Yeah, *Un*, *Oh*: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection

Koji Inoue, Divesh Lala, Gabriel Skantze[*], Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan,
[*]KTH Royal Institute of Technology, Sweden
Correspondence: inoue@sap.ist.i.kyoto-u.ac.jp

Abstract

In human conversations, short backchannel utterances such as “yeah” and “oh” play a crucial role in facilitating smooth and engaging dialogue. These backchannels signal attentiveness and understanding without interrupting the speaker, making their accurate prediction essential for creating more natural conversational agents. This paper proposes a novel method for real-time, continuous backchannel prediction using a fine-tuned Voice Activity Projection (VAP) model. While existing approaches have relied on turn-based or artificially balanced datasets, our approach predicts both the timing and type of backchannels in a continuous and frame-wise manner on unbalanced, real-world datasets. We first pre-train the VAP model on a general dialogue corpus to capture conversational dynamics and then fine-tune it on a specialized dataset focused on backchannel behavior. Experimental results demonstrate that our model outperforms baseline methods in both timing and type prediction tasks, achieving robust performance in real-time environments. This research offers a promising step toward more responsive and human-like dialogue systems, with implications for interactive spoken dialogue applications such as virtual assistants and robots.

1 Introduction

In natural human conversations, short backchannels, such as “*yeah*” and “*right*,” play a crucial role in facilitating smooth and engaging interactions (Clark, 1996; Clancy et al., 1996; Maddrell and Watson, 2012). They function as feedback mechanisms, signaling attentiveness, understanding, and agreement without interrupting the speaker. Accurate prediction and generation of backchannels in spoken dialogue systems are essential for creating more natural and human-like interactions (Schroder et al., 2011; DeVault et al., 2014; Inoue et al., 2020b). Although some definitions of backchannels include longer and more linguistic tokens such as “*I see*,” this work focuses on short tokens that are frequently and dynamically used by listeners.

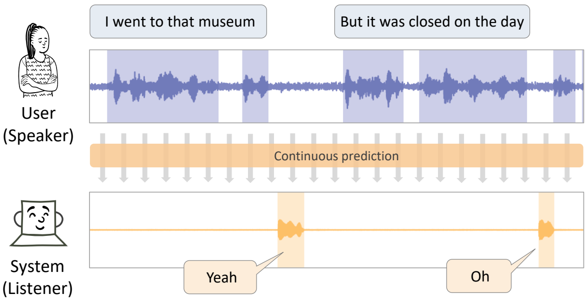

Figure 1: Conceptual diagram of continuous backchannel prediction

Backchannel modeling remains a significant challenge due to their subtle and context-dependent characteristics. Given the dynamic nature of backchannels, it is essential to predict them on a frame-by-frame basis and in real-time for live spoken dialogue systems, as depicted in Figure 1. However, most previous studies have focused on utterance-based systems, or, in the case of frame-based systems, have artificially balanced the test data by reducing non-backchannel samples. This data manipulation introduces a discrepancy between the training models and real-world systems. Consequently, for practical applications, it is necessary to develop models capable of real-time, continuous frame-wise prediction and evaluate them using unbalanced, real-world datasets.

Transformer-based architectures have emerged as powerful tools for a broad range of sequential prediction tasks, such as language modeling and speech recognition. Among these, the Voice Activity Projection (VAP) model, a Transformer-based architecture, has shown its efficacy in predicting future voice activity within dialogues (Ekstedt and Skantze, 2022b,a). Since voice activity is closely intertwined with turn-taking dynam-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7171–7181
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

ics and backchannel behaviors, the VAP model seems to have the potential to enable more accurate backchannel prediction. Furthermore, as a previous study implemented the real-time VAP model (Inoue et al., 2024b), a VAP-based backchannel prediction model could serve as a promising candidate for real-time backchannel generation systems. In its initial application, the VAP model was employed for backchannel prediction in a zero-shot manner. However, this approach exhibited a sensitivity to threshold selection and depended on a balanced dataset. These findings suggest that while the VAP model shows promise for backchannel prediction, further research is necessary to refine its training methodology for improved performance.

In this study, we propose a novel approach to backchannel prediction by utilizing the VAP model as a foundational framework. We first train the VAP model on a large corpus of general dialogue data to capture the fundamental patterns of conversational dynamics. Subsequently, we fine-tune the model on a specialized dataset focused on backchannel prediction. This two-stage training process is analogous to the pre-training and fine-tuning paradigm employed by models like BERT, aiming to demonstrate the VAP model's versatility as a general-purpose base model. Moreover, to the best of our knowledge, our model is the first to predict both the timing and type of backchannels in a continuous and real-time manner.

The contributions of this paper are twofold.

- **Real-Time Continuous Backchannel Prediction**: A method for real-time, continuous backchannel prediction based on the VAP model is developed and evaluated on real-world, unbalanced test data.
- **Two-Stage Training for Generalization**: A two-stage training process is introduced for the VAP model, demonstrating its potential as a fundamental model for predicting conversational dynamics.

Note that the source codes and trained models are publicly available [1].

2 Related Work

Effective backchannel generation necessitates accurate prediction of three key elements: temporal placement (timing), linguistic form (type), and prosodic patterns. The majority of prior studies have focused exclusively on predicting the timing and types of backchannels. The definition and functions of backchannel types have been explored in conversational analysis and linguistic studies (Drummond and Hopper, 1993; Wong and Peters, 2007; Tang and Zhang, 2009; Den et al., 2011). Despite the critical role of prosody in entrainment, existing research on this remains scarce (Kawahara et al., 2015; Ochi et al., 2024). This review primarily summarizes the current state of research on predicting the timing and form of backchannels.

Before recent advances in machine learning technologies, backchannel prediction models were primarily based on hand-crafted features with heuristic rules or simpler models (Koiso et al., 1998; Ward and Tsukahara, 2000; Fujie et al., 2005; Morency et al., 2008, 2010; Ozkan and Morency, 2011; Blache et al., 2020). With the advent of dataset creation paradigms and machine learning advancements, deep learning models began to be employed for backchannel prediction. Initial models were built using long-short-term memory (LSTM) networks (Ruede et al., 2017a,b; Adiba et al., 2021; Jain et al., 2021), while more recent approaches leverage Transformer-based models (Jang et al., 2021; Liermann et al., 2023).

Most previous studies focused on timing prediction, with some also addressing type prediction. The most conventional approach involves framing the prediction task as a three-class classification problem: non-backchannel, continuer, and assessment (Choi et al., 2024), as described in Section 4.2. The present work adopts this three-class classification scheme for both timing and type prediction. Other research has explored a four-type classification system, encompassing non-backchannel, continuer, understanding, and empathy backchannels (Jang et al., 2021). Another study introduced a five-type classification for single continuer, double continuers (e.g., "yeah yeah"), triple continuers, assessment, and non-backchannel (Kawahara et al., 2016). Furthermore, a different approach proposed a two-step classification method where the first model predicts the timing, followed by a second model that determines the type (Adiba et al., 2021).

In terms of prediction unit, utterance-based or continuous, utterance-based models tend to incorporate linguistic features such as word embeddings (Jang et al., 2021; Park et al., 2024). Conversely, previous continuous models were generally restricted to using prosodic features (Ruede et al.,

[1] We will put the link here after the reviewing process.

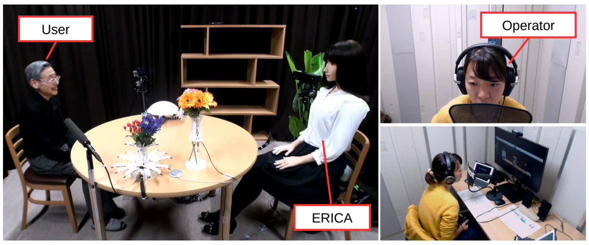

Figure 2: Setup for dialogue recording

2017a,b). Recent models have begun utilizing audio encoders that can theoretically capture both linguistic and prosodic information in an end-to-end manner (Park et al., 2024; Choi et al., 2024). The VAP model used in this study similarly employs a pre-trained contrastive predictive coding (CPC) model as its audio encoder.

To address the issue of imbalanced data, recent studies have integrated multi-task learning with the primary task of backchannel prediction. For instance, subtasks such as turn-taking prediction (Hara et al., 2018; Ishii et al., 2021), sentiment score analysis (Jang et al., 2021), dialogue act recognition (Liermann et al., 2023), and streaming automatic speech recognition (ASR) (Choi et al., 2024) have been considered. Notably, a recent model (Choi et al., 2024) was evaluated on a real-world imbalanced dataset, demonstrating reasonable performance with F1-scores of 26% and 22% for continuer and assessment backchannels, respectively, in a frame-wise manner. This work also proposes the incorporation of multi-task learning for both backchannel prediction and VAP tasks.

3 Dataset

We employ two types of datasets: one specialized for backchannel prediction and the other for pre-training the proposed model. Note that all the dialogue datasets were in Japanese.

3.1 Attentive Listening Dataset

We have collected spoken dialogue data using a Wizard-of-Oz (WOZ) setup. In this experiment, the android ERICA (Inoue et al., 2016) was employed, with a human operator remotely controlling it, which was transmitted and played through ERICA's speaker system (Figure 2).

The dialogue task focused on attentive listening, where human participants (speaker) shared personal experiences, and ERICA actively engaged as a listener (Inoue et al., 2020b; Lala et al., 2017). This task was advantageous because it allowed for the collection of numerous backchannel responses by ERICA. The participants comprised two demographic groups: students and older adults. Each group was provided with a prompt to guide their conversation; for instance, students discussed "challenges during the COVID-19 pandemic," while the elderly participants reflected on "memorable travel experiences and recent favorite meals."

ERICA's operators were three actresses, who had experiences of our past attentive listening dialogue experiments. While backchannel behaviors can be subjective and vary among individuals, the few operators were selected for their experience, ensuring the collection of high-quality backchannel data. Furthermore, the operators participated in a sufficient number of dialogue sessions in this experiment, ensuring both the quantity and quality of the training data, despite the varying speakers in each session.

In total, we recorded 109 dialogue sessions, each lasting approximately 7 to 8 minutes. The data were randomly divided into 87, 11, and 11 sessions for training, validation, and testing purposes, respectively. We subsequently transcribed the dialogues and annotated ERICA's backchannel responses.

3.2 Pre-training Data for VAP

In this study, we introduce a two-step training approach where the original VAP model is initially trained, followed by fine-tuning for backchannel prediction. The first step requires a larger dataset to train the VAP model effectively to continuously predict future voice activities. To support this, in addition to the attentive listening dialogue dataset mentioned earlier, we incorporated additional training data at this stage. Using the same configuration as ERICA, we recorded data across various scenarios, such as job interviews (Inoue et al., 2020a) and first-meeting dialogues (Inoue et al., 2022). These diverse tasks provide different dialogue styles, enhancing the VAP model's robustness and enabling it to adapt to various behaviors, including backchannels. In total, the pre-training data amounted to about 35 hours, which includes the training set from the aforementioned backchannel prediction dataset.

4 Task Definition

In this work, we address two distinct backchannel prediction tasks as outlined below.

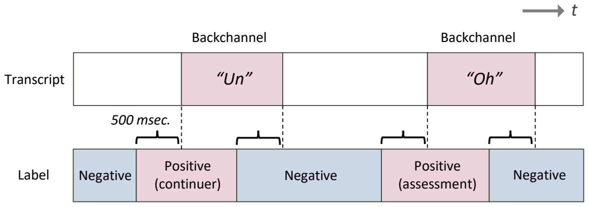

Figure 3: Definition of positive (backchannel) and negative (non-backchannel) frames

4.1 Timing Prediction

The primary objective of this task is to predict the occurrence of a backchannel, framing it as a binary classification problem. We manually annotated the backchannels in the aforementioned dialogue dataset, identifying two distinct types of short tokens as backchannels: continuer and assessment. The continuer tokens include expressions such as "*un*" and "*hai*" in Japanese, which correspond to "*yeah*" and "*right*" in English. On the other hand, the assessment tokens include utterances such as "*he-*" and "*oh*" in Japanese, equivalent to "*wow*" and "*oh*" in English. It is important to note that in the current task, we do not differentiate between these two token types, whereas such a distinction is made in the second task. To facilitate the implementation of the model in real-time spoken dialogue systems, we marked the positive sample frames as occurring 500 milliseconds before the actual backchannel utterances, as illustrated in Figure 3.

The total number of annotated backchannel utterances amounted to 13,601, with a cumulative duration of 5,912.6 seconds. These were split into 11,371 utterances for training, 1,139 for validation, and 1,091 for testing. For the negative samples, the cumulative time of non-backchannel segments was 56,467.3 seconds, resulting in a ratio of approximately 10% for positive samples.

4.2 Timing and Type Prediction

In the second task, the prediction process becomes more refined by incorporating different types of backchannels. Although numerous definitions of backchannel types or categories exist in prior research, we adopt the two basic types: continuers and assessments, as defined earlier. This distinction is crucial for conveying different listener intentions, and most previous studies have primarily addressed continuers, as assessment backchannels may require comprehension of both the prosodic and linguistic aspects of the user's utterances.

After reclassifying the previously annotated backchannels into these two categories, we found that there were 10,081 instances of continuers and 3,506 instances of assessment backchannels. This ratio means that the prediction of assessments seems to be more difficult than those for continuers. It is important to note that 14 tokens could not be classified into either category, and thus they were excluded from this task. Consequently, the classification problem in this task becomes a three-class classification: continuers, assessments, and non-backchannels. The definition of timing remains consistent with the previous task, as illustrated in Figure 3.

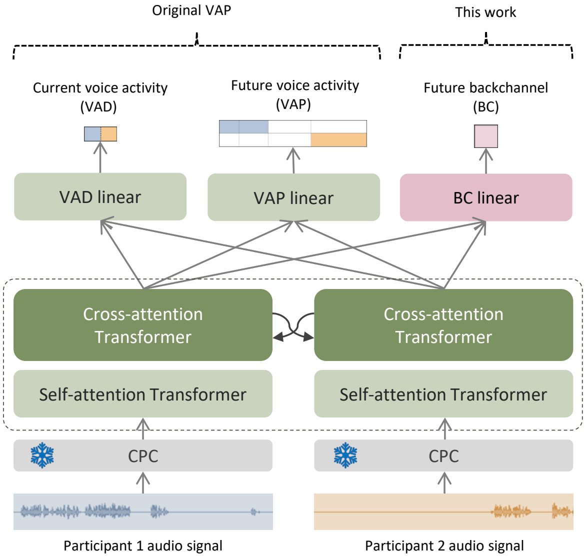
Original VAP
This work
Current voice activity (VAD)
Future voice activity (VAP)
Future backchannel (BC)
VAD linear
VAP linear
BC linear
Cross-attention Transformer
Cross-attention Transformer
Self-attention Transformer
Self-attention Transformer
CPC
CPC
Participant 1 audio signal
Participant 2 audio signal

Figure 4: Architecture of the VAP model

5 Proposed Method

In this section, we begin by explaining the voice activity projection (VAP) model, which serves as the foundational model and is trained to predict future voice activities by using the largest spoken dialogue dataset explained in Section 3.2. Following that, we discuss how to adapt a pre-trained VAP model for use in the current backchannel prediction task, using the data introduced in Section 3.1.

5.1 Voice Activity Projection

The VAP model employed in this study is constructed upon a Transformer-based architecture designed to emulate human-like predictive capabilities. As illustrated in Figure 4, the architecture of the VAP model processes stereo audio signals originating from two participants[2], consistent with

[2] A detailed explanation of the model's architecture is provided in a previous work (Inoue et al., 2024a)

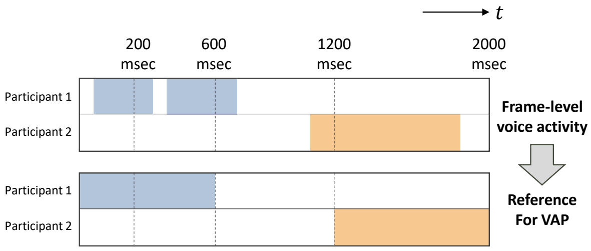

Figure 5: Definition of the VAP state

the operation of full-duplex spoken dialogue systems. Note that this model integrates the listener's audio as one of the input channels. This allows self-generated backchannels to be fed back into the model, distinguishing it from other existing models. This functionality plays a crucial role in preventing multiple consecutive backchannels, which may appear unnatural.

Each audio channel is initially processed independently through a Contrastive Predictive Coding (CPC) audio encoder and a channel-wise Transformer. The CPC was pre-trained with the Librispeech dataset (Riviere et al., 2020) and is frozen during the VAP training. The resulting outputs are then input into a cross-attention Transformer, where one channel serves as the query, while the other functions as the key and value. The output, a concatenation of both channels, produces a 512-dimensional vector. Note that we used the same parameters as defined in the original work (Inoue et al., 2024a) where the numbers of layers for the channel-wise and the cross-attention Transformers were 1 and 3, and the number of attention heads was 4, respectively.

The concatenated vector is subsequently processed by linear layers for two distinct tasks: voice activity projection (VAP) and voice activity detection (VAD). The primary task, voice activity projection, yields a 256-dimensional state vector that predicts the voice activity of the two participants over the next two seconds (Ekstedt and Skantze, 2022b), as illustrated in Figure 5. This two-second period is divided into four time intervals: 0-200 ms, 200-600 ms, 600-1200 ms, and 1200-2000 ms. Consequently, there are eight binary bins in total, four for each participant, resulting in 256 ($= 2^8$) possible combinations of speaking/non-speaking states for each participant within these bins. The voice activity detection task, conversely, focuses on the current frame, producing two binary vectors representing the present voice activity of each participant. Both tasks are trained using the cross-entropy loss function, denoted as L_{vap} and L_{vad} in the following subsection.

5.2 Fine-Tuning for Backchannel Prediction

Following the pre-training of the VAP model, an additional training phase is conducted using data specific to backchannel prediction. To facilitate this, a new linear layer is introduced on top of the VAP model, complementing the existing layers for VAP and VAD, as depicted in Figure 4. The loss function for this fine-tuning process, denoted as L, is formulated as follows:

$$L = \alpha\, L_{vap} + \beta\, L_{vad} + \gamma L_{bc}\,, \tag{1}$$

where α, β, and γ are the hyperparameters used to adjust the balance between the three tasks, with γ typically assigned a higher value due to the primary focus on backchannel prediction. The first two terms are consistent with those used in the original VAP model (Ekstedt and Skantze, 2022b), while the final term, L_{bc}, is newly introduced in this work. This term represents the cross-entropy loss associated with backchannel prediction and is defined as:

$$L_{bc} = -\log \sigma(\mathbf{o}_{bc}(r_{bc}))\,, \tag{2}$$

where $\mathbf{o}_{bc}$ represents the output from the linear head associated with backchannel prediction, and r_{bc} denotes the reference label. It is important to note that the dimensionality of these vectors is dependent on the specific task. For instance, in binary classification tasks, such as predicting the presence or absence of a backchannel (Section 4.1 and Section 6.1), the dimensionality would be 2. Conversely, in multi-class classification tasks, which involve predicting both the timing and type of backchannel (Section 4.2 and Section 6.2), the dimensionality would exceed 3.

6 Experiment

To evaluate the effectiveness and applicability of the proposed method, we conducted the four experiments described below.

6.1 Timing Prediction

The first experiment focuses on the initial tasks outlined in Section 4.1, which involve a binary classification of backchannels or non-backchannels. We prepared several comparative methods, including random classification (always outputs positive), as detailed below:

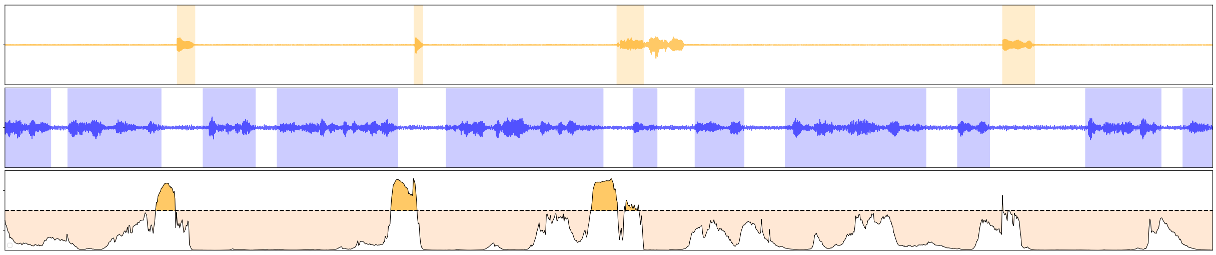

Figure 6: Example of backchannel timing prediction. The top section represents the listener's activity (orange: backchannel), the middle shows the speaker's speech, and the bottom section illustrates the model's predicted probabilities for backchannel occurrence.

(i) **Baseline** consists solely of the audio encoder CPC and a linear head. While other methods in this study freeze the CPC during training, this baseline approach fine-tunes the CPC model itself. As the two input audio channels are separately fed into the CPC, their distinct output vectors are concatenated and then passed to the linear head.

(ii) **Zero-shot** is based on the direct use of the VAP output proposed in the original VAP work (Ekstedt and Skantze, 2022b). It is a normalized value obtained by adding the probability of immediate voice activity of the listener (system) (0.0 to 1.2 seconds) and the probability of slightly later voice activity of the speaker (user) (1.2 to 2.0 seconds). A higher value indicates that the immediate system utterance would be short, corresponding to a backchannel.

(iii) **ST w/o PT** refers to the single-task (ST) model, where the loss function only includes L_{bc} from Equation (1). Moreover, this model does not involve any pre-training (PT) of the VAP model.

(iv) **ST w/ PT** introduces pre-training of the VAP model in addition to the single-task learning.

(v) **MT-ASR** performs another type of multi-task learning, where backchannel prediction is trained together with automatic speech recognition, inspired by a recent backchannel prediction model (Choi et al., 2024). This subtask was trained to recognize phonemes (19 phones) using the CTC loss function.

(vi) **MT w/ PT** represents the proposed method, which incorporates both multi-task (MT) learning, as described in Equation 1, and the pre-training of the VAP model.

Table 1: Result on timing prediction (ST: Single-task, MT: Multi-task, PT: Pre-training)

Method	F1-score	Precision	Recall
Random	13.76	7.39	100.00
Zero-shot	15.11	8.22	93.11
Baseline	36.37	26.43	58.32
ST w/o PT	36.34	25.04	66.24
ST w/ PT	41.65	31.31	62.18
MT-ASR	39.34	28.25	63.39
MT w/ PT	42.85	32.52	62.80

The evaluation metrics used are F1-score, precision, and recall, calculated in a frame-wise manner, and the F1-score is the most priority indicator. For the hyperparameters in Equation (1), to emphasize the importance of backchannel prediction, we empirically set them as $\alpha = 1$, $\beta = 1$, and $\gamma = 5$. Additionally, to mitigate the impact of the imbalanced dataset, we adjusted the loss weight, assigning a weight five times larger to positive (backchannel) samples compared to negative (non-backchannel) samples.

Table 1 presents a summary of the results for this task. Firstly, all trained methods demonstrated significantly higher scores when compared to the random and zero-shot approaches. In particular, the proposed method (MT w/ PT) achieved the highest scores in both the F1 score and precision metrics. These findings suggest that both pre-training and multi-task learning play a crucial role in improving backchannel prediction performance, indicating that this task requires a more robust approach than conventional target-specific training or fine-tuning. Moreover, the VAP model, along with its original loss function, exhibits better generalizability and applicability to other non-linguistic behavior predictions, such as the current backchannel prediction task.

Figure 6 illustrates a sample output generated by

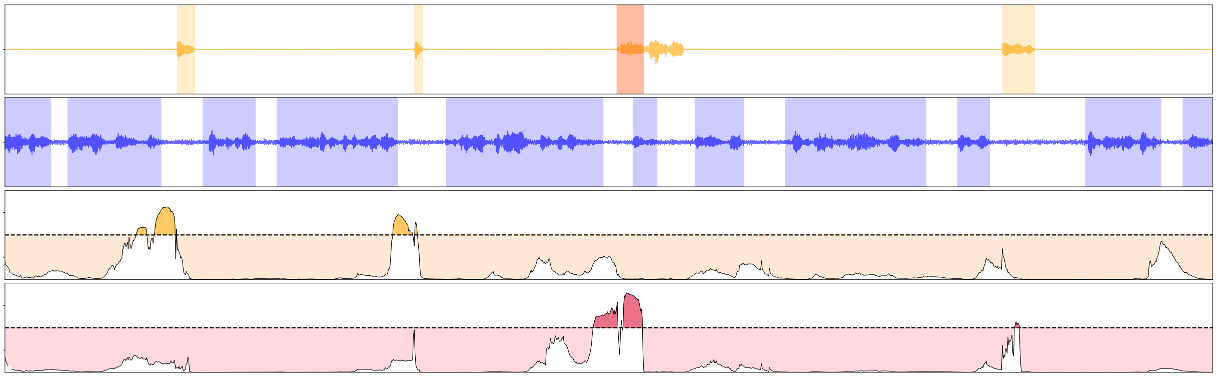

Figure 7: Example of model predictions for backchannel timing and type. The top two sections show the listener (orange: continuer, pink: assessment) and speaker activities. The third section displays the model's prediction probabilities for continuer (orange), while the bottom section shows the probabilities for assessment (pink).

Table 2: Result on time and type prediction on **continuer** (ST: Single-task, MT: Multi-task, PT: Pre-training)

Method	F1-score	Precision	Recall
Random	10.19	5.37	100.00
Baseline	34.13	26.59	47.63
ST w/o PT	36.10	28.65	48.77
ST w/ PT	36.47	29.08	48.90
MT-ASR	33.45	25.33	49.26
MT w/ PT	38.11	29.89	52.58

Table 3: Result on time and type prediction on **assessment** (ST: Single-task, MT: Multi-task, PT: Pre-training)

Method	F1-score	Precision	Recall
Random	3.57	1.82	100.00
Baseline	19.74	32.71	14.13
ST w/o PT	23.72	26.11	21.73
ST w/ PT	30.09	30.36	29.82
MT-ASR	20.27	25.33	16.90
MT w/ PT	31.76	29.95	33.81

the model. As shown in the graph, even while the Blue speaker is still speaking, the model is capable of predicting multiple backchannel points just prior to their occurrence.

6.2 Timing and Type Prediction

The second experiment involves the prediction of backchannel types, as outlined in Section 4.2. We employed the same comparative methods as in the previous experiment, but adjusted the output dimension of the linear head from 2 to 3 to accommodate the classification of continuers, assessments, and non-backchannels. The evaluation metric remained unchanged; however, we conducted separate evaluations for continuers and assessments. Note that the zero-shot approach (defined in Section 6.1) was not applicable to this task and was therefore excluded.

Table 2 and Table 3 present the outcomes of this task. As with the previous results, both tables demonstrate that the combination of multi-task learning and pre-training significantly enhanced performance, with the proposed method (MT w/ PT) achieving the highest F1-score. When comparing these two types, as anticipated, the prediction of assessment backchannels yielded lower scores. While random prediction offers no meaningful insight, the proposed method exceeded an F1-score of 30.

Figure 7 illustrates a sample output generated by the model. In this example, the listener uttered two continuer backchannels (orange) followed by an assessment (pink). The model can correctly predict the first two continues and then also properly predict the assessment. From this result, the model can be trained properly to predict both two types of backchannels.

6.3 Prosody Sensitivity

We further examined the extent to which the model depends on prosodic information. Previous work on the VAP model conducted a similar experiment by flattening the pitch (Figure 8) and intensity (Figure 9) of the test input (Ekstedt and Skantze, 2022a). In this study, we similarly utilized Praat[3] to flatten both pitch and intensity, respectively. If such manipulations significantly degrade performance, it would suggest that the model both relies on and effectively captures the prosodic information. We subsequently analyzed the performance changes

[3] https://www.fon.hum.uva.nl/praat/

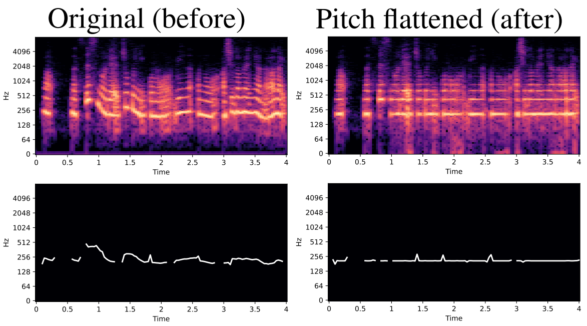

Figure 8: Input example of pitch flattening test (Top: Spectrogram, Bottom: Automatically estimated F0)

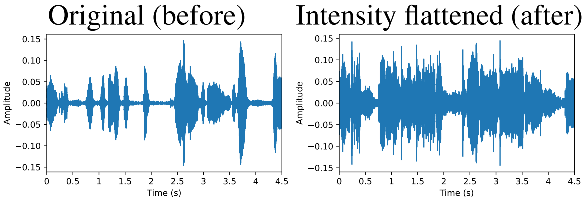

Figure 9: Input example of intensity flattening test

before and after applying the flattening manipulations across the three classification models. Due to space limitations, we report only the F1-score of the proposed model (**MT w/PT**).

Table 4 presents the results of this experiment. Overall, neither manipulation significantly degraded performance, suggesting that the model may rely more heavily on other factors, such as linguistic information. Both types of backchannels exhibited a similar trend, with intensity flattening causing greater degradation than pitch flattening. This finding indicates that the current backchannel prediction model captures the intensity dynamics of preceding user utterances more effectively. When comparing the two types of backchannels, the assessment revealed a higher dependence on intensity, though the difference was not substantial.

6.4 Real-time Processing Performance

To validate the applicability of live spoken dialogue systems, we also examined the relationship between the model's input context length and its prediction performance. As the CPC audio encoder is composed of an autoregressive model, we provided the entire context audio input to the encoder. Subsequently, we constrained the input length for the Transformer layers. In addition, we adjusted the frame rate to 10 Hz, which is sufficient for real-time prediction systems, and retrained the models accordingly. Therefore, note that the results in this section would be different from the ones so far.

Table 4: Pitch and intensity flattening result

Manipulation	F1-score	
	Continuer	Assessment
None (original)	38.11	31.76
Pitch flat	37.20 (-0.91)	31.09 (-0.67)
Intensity flat	35.48 (-2.63)	28.73 (-3.03)

Table 5: Real-time processing performance (RTF: Real-time factor)

Context [sec.]	F1-score		RTF
	Continuer	Assessment	
20	36.17	28.75	0.229
10	36.51	30.46	0.220
5	36.57	30.08	0.194
3	35.79	29.51	0.172
1	35.25	27.67	0.157

In this experiment, we employed the second task, which involves predicting both the timing and type of backchannels. For this evaluation, only a CPU was utilized, specifically an Intel Core i7-11700 @ 2.50 GHz.

The result for the continuer and assessment backchannels in the different input context lengths is presented in Table 5. Also, in this experiment, we only reported the F1-score of the proposed model (**MT w/PT**). Overall, due to the compact design of the VAP model, the real-time factor (RTF) was consistently below 1.0 in all cases, indicating that real-time processing is achieved. Regarding the effect of input context length on the Transformer layers, approximately 5 seconds of input context yielded optimal results for both types of backchannels. When comparing the two types, while the continuer backchannels could be predicted even with a 1-second input context, the performance for assessment backchannels decreased significantly with shorter contexts, such as 1 or 3 seconds. This disparity suggests that the prediction of assessment backchannels requires a longer input context.

7 Integration with a CG Agent

Based on the results from the previous experiment, we have developed a VAP-based real-time backchannel prediction system and implemented it with a conversational CG agent[4]. Figure 10 illustrates the system in operation with the agent, as well as its real-time GUI visualization tool. Note

[4] CG-CA Takumi (c) 2023 by Nagoya Institute of Technology, Moonshot R&D Goal 1 Avatar Symbiotic Society

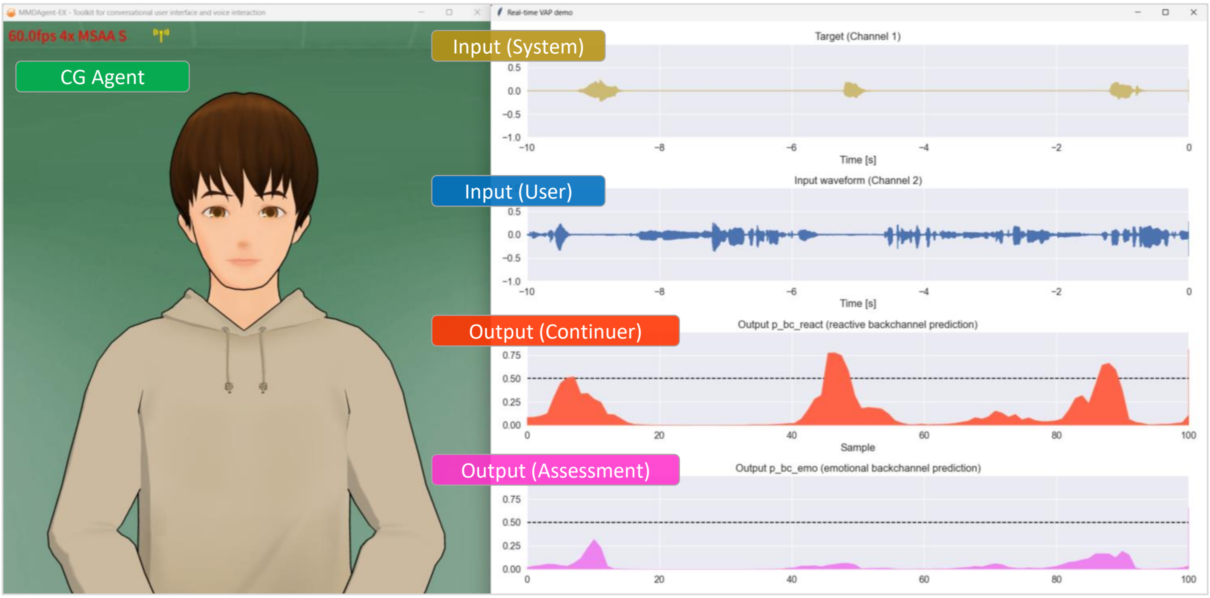

Figure 10: A conversational agent with VAP-based backchannel prediction and its GUI visualization

that since this backchannel generation system operates independently of the interface, it can be applied to various other interfaces, including physically embodied robots. We plan to conduct a user dialogue experiment with this dialogue system to evaluate the naturalness and effectiveness of the backchannel generation system.

8 Conclusion

This paper presents a method for real-time, continuous backchannel prediction using a fine-tuned Voice Activity Projection (VAP) model. Our approach combines pre-training on a larger dialogue dataset with fine-tuning on a specialized backchannel dataset, leveraging the VAP architecture's generalizability. Experimental results showed that our two-stage and multi-task training process improves the model's ability to predict both the timing and type of backchannels, demonstrating its adaptability to real-world, unbalanced data. We also validated the model for real-time use, confirming its effectiveness in live systems without compromising accuracy, especially for continuer backchannels. The results also highlight the need for longer input contexts for accurate assessment backchannel predictions.

This study represents a step forward in enhancing conversational agents' interactivity by providing a more human-like and responsive backchanneling system. Future research will concentrate on evaluating the effectiveness of the backchannel generation system through user dialogue experiments, as well as further refining backchannel prediction for more complex conversational contexts.

Acknowledgments

This work was supported by JST PREST JPMJPR24I4, JST Moonshot R&D JPMJPS2011, and JSPS KAKENHI JP23K16901.

Limitations

This study was evaluated solely on a Japanese dialogue dataset, which limits the generalizability of the model to other languages. Future work should assess its performance on common other datasets like Switchboard to ensure broader applicability. Additionally, while our model shows promise for real-time backchannel prediction, it has not been evaluated in practical settings with conversational agents or robots. Future experiments involving user interactions with such systems are necessary to evaluate the model's effectiveness and user impressions in real-world scenarios.

Ethical Considerations

In the process of collecting dialogue data, all participants were informed about the purpose of the research, and their explicit consent was obtained for the use of their data. The data collection process was designed to ensure the protection of participants' privacy, and any personal information was anonymized or excluded from the dataset to prevent identification. The study was conducted in accordance with ethical guidelines, and approval was obtained from the appropriate ethics committee prior to data collection. The participants' privacy and confidentiality were strictly maintained throughout the research process.

References

Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards immediate backchannel generation using attention-based early prediction model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7408–7412.

Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, and Roxane Bertrand. 2020. An integrated model for predicting backchannel feedbacks. In *International Conference on Intelligent Virtual Agents (IVA)*, pages 1–3.

Yong-Seok Choi, Jeong-Uk Bang, and Seung Hi Kim. 2024. Joint streaming model for backchannel prediction and automatic speech recognition. *ETRI Journal*, 46(1):118–126.

Patricia M Clancy, Sandra A Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of pragmatics*, 26(3):355–387.

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

Yasuharu Den, Nao Yoshida, Katsuya Takanashi, and Hanae Koiso. 2011. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 168–173.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis P. Morency. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1061–1068.

Kent Drummond and Robert Hopper. 1993. Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on language and Social Interaction*, 26(2):157–177.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? Investigations using voice activity projection models. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 541–551.

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

Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *INTERSPEECH*, pages 889–892.

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. In *INTERSPEECH*, pages 991–995.

Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020a. Job interviewer android with elaborate follow-up question generation. In *International Conference on Multimodal Interaction (ICMI)*, pages 324–332.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. Multilingual turn-taking prediction using voice activity projection. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 11873–11883.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024b. Real-time and continuous turn-taking prediction using voice activity projection. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*, 9.

Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020b. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *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 *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 212–215.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In *International Conference on Intelligent Virtual Agents (IVA)*, pages 131–138.

Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring semi-supervised learning for predicting listener backchannels. In *Conference on Human Factors in Computing Systems (CHI)*.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3447–3452.

Tatsuya Kawahara, Miki Uesato, Yoshino Koichiro, and Katsuya Takanashi. 2015. Toward adaptive generation of backchannels for attentive listening agents. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G Ward. 2016. Prediction and generation of backchannel form for attentive listening systems. In *INTERSPEECH*, pages 2890–2894.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and speech*, 41(3-4):295–321.

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 127–136.

Wencke Liermann, Yo-Han Park, Yong-Seok Choi, and Kong Lee. 2023. Dialogue act-aided backchannel prediction using multi-task learning. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP) Findings*, pages 15073–15079.

Jennifer A Maddrell and Ginger S Watson. 2012. The influence of backchannel communication on cognitive load. In *The Next Generation of Distance Education: Unconstrained Learning*, pages 171–180. Springer.

L. P. Morency, I. D. Kok, and J. Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 20(1):70–84.

Louis-Philippe Morency, Iwan De Kok, and Jonathan Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *International workshop on intelligent virtual agents (IVA)*, pages 176–190.

Keiko Ochi, Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2024. Entrainment analysis and prosody prediction of subsequent interlocutor's backchannels in dialogue. In *INTERSPEECH*, pages 462–466.

Derya Ozkan and Louis-Philippe Morency. 2011. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 335–340.

Yo-Han Park, Wencke Liermann, Yong-Seok Choi, and Kong Joo Lee. 2024. Improving backchannel prediction leveraging sequential and attentive context awareness. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL) Findings*, pages 1689–1694.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017a. Enhancing backchannel prediction using word embeddings. In *INTERSPEECH*, pages 879–883.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017b. Yeah, right, uh-huh: A deep learning backchannel predictor. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages 247–258.

Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. 2011. Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183.

Chen-Hsin Tang and Grace Qiao Zhang. 2009. A contrastive study of compliment responses among Australian English and Mandarin Chinese speakers. *Journal of pragmatics*, 41(2):325–345.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics*, 32(8):1177–1207.

Deanna Wong and Pam Peters. 2007. A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification. *International Journal of Corpus Linguistics*, 12(4):479–510.