A comprehensive evaluation of incremental speech recognition and diarization for conversational ai. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 3492–3503.

Bishnu S Atal and Lawrence Rabiner. 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):201–212.

Paul T. Brady. 1965. A technique for investigating on-off patterns of speech. *The Bell System Technical Journal*, 44(1):1–22.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Interspeech*, pages 5190–5194.

Kallirroi Georgila, Anton Leuski, Volodymyr Yanov, and David Traum. 2020. Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 6469–6476.

Kallirroi Georgila and David Traum. 2024. Evaluation of off-the-shelf whisper models for speech recognition across diverse dialogue domains. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Gökhan Ince, Keisuke Nakamura, Futoshi Asano, Hirofumi Nakajima, and Kazuhiro Nakadai. 2011. Assessment of general applicability of ego noise estimation - applications to automatic speech recognition and sound source localization -. In *International Conference on Robotics and Automation (ICRA)*, pages 3517–3522.

Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*, pages 118–127.

Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with ERICA, an autonomous android. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*, pages 212–215, Los Angeles. Association for Computational Linguistics.

Brian Kingsbury, George Saon, Lidia Mangu, Mukund Padmanabhan, and Ruhi Sarikaya. 2002. Robust speech recognition in noisy environments: The 2001 ibm spine evaluation system. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–53–I–56.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko. 2020. Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario. In *INTERSPEECH*, pages 274–278.

Takashi Minato, Ryuichiro Higashinaka, Kurima Sakai, Tomo Funayama, Hiromitsu Nishizaki, and Takayuki Nagai. 2024. Overview of dialogue robot competition 2023. In *Proceedings of Dialogue Robot Competition 2023*.

Yoshitaka Nishimura, Mikio Nakano, Kazuhiro Nakadai, Hiroshi Tsujino, and Mitsuru Ishizuka. 2006. Speech recognition for a robot under its motor noises by selective application of missing feature theory and mllr. In *Statistical and Perceptual Audio Processing (SAPA)*, pages 53–58.

Sashi Novitasari, Takashi Fukuda, and Gakuto Kurata. 2022. Improving ASR Robustness in Noisy Condition Through VAD Integration. In *Interspeech*, pages 3784–3788.

Hannaneh B. Pasandi and Haniyeh B. Pasandi. 2016. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

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

Ryu Takeda and Kazunori Komatani. 2024. Scale-invariant online voice activity detection under various environments. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conerence (APSIPA ASC)*.