Multilingual and Continuous Backchannel Prediction: A Cross-lingual Study

Koji Inoue, Mikey Elmers, Yahui Fu, Zi Haur Pang, Taiga Mori, Divesh Lala, Keiko Ochi, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan,

Correspondence: inoue@sap.ist.i.kyoto-u.ac.jp

Abstract

We present a multilingual, continuous backchannel prediction model for Japanese, English, and Chinese, and use it to investigate cross-linguistic timing behavior. The model is Transformer-based and operates at the frame level, jointly trained with auxiliary tasks on approximately 300 hours of dyadic conversations. Across all three languages, the multilingual model matches or surpasses monolingual baselines, indicating that it learns both language-universal cues and language-specific timing patterns. Zero-shot transfer with two-language training remains limited, underscoring substantive cross-lingual differences. Perturbation analyses reveal distinct cue usage: Japanese relies more on short-term linguistic information, whereas English and Chinese are more sensitive to silence duration and prosodic variation; multilingual training encourages shared yet adaptable representations and reduces overreliance on pitch in Chinese. A context-length study further shows that Japanese is relatively robust to shorter contexts, while Chinese benefits markedly from longer contexts. Finally, we integrate the trained model into a real-time processing software, demonstrating CPU-only inference. Together, these findings provide a unified model and empirical evidence for how backchannel timing differs across languages, informing the design of more natural, culturally-aware spoken dialogue systems.

1 Introduction

Smooth human conversation is supported by brief listener responses such as "uh-huh" and "oh," known as *backchannels*, produced at appropriate moments (Schegloff, 1982; Clark, 1996; Clancy et al., 1996). Backchannels serve not only to regulate turn-taking but also to signal interest and understanding, and are thus essential for spoken dialogue systems that aim to interact in a human-like manner (Schroder et al., 2011; DeVault et al., 2014; Inoue et al., 2016). Their importance is also recognized in emerging full-duplex spoken dialogue systems, for which modeling and evaluation foundations are being established (Défossez et al., 2024; Lin et al., 2025).

Automatic backchannel generation involves predicting three factors: timing, form, and prosody. Among these, *timing*—when to produce a backchannel—is fundamental. Prior work has explored utterance-level and frame-level (continuous) prediction (Jang et al., 2021; Ruede et al., 2017). Because humans often insert backchannels before a speaker's utterance is complete, continuous frame-level models are preferable for reproducing human-like behavior. However, continuous models face challenges such as label imbalance. Recent approaches improve performance via multi-task learning with related tasks such as turn-taking prediction (Hara et al., 2018; Choi et al., 2024; Inoue et al., 2025).

Most backchannel studies to date have targeted a single language (often Japanese or English), and cross-linguistic analyses remain limited. Backchannel frequency and timing vary by language and culture; for example, in Japanese, backchannels often occur during the speaker's ongoing utterance, whereas in Chinese they more frequently appear after utterance completion (Clancy et al., 1996). Quantifying these differences and modeling both universal and language-specific features are crucial steps toward dialogue technologies that are robust across diverse linguistic cultures.

To this end, we conduct a comparative analysis of backchannel timing in Japanese, English, and Chinese. We first compile a large-scale, 300-hour multilingual conversational dataset. We then build a Transformer-based multilingual backchannel prediction model that continuously outputs frame-level probabilities. The model is designed

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 222–230
February 26–March 1, 2026.

Table 1: Statistics of backchannel data by language

	Japanese	English	Chinese
# Dialogues	299	300	298
Total dialogue time	108:13:34	119:56:12	108:05:12
Total used dialogue time	49:13:39	27:20:31	25:04:53
# Backchannel utterances	58800 (34.4%)	24612 (28.4%)	21182 (27.5%)
# Non-backchannel utterances	112177	62158	55955
Total BC time [s]	29253.73 (16.5%)	11006.73 (11.2%)	7695.41 (8.5%)
Total non-BC time [s]	147965.49	87424.67	82598.55

to learn features that are shared across languages while also capturing language-specific patterns. Finally, we compare monolingual and multilingual settings and analyze which input aspects are important for predicting backchannels across languages, highlighting commonalities and differences.

2 Dataset

We analyze first-encounter dyadic conversations recorded over an online conferencing tool (Zoom). The total recording time is nearly 300 hours: about 100 hours each for Japanese, English, and Chinese. Utterances were segmented manually into Inter-Pausal Units (IPUs) using a 200 ms silence threshold. We then applied automatic speech recognition (ASR) to each segment to obtain transcripts. Whisper was used for ASR: *kotoba-tech/kotoba-whisper-v2.2*[1] for Japanese, and *large-v3*[2] for English and Chinese.

Using a manually curated surface-form list of backchannels, we identified backchannel utterances from the ASR outputs. Following prior works (Choi et al., 2024; Inoue et al., 2025), our target forms comprise interjections from the *continuer* class (e.g., "うん," "yeah," and "嗯") and the *assessment* class (e.g., "へー," "wow," and "哦"). For each language, the list was verified by native-speaking authors and consolidated to account for variants and dialectal forms. Consecutive backchannels, such as "yeah yeah," were merged into a single instance. If the preceding utterance was produced by the same person who produced the candidate backchannel, we filtered it out (i.e., it was not treated as a listener backchannel).

Since VAP training operates on 20-second segments, we split the dialogues accordingly. In each segment, the participant who produced more backchannels was labeled the *listener*, and the other the *speaker*. We then predicted backchannels for the listener. Segments with no backchannels were excluded.

[1] https://huggingface.co/kotoba-tech/kotoba-whisper-v2.2

[2] https://huggingface.co/openai/whisper-large-v3

Table 2: Share of backchannels occurring during vs. after the preceding utterance

	During utt.	After utt.
Japanese	40804 (69.4%)	17996 (30.6%)
English	14981 (60.9%)	9631 (39.1%)
Chinese	10038 (47.4%)	11144 (52.6%)

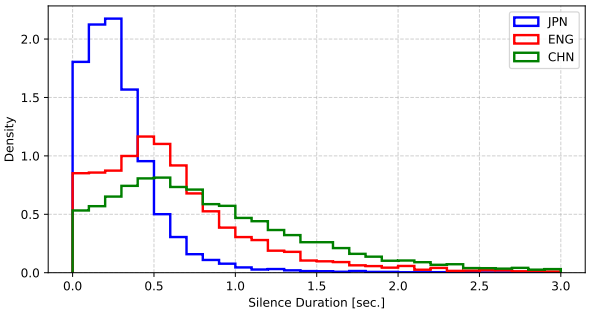

Figure 1: Probability density histograms of the time lag between the end of the preceding utterance and the onset of the backchannel

Table 1 summarizes the annotations. Japanese shows the highest backchannel rate: approximately 34.4% of utterances and 16.5% of total time. English and Chinese exhibit lower rates (28.4% / 11.2% and 27.5% / 8.5%, respectively), suggesting cross-linguistic differences in backchannel behavior. The higher frequency in Japanese aligns with prior reports (Maynard, 1986; Clancy et al., 1996; Miller, 2011).

We further analyzed whether backchannels overlap with the speaker's ongoing utterance or occur after a silence following utterance completion. As shown in Table 2, 69.4% of Japanese backchannels occur during the speaker's utterance, which is higher than in Chinese (47.4%). This reflects a conversational tendency in Japanese to insert sup-

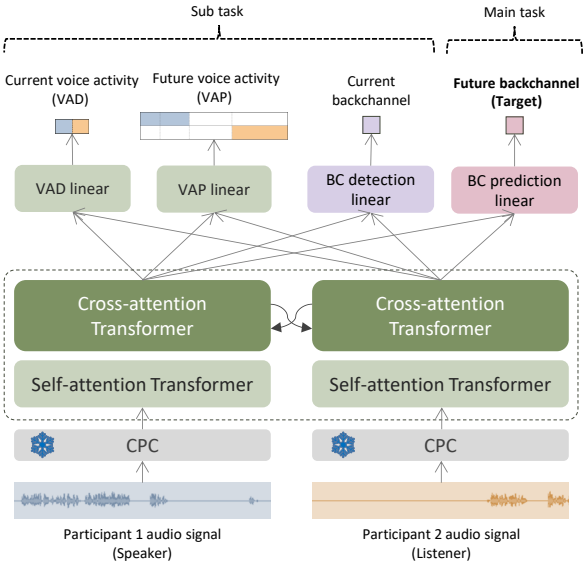

Figure 2: Architecture of the backchannel prediction model

portive responses mid-utterance, with such insertions less likely to be perceived as interruptions. By contrast, in Chinese, 52.6% of backchannels occur after the speaker's utterance, indicating a preference to respond following clear completion; this points to more explicit turn boundaries and a sharper separation of speaker and listener roles.

Figure 1 plots the probability density of post-utterance silence (the time from the end of the preceding utterance to the start of a backchannel). The peak for Japanese is around 0.2-0.3 s, while English and Chinese both peak near 0.5 s. Thus, Japanese backchannels tend to occur after shorter silences. Comparing English and Chinese, the Chinese distribution has a heavier tail with a longer mean silence (about 0.9 s vs. about 0.6 s for English), suggesting that Chinese backchannels are more sensitive to the duration of silence.

3 Backchannel Prediction Model

We build upon the Voice Activity Projection (VAP) model (Ekstedt and Skantze, 2022; Inoue et al., 2025), which supports continuous prediction (Figure 2). The inputs are the separated waveforms of two interlocutors—one for the *speaker* and the other for the *listener*. Each input is encoded by a Contrastive Predictive Coding (CPC) encoder into a feature sequence. We use a CPC model pre-trained on the Libri-light dataset (about 60k hours) (Riviere et al., 2020) and keep its parameters frozen.

Encoded features are first processed by separate Transformers and then fused via a Cross-Attention

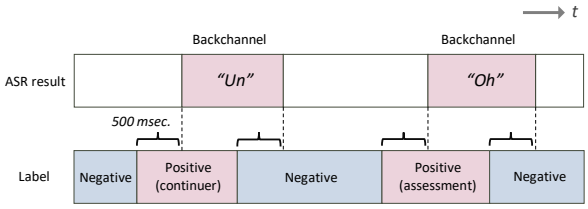

Figure 3: Definition of the target label window

Transformer to capture inter-speaker interactions. The resulting representation is fed to four linear heads (top of Figure 2), each corresponding to a subtask. Following (Ekstedt and Skantze, 2022) and (Inoue et al., 2025), we design the tasks, which are helpful for stabilizing learning under label imbalance (backchannels can be sparse in some languages), as follows:

- **Voice Activity Detection (VAD)** estimates the probability of speaking vs. non-speaking for each interlocutor at the current frame. This is the subtask in the original VAP model.
- **Voice Activity Projection (VAP)** predicts the joint speaking states of both interlocutors over the next 2 s, as a proxy for turn-taking prediction. We discretize into four bins: 0-200 ms, 200-600 ms, 600-1200 ms, and 1200-2000 ms, and represent the joint state (speak/non-speak for each person) in each bin, yielding a 256-class output.
- **Backchannel Detection (BD)** estimates whether the listener is currently producing a backchannel. This task is expected to supplement backchannel prediction by explicitly identifying backchannel instances.
- **Backchannel Prediction (BP)** estimates whether the listener will produce a backchannel 0.5 s in the future. This is our main task. During training, we shift annotated backchannel onsets by 0.5 s (Figure 3) to create supervision targets.

The overall loss is

$$L = \alpha_1 L_{\text{VAD}} + \alpha_2 L_{\text{VAP}} + \alpha_3 L_{\text{BD}} + \alpha_4 L_{\text{BP}} \,, \quad (1)$$

where L_{VAD}, L_{VAP}, L_{BD}, and L_{BP} are the losses for VAD, VAP, backchannel detection, and backchannel prediction, respectively. We set $\alpha_1 = \alpha_2 = 1.0$ and $\alpha_3 = \alpha_4 = 5.0$ to emphasize the backchannel-related tasks, following (Inoue et al., 2025).

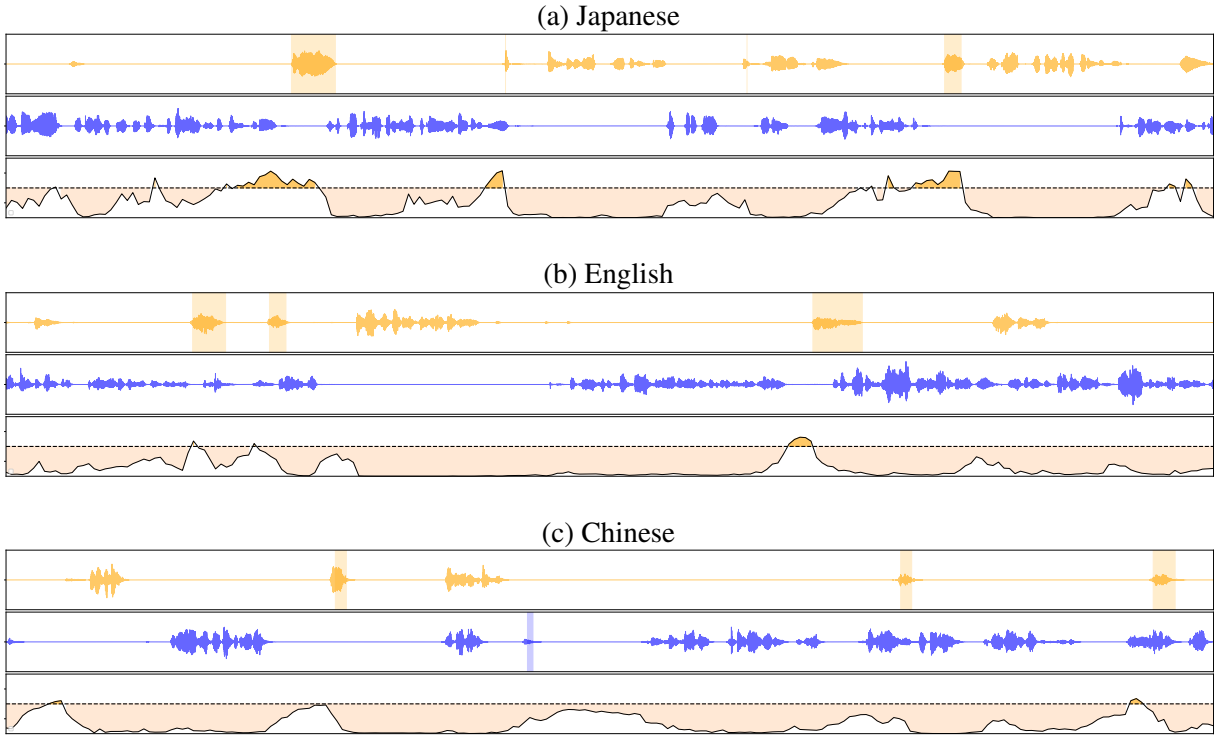

Figure 4: Examples of multilingual model behavior on test data for each language (top to bottom in each panel: listener waveform, speaker waveform, listener backchannel prediction probability; highlighted regions indicate backchannel intervals).

Table 3: Cross-lingual backchannel prediction performance (F1 score [%])

Training	Test		
	Japanese	English	Chinese
Japanese	33.27	15.41	10.32
English	7.92	22.85	19.52
Chinese	10.32	19.52	21.37
Multilingual	**33.69**	**23.96**	**22.65**

Table 4: Zero-shot performance of two-language models

Training	Test	F1 score [%]
English-Chinese	Japanese	8.02
Chinese-Japanese	English	12.33
Japanese-English	Chinese	17.02

4 Experiments

We evaluate the proposed multilingual backchannel prediction model and analyze cross-linguistic differences.

4.1 Setup

We train three monolingual models (Japanese, English, Chinese) and one multilingual model (all three languages). Details are as follows:

- **Model:** same architecture as Figure 2; 1 transformer layer for each speaker and 3 cross-attention transformer layers; model dimension 256; 4 attention heads. Note that no language information (e.g., language ID) is explicitly provided to the model.
- **Dataset:** the corpus in Section 2; for each language, dialogues are randomly split into train (80%), validation (10%), and test (10%). The multilingual model is trained on the combined training sets of all three languages.
- **Training:** AdamW optimizer; learning rate 3.63×10^{-4}, batch size 8, max 25 epochs.
- **Metric:** frame-level (100 ms) F1 score, following (Inoue et al., 2025), with a decision threshold of 0.5 on predicted probabilities.

4.2 Cross-Lingual Performance

Table 3 presents F1 scores for monolingual and multilingual models. As expected, monolingual models perform best on their own language (matched) but degrade substantially in zero-shot transfer. This mirrors the cross-linguistic differences observed in Section 2. For instance, a model trained on Japanese–where many backchannels occur during the speaker's utterance–struggles on Chi-

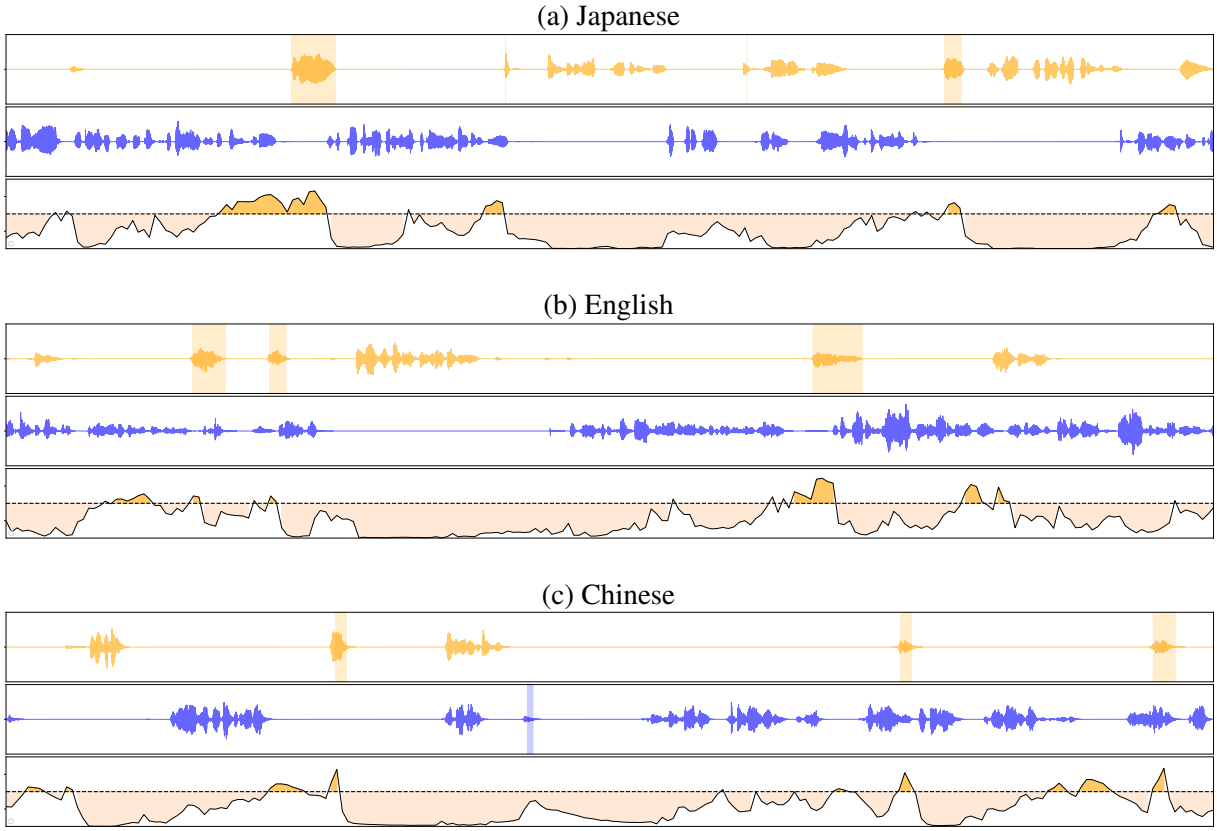

Figure 5: Examples of Japanese monolingual model behavior on test data for each language (top to bottom in each panel: listener waveform, speaker waveform, listener backchannel prediction probability; highlighted regions indicate backchannel intervals).

nese, where backchannels more often follow utterance completion, and vice versa. English tends to fall between Japanese and Chinese both in frequency and silence timing, and accordingly shows intermediate cross-lingual transfer. Nevertheless, the persistent degradation indicates that English does not simply subsume the other two languages.

By contrast, the multilingual model achieves performance comparable to or better than the matched monolingual models in all three languages. This suggests that the model effectively learns universal cues while adapting to language-specific patterns based on the input. The finding aligns with prior work on multilingual turn-taking prediction (Inoue et al., 2024).

We also evaluated two-language models in zero-shot settings (Table 4). These models, trained on pairs of languages, performed poorly when tested on the unseen third language, underperforming both the matched monolingual models and the three-language multilingual model. This suggests that backchannel behaviors differ substantially across all three languages, making it difficult to learn universal and emergent prediction capabilities for backchannel behaviors.

Figure 4 illustrates the behavior of the multilingual model across languages. In (a) Japanese, the prediction probability rises just before true backchannel intervals, consistent with the relatively high F1. In (b) English and (c) Chinese, while the peaks are less sharp, the model still captures backchannel timings reasonably well. On the other hand, Figure 5 shows that the Japanese monolingual model struggles more with English and Chinese test data, producing many false positives and failing to capture backchannel timings. This further highlights the limitations of monolingual models in cross-lingual settings.

4.3 Ablation Study

We also conducted an ablation study to assess the contributions of auxiliary tasks (VAD, VAP, backchannel detection) to the main backchannel prediction task. Table 5 shows the results for monolingual models. Overall, we did not see any large degradation when removing auxiliary tasks, suggesting that monolingual models can learn backchannel prediction reasonably well on

Table 5: Ablation result of (matched) **monolingual** models (F1 score [%] and drop)

Ablation	Japanese		English		Chinese	
Original	33.27		22.85		21.37	
w/o L_{BD}	33.14	(−0.13)	22.91	(+0.06)	22.67	(+1.30)
w/o L_{VAP}	33.09	(−0.18)	22.04	(−0.81)	21.22	(−0.15)
w/o L_{VAD}	33.81	(+0.54)	21.75	(−1.10)	22.02	(+0.65)

Table 6: Ablation result of **multilingual** model (F1 score [%] and drop)

Ablation	Japanese		English		Chinese	
Original	33.69		23.96		22.65	
w/o L_{BD}	33.41	(−0.28)	23.99	(+0.03)	23.25	(+0.60)
w/o L_{VAP}	32.57	(−1.12)	21.11	(−2.85)	20.02	(−2.63)
w/o L_{VAD}	33.79	(+0.10)	23.29	(−0.67)	21.93	(−0.72)

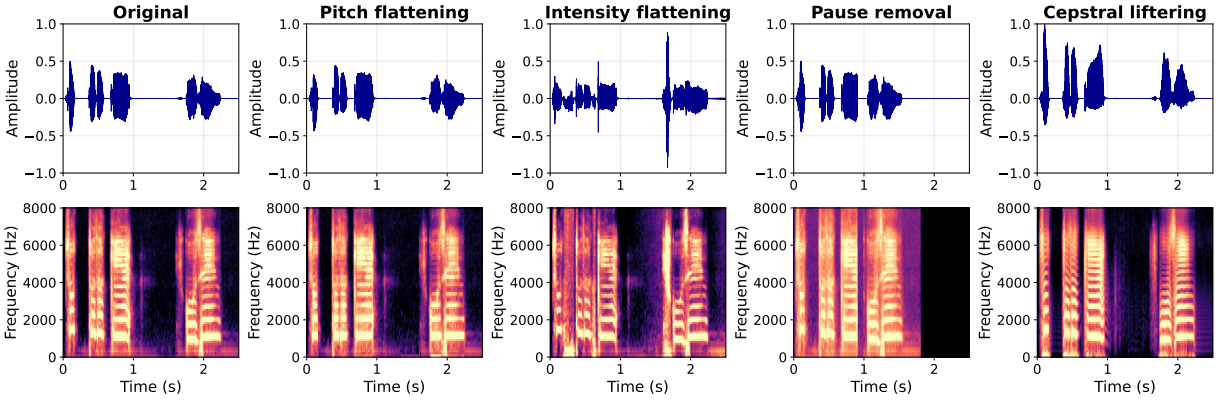

Figure 6: Example of perturbation processing

their own. Rather, in some cases, removing auxiliary tasks slightly improved performance (e.g., removing L_{BD} in Chinese and L_{VAD} in Japanese). This may be because the monolingual models can already capture language-specific cues effectively, and auxiliary tasks may introduce noise or conflicting signals.

On the other hand, the case of the multilingual model is different and showed consistent trends across languages. Table 6 shows the results for the multilingual model. Removing the VAP loss (L_{VAP}) causes the largest performance drop across all languages (−0.77 to −3.59 points), indicating that learning turn-taking dynamics also contributes to backchannel timing prediction. Removing the backchannel detection loss (L_{BD}) also degrades performance, but to a lesser extent (−0.08 to −1.88 points), suggesting that explicit backchannel identification also aids prediction. Interestingly, removing the VAD loss (L_{VAD}) slightly improves performance in all languages (+0.19 to +0.99 points), possibly because VAD may introduce noise when speech activity is not strongly correlated with backchannel timing prediction. These findings indicate that auxiliary tasks play a more critical role in the multilingual setting, helping the model learn shared representations that generalize across languages.

4.4 Perturbation Analysis

To identify which input aspects the models exploit, we perform perturbation analyses by applying controlled manipulations to test audio and measuring performance changes. As depicted in Figure 6, we consider four perturbations:

- **Pitch flattening:** remove pitch variation to test reliance on F0 dynamics.
- **Intensity flattening:** flatten amplitude dynamics to test reliance on energy contours.
- **Pause removal:** remove up to 0.5 s of post-utterance silence to test reliance on silent gaps.

Table 7: Perturbation analysis for (matched) **monolingual** models (F1 score [%] and drop)

Perturbation	Japanese		English		Chinese	
None	33.27		22.85		21.37	
Pitch flattening	31.84	(−1.43)	19.19	(−3.36)	20.46	(−0.91)
Intensity flattening	30.16	(−3.11)	19.99	(−2.86)	19.81	(−1.56)
Pause removal	30.32	(−2.95)	16.46	(−6.39)	15.32	(−6.05)
Cepstral liftering	17.02	(−16.25)	9.58	(−13.27)	5.45	(−15.92)

Table 8: Perturbation analysis for **multilingual** model (F1 score [%] and drop)

Perturbation	Japanese		English		Chinese	
None	33.69		23.96		22.65	
Pitch flattening	30.36	(−3.33)	19.53	(−4.43)	21.57	(−1.08)
Intensity flattening	27.38	(−6.31)	20.15	(−3.81)	19.85	(−2.80)
Pause removal	30.77	(−2.92)	17.57	(−6.39)	16.13	(−6.52)
Cepstral liftering	8.81	(−24.88)	8.28	(−15.68)	10.94	(−11.71)

- **Cepstral liftering:** retain only low-order cepstral components to suppress phonetic content and test reliance on linguistic information.

Table 7 shows results for monolingual models. The magnitude of degradation differs by language, indicating different feature usage. The Japanese model is most affected by cepstral liftering (−16.25 points), suggesting strong reliance on linguistic information. The English and Chinese models show large drops for both pause removal (−6.39 / −6.05) and cepstral liftering (−13.27 / −15.92), indicating sensitivity to both silence and linguistic cues. The Chinese model is also relatively robust against pitch and intensity flattening (−0.91 / −1.56), suggesting less dependence on prosodic variation.

Results for the multilingual model (Table 8) broadly follow similar trends, but with notable differences for Japanese and English: the impact of cepstral liftering increases to −24.88 and −15.68, respectively. Note that the Chinese case shows a smaller drop (−11.71) compared to the monolingual one. Pitch and intensity flattening also cause larger drops in the three languages, compared to the monolingual case, indicating increased reliance on prosodic cues. This suggests that, when trained jointly, the model acquires a more language-aware strategy that emphasizes linguistic and prosodic information for all languages, while maintaining a similar level of sensitivity to silence cues.

4.5 Context Length Dependency

We further analyze how varying the input context length affects performance. In the default setting, the model processes 20 s of past audio for both speaker and listener as the input context. In this experiment, we again trained and used the multilingual model, but varied the input context length for the Transformer layers ranging from 1 s to 20 s during both training and inference. Note that since the CPC encoder consists of CNN and GRU layers, it always processes the full 20 s input. Table 9 and Table 10 present the results for monolingual and multilingual models, respectively. Reducing the context length generally degrades performance, but the extent varies by language. Japanese is relatively robust, with only a small drop (−0.55 and −1.54 points) even at 1 s context, suggesting that short-term cues suffice for backchannel prediction. By contrast, English and Chinese show larger drops at 1 s context, indicating greater reliance on longer-term context. Especially for Chinese, performance degrades sharply when context is reduced below around 3 s, suggesting that longer context is crucial for capturing relevant cues. These differences may reflect language-specific conversational dynamics, such as the timing and distribution of backchannels.

5 System Integration for Real-time Processing

Finally, we integrated the trained backchannel prediction models into a real-time spoken dialogue system. We implemented and released an open-source Python package, MaAI[3], which supports real-time execution of VAP-based models (e.g., turn-taking, backchannel, and nodding prediction).

[3] https://github.com/MaAI-Kyoto/MaAI

Table 9: Context length analysis for (matched) **monolingual** models (F1 score [%] and drop against 20 sec.)

Context len. [sec.]	Japanese		English		Chinese	
20	33.27		22.85		21.37	
10	33.75	(+0.48)	22.39	(−0.46)	22.11	(+0.74)
5	33.18	(−0.09)	22.49	(−0.36)	21.63	(+0.26)
3	33.46	(+0.19)	21.60	(−1.25)	19.52	(−1.85)
1	32.72	(−0.55)	18.52	(−4.33)	11.90	(−9.47)

Table 10: Context length analysis for **multilingual** model (F1 score [%] and drop against 20 sec.)

Context len. [sec.]	Japanese		English		Chinese	
20	33.69		23.96		22.65	
10	33.63	(−0.06)	24.25	(+0.29)	22.31	(−0.34)
5	33.48	(−0.21)	24.20	(+0.24)	22.79	(+0.14)
3	33.20	(−0.49)	23.00	(−0.96)	19.96	(−2.69)
1	32.15	(−1.54)	20.58	(−3.38)	15.69	(−6.96)

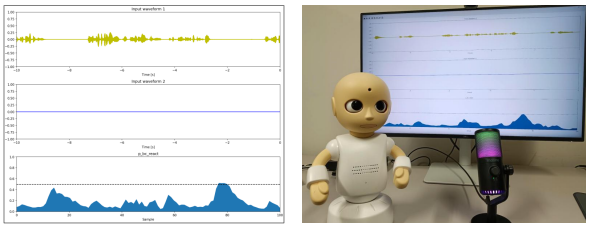

Figure 7: Real-time processing software integrated with a dialogue robot

This package modularizes audio input/output (microphone, network, etc.), VAP processing, and visualization of VAP results, enabling straightforward integration into existing spoken dialogue systems and robots. The trained backchannel prediction models are already integrated into the software; an example of its operation is shown in Figure 7. Thanks to the models' relatively small parameter counts and an efficient caching architecture, the 10 Hz backchannel predictor runs in real time on CPU only (Intel Core Ultra 9 285K).

6 Conclusion

We presented a multilingual backchannel prediction model for Japanese, English, and Chinese, enabling a cross-linguistic analysis of backchannel timing. Trained on a large-scale multilingual conversational corpus, the proposed Transformer-based model achieved comparable or superior performance to monolingual models across all three languages. These results demonstrate that the model successfully captures both language-universal cues and language-specific timing patterns.

Perturbation analyses revealed that the input cues underlying backchannel prediction differ across languages, highlighting distinct conversational strategies: Japanese listeners rely more on linguistic and short-term cues, while English and Chinese listeners are more sensitive to silence duration and prosodic variation. The multilingual model effectively integrates these tendencies, suggesting that cross-lingual training encourages the emergence of shared yet adaptable representations of conversational feedback behavior.

Future work will focus on refining the annotation quality and expanding the scope of analysis. Although utterance segmentation was performed manually, backchannel identification relied on ASR and surface-form matching; incorporating human-verified annotations would enable more precise modeling of backchannel types and functions. We also plan to perform deeper interpretability analyses to elucidate the internal mechanisms by which the model captures language-universal backchannel cues. Finally, by integrating the predictor into real-time spoken dialogue systems and evaluating it through human–machine interaction studies, we aim to quantify its impact on perceived naturalness, engagement, and conversational flow.

Acknowledgments

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

References

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.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

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.

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

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, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. 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, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2025. Yeah, Un, Oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 7171–7181.

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.

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.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung yi Lee. 2025. Full-duplex-bench v1.5: Evaluating overlap handling for full-duplex speech models. *arXiv preprint arXiv:2507.23159*.

Senko K Maynard. 1986. On back-channel behavior in japanese and english casual conversation. *Linguistics*, 24(6):1079–1108.

Laura Miller. 2011. Verbal listening behavior in conversations between japanese and americans. In *The Pragmatics of International and Intercultural Communication: Selected papers from the International Pragmatics Conference, Antwerp, August 1987. Volume 3*, pages 111–130. John Benjamins Publishing Company.

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. 2017. Yeah, right, uh-huh: A deep learning backchannel predictor. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages 247–258.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71(93).

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.