Sensor-Augmented Voice Activity Projection for Enhancing Turn-Taking Prediction

Satoki Hamanaka[1], Yasue Kishino[2], Yuiko Tsunomori[2], Shin Mizutani[2], Yuya Chiba[2], Tadashi Okoshi[1], Jin Nakazawa[1]

[1]Keio University, Japan, [2]NTT Inc., Japan

Correspondence: 0216hamanaka@keio.jp

Abstract

Voice Activity Projection (VAP) has been actively studied to enable natural turn-taking in spoken dialogue systems, relying primarily on acoustic features. Visual cues such as head movements are also known to contribute to turn-taking prediction; however, camera-based approaches are affected by placement and lighting conditions and are not always reliably available to dialogue systems. As a camera-independent approach for directly capturing head motion, earable devices offer a promising solution. In this study, we propose Sensor-Augmented VAP, a framework that integrates in-ear inertial measurement unit (IMU) signals with a pre-trained VAP model via a lightweight residual fusion module. To validate our proposed method, we collected a dataset pairing conversational audio with in-ear IMU data, comprising 12 dyadic Japanese dialogues recorded using microphones and earbuds. Experiments in speaker-independent and speaker-dependent settings demonstrate that IMU fusion consistently improves weighted F1 score for shift detection and reduces VAP loss over the audio-only baseline. These results confirm that head-motion cues are effective for enhancing turn-taking prediction.

1 Introduction

Turn-taking is the process by which conversational participants coordinate the right to speak (Skantze, 2021). Smooth turn-taking requires participants to anticipate when the current speaker will finish speaking and to prepare their own turn and next utterance accordingly. Modeling this process computationally is crucial for building spoken dialogue systems that exhibit human-like turn-taking behavior.

Voice Activity Projection (VAP) (Ekstedt and Skantze, 2022b) is a state-of-the-art framework for real-time, continuous turn-taking prediction. VAP is a model that takes the raw audio waveforms of two conversation participants as input and predicts the probability of future voice activity for each speaker at every time step (Ekstedt and Skantze, 2022a,b). This enables the fine-grained inference of conversational events, such as turn shifts and backchannels using acoustic features.

Several recent studies have demonstrated that non-verbal information can improve the predictive accuracy of VAP. In particular, visual cues such as head nods, gaze direction, and facial expression have been reported to contribute to the detection of turn shifts and turn-hold behaviors (Dermouche and Pelachaud, 2019; Danner et al., 2021; Kurata et al., 2023; Mizuno et al., 2023; Onishi et al., 2024; Russell and Harte, 2025a,b; Ishii et al., 2025; Saga and Pelachaud, 2025). However, camera-based approaches are sensitive to placement and lighting conditions, limiting their reliability in real-world settings. To address these limitations, wearable sensors offer an alternative approach. In particular, earable devices equipped with inertial measurement units (IMUs) can capture head motion directly without relying on cameras or lighting conditions, making them a promising alternative (Röddiger et al., 2022).

In this paper, we propose **Sensor-Augmented VAP**, a framework that integrates in-ear IMU signals with a pre-trained VAP model via a lightweight residual fusion module. To validate our method, we collected a dataset of 12 dyadic Japanese dialogues. To our knowledge, this is the first work to leverage both audio and in-ear IMU signals from a consumer earable device for VAP-based turn-taking prediction.

2 Sensor-Augmented VAP

Sensor-Augmented VAP extends the VAP framework (Inoue et al., 2024b) by incorporating in-ear IMU signals alongside the audio waveform. As illustrated in Figure 1, the proposed model builds

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 164–169
August 2–5, 2026.

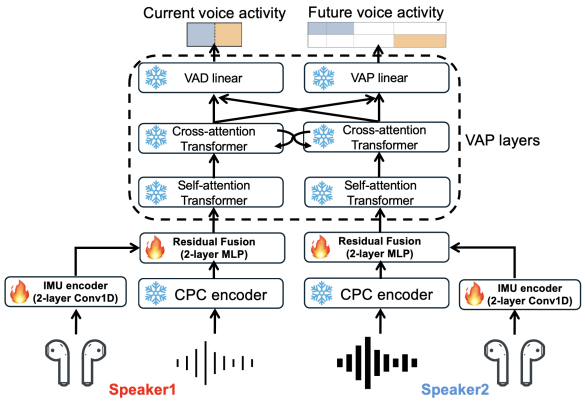

Figure 1: Overview of Sensor-Augmented VAP

upon the original VAP architecture by incorporating an additional IMU encoder and a lightweight residual fusion module that integrates the two modalities. During training, only the IMU encoder and fusion module are updated; all other parameters are kept frozen.

2.1 Base VAP Model Architecture

VAP is a model that takes the audio waveforms of two speakers as input and estimates the probability of future voice activity for each speaker. VAP consists of an audio encoder and a transformer-based sequence model (Inoue et al., 2024b). The audio encoder employs a pre-trained Contrastive Predictive Coding (CPC) model (Riviere et al., 2020) to extract frame-level speech representations from the raw waveform of each speaker. These representations are then passed to the VAP layers, which consist of a per-speaker self-attention transformer that captures temporal dependencies within each speaker's utterance, followed by a cross-attention transformer that models inter-speaker dynamics. The output of the cross-attention transformer is fed into two heads that predict the probability of current and future voice activity, respectively.

The VAP output is converted into two summary values, p_{now} and p_{future} (Ekstedt and Skantze, 2022b). Specifically, p_{now} represents the probability that one speaker holds the turn over a short upcoming window (0-600 ms), while p_{future} covers a longer horizon (600-2000 ms). Values below a decision threshold indicate that one speaker holds the turn, while values above the threshold indicate that another speaker holds the turn. In our proposed model, encoded IMU features are concatenated with the CPC audio representations before being passed to the VAP layers.

2.2 IMU Encoder

The IMU encoder is a module that transforms raw IMU signals into fixed-dimensional embeddings. This encoder is based on a lightweight two-layer 1D convolutional neural network (Conv1D) architecture, with only 21K trainable parameters (0.38% of the total 5.6M parameters). We trained the IMU encoder end-to-end alongside a fusion projection module, allowing its representations to be optimized directly.

2.3 Residual Fusion Module

The Residual Fusion Module integrates IMU features into the audio representations. The module learns a residual $\mathbf{\Delta}$, an additive correction that compensates for errors in the audio-only predictions. The fusion layer is a two-layer MLP with ReLU activation and only 37K trainable parameters (0.66% of the total 5.6M parameters). This lightweight design minimizes trainable parameters, making the approach feasible with limited data.

The residual fusion module takes the encoded audio and IMU representations as input. Let $\mathbf{z}_{\text{audio}} \in \mathbb{R}^{T_a \times d_a}$ denote the audio embeddings and $\mathbf{z}_{\text{imu}} \in \mathbb{R}^{T_i \times d_i}$ denote the IMU embeddings, where T_m and d_m denote the number of frames and the feature dimension of modality $m \in \{\text{audio},\ \text{imu}\}$. Since the IMU encoder operates at a different frame rate from the audio encoder, $\mathbf{z}_{\text{imu}}$ is first temporally resampled via nearest-neighbor interpolation to match the audio frame length T_a, resulting in the aligned IMU features $\mathbf{z}'_{\text{imu}} \in \mathbb{R}^{T_a \times d_i}$. To integrate the two modalities, $\mathbf{z}_{\text{audio}}$ and $\mathbf{z}'_{\text{imu}}$ are concatenated along the feature dimension, producing a joint representation $[\mathbf{z}_{\text{audio}}; \mathbf{z}'_{\text{imu}}] \in \mathbb{R}^{T_a \times (d_a + d_i)}$. A two-layer MLP projects the combined feature dimension $(d_a + d_i)$ back to the original audio feature dimension d_a, computing a residual correction term $\mathbf{\Delta} \in \mathbb{R}^{T_a \times d_a}$ as $\mathbf{\Delta} = \text{MLP}([\mathbf{z}_{\text{audio}}; \mathbf{z}'_{\text{imu}}])$. The fused representation is then obtained as $\mathbf{z}_{\text{fused}} = \mathbf{z}_{\text{audio}} + \mathbf{\Delta}$. The fused representation $\mathbf{z}_{\text{fused}}$ is then passed to the VAP layers in place of the original audio representation $\mathbf{z}_{\text{audio}}$.

3 Data Collection

To train and evaluate our proposed framework, we collected a dataset pairing conversational audio with synchronized in-ear IMU signals. This was necessary because, to our knowledge, no public dataset provides synchronized conversational audio and in-ear IMU recordings from both speakers in a

dyadic dialogue setting.

3.1 Data Collection Protocol

Nine Japanese university students participated in a data collection experiment. Each participant took part in at least two dialogues conducted remotely via Zoom[1] between two adjacent soundproof booths. Argumentative dialogue was chosen as the target scenario because it is expected to elicit frequent turn-taking between speakers, and participants with opposing stances were assigned to each dialogue. This resulted in a total of 12 collected dialogues.

During the recording, participants completed a brief questionnaire on basic demographic information and their stance on each topic prior to the dialogue. Based on their responses, participants were paired with others holding opposing stances. They then engaged in a dialogue lasting approximately 10–15 minutes. After each dialogue, participants completed an eight-item questionnaire (7-point Likert scale) regarding their subjective experience.

3.2 Recording Setup

Audio was recorded with pin microphones at 48 kHz on a single monaural channel per participant. Each participant wore earbuds (Apple's AirPods3) in the right ear to record IMU signals. This device captures 6-DoF IMU signals (3-axis accelerometer and 3-axis gyroscope) at a maximum sampling rate of 25 Hz (Mollyn et al., 2023) in background mode. Camera views were disabled in Zoom to simulate camera-unavailable conditions.

3.3 Pre-Processing and VA Label Annotation

All audio recordings were downsampled to 16 kHz. We adopted Silero VAD[2] with default settings to obtain Voice Activity Detection (VAD) labels. Dialogues were segmented into 20-second chunks with a 2-second overlap on both sides following the VAP evaluation protocol.

4 Experiments

We evaluated whether Sensor-Augmented VAP improves turn-taking prediction over an audio-only baseline. In addition, we conducted an ablation study to examine the contribution of the IMU signals.

[1] https://www.zoom.com/

[2] https://github.com/snakers4/silero-vad

4.1 Training Conditions

IMU data were preprocessed using per-channel standardization before being fed into the IMU encoder. We used only the accelerometer signals, which directly capture the low-frequency nodding and postural components of head motion. The IMU encoder consists of two layers of Conv1D with kernel size 5 and hidden dimension 64, separated by a ReLU activation. A separate IMU encoder and fusion module were instantiated for each speaker. Both models were trained on NVIDIA L40S GPU for up to 50 epochs using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1×10^{-4}, ReduceLROnPlateau scheduling (Paszke et al., 2019), and early stopping on validation loss. Experiments follow a three-fold cross-validation protocol with the best checkpoint per fold selected for evaluation.

We used the pre-trained VAP model[3] (Inoue et al., 2024a) as a baseline. This model was pre-trained using Japanese dialogue datasets from three different domains.

4.2 Experimental Conditions

For the three-fold cross-validation, dialogues were split at the dialogue level for evaluation. A fixed validation set was used across all folds, while the training and test splits varied per fold. Overlapping speech segments were excluded from evaluation in this study.

We conducted the experiment under both speaker-independent and speaker-dependent conditions. In the speaker-independent setting, we reserved three dialogues whose speakers did not appear in the training data, and used one of these as the test set in each of the three folds. In the speaker-dependent setting, dialogues were split without speaker-based constraints, allowing test speakers to appear in training data. For fair comparison, both settings used three test dialogues per fold, differing only in whether test speakers appeared in training.

Following VAP evaluation practice, we report weighted F1 score, precision, and recall for the shift class (with F1 weighted by class frequency), balanced accuracy, and VAP test loss. Balanced accuracy is the average of per-class recall. VAP loss is the cross-entropy loss over the VAP objective.

4.3 Experimental Results

Table 1 reports results averaged over three test folds. Our Sensor-Augmented VAP improved the

[3] https://huggingface.co/maai-kyoto/vap_jp

	Audio-only (baseline)	w/o IMU (Spk-indep.)	Sensor-Augmented VAP (ours)	
			Spk-indep.	Spk-dep.
Weighted F1 (↑)	0.520	0.598	<u>0.622</u>	**0.653**
Precision (↑)	0.420	<u>0.637</u>	**0.710**	0.624
Recall (↑)	**0.746**	0.573	0.592	<u>0.695</u>
Balanced Acc. (↑)	<u>0.821</u>	0.764	0.775	**0.826**
Loss (VAP) (↓)	2.675	<u>2.476</u>	2.500	**2.433**

Table 1: Turn-taking prediction performance for shift detection. **Bold** and underlined values indicate the best and second-best performance across all settings. The w/o IMU column is an ablation result in the speaker-independent setting where the IMU input is replaced with a zero-masked tensor prior to standardization. Spk-indep. and Spk-dep. denote speaker-independent and speaker-dependent evaluation settings.

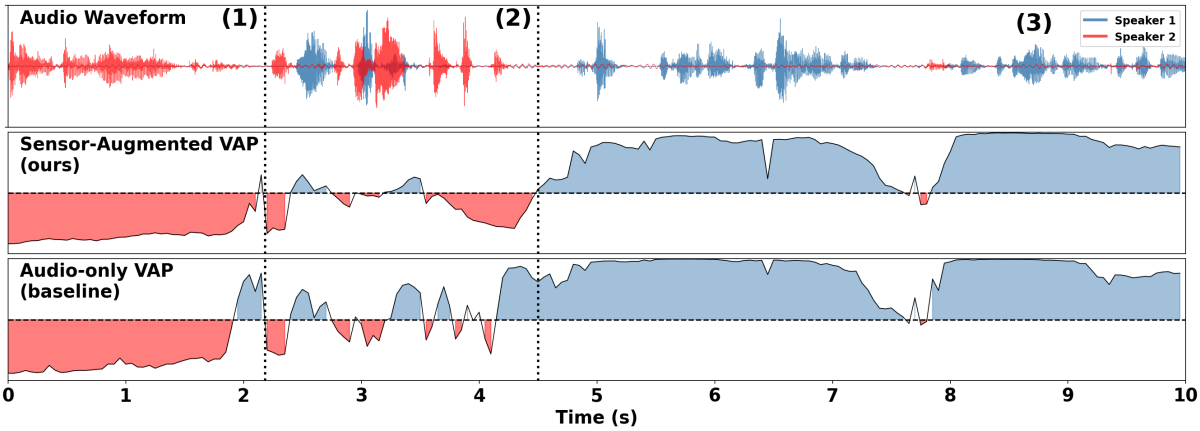

Figure 2: Comparison of p_{now} predictions for a 10-second dialogue segment in a speaker-independent setting. The middle and bottom rows show the proposed Sensor-Augmented VAP and the baseline audio-only VAP, respectively.

weighted F1 score over the audio-only baseline by 0.102 and 0.133 in the speaker-independent and speaker-dependent settings, respectively. The VAP loss also decreased in both settings. Balanced accuracy declined with IMU integration in the speaker-independent setting, which is attributable to the reduced recall for the shift class. The audio-only baseline exhibited low precision. This result is consistent with the report that dialogue systems employing VAP-based turn-taking decisions frequently interrupt user utterances due to false-positive shift detections (Chiba and Higashinaka, 2025). These results suggest that IMU fusion trades recall for precision.

Since our method adds trainable parameters (the IMU encoder and fusion module), the improvement may partly come from adapting the frozen audio representations to our dataset rather than from the head-motion cues. To disentangle the contribution of the IMU signals from that of the fusion module, we conducted an ablation in the speaker-independent setting. In this condition, denoted w/o IMU, the three-axis accelerometer input is replaced with a zero-masked tensor of the same shape, while the per-channel standardization and all trained parameters are kept unchanged. As shown in Table 1, relative to the audio-only baseline, the w/o IMU condition already improves weighted F1 from 0.520 to 0.598 and precision from 0.420 to 0.637. However, the w/o IMU condition still falls short of our full model. This indicates that the head-motion cues contribute beyond the domain adaptation provided by the fusion module. While the IMU-specific improvement in weighted F1 is modest, the IMU input further improves shift-class precision, the metric most relevant to suppressing false-positive shift detections. These results show that, while the fusion module contributes domain adaptation, the IMU encoder provides head-motion cues that further refine shift detection.

To further illustrate the effect of IMU integration, Figure 2 shows p_{now} predictions at the frame level for a 10-second dialogue segment from one test sample in a speaker-independent setting. p_{now} values above and below the threshold indicate that Speaker1 and Speaker2 hold the turn, respectively. Region (1) (0–2 s) corresponds to an interval in which Speaker2 holds the turn; the Sensor-

Augmented VAP maintains stable p_{now} values below the decision threshold. Region (2) (2–4.5 s) corresponds to overlapping utterances from both speakers; the baseline oscillates rapidly near the threshold. In contrast, the proposed model maintains more stable p_{now} values during overlapping speech when both speakers are active. Region (3) (4.5–10 s) corresponds to a clear turn shift to Speaker1. Near the onset at around 5 s, the baseline exceeds the threshold roughly 1 s before Speaker1 speaks. In contrast, the proposed model crosses the threshold closer to the actual speech onset. This qualitative analysis indicates that the advantages of Sensor-Augmented VAP are concentrated in unambiguous single-speaker intervals and around turn-shift boundaries. These results suggest that IMU-derived head-motion signals provide reliable cues for enhancing turn-taking prediction.

5 Conclusion

We proposed Sensor-Augmented VAP, which integrates in-ear IMU signals with a pre-trained VAP model using a lightweight residual fusion module. Experiments show that IMU fusion improves the weighted F1 score for shift detection by 0.102 and 0.133 in speaker-independent and speaker-dependent settings, and reduces VAP loss. The qualitative analysis further suggests that IMU signals contribute to more stable prediction of future voice activity, particularly in single-speaker intervals and around turn-shift boundaries.

Limitations

This study has several limitations. First, we rely solely on accelerometer signals from the earable device. Incorporating gyroscope data would enable more precise characterization of head motion, which we plan to explore in future work. More broadly, since gestures and body posture also cue turn-taking, integrating sensors that capture them may further improve performance. We also did not compare the predictive contribution of IMU-based head-motion cues with that of other non-verbal modalities, such as camera-based visual cues, which we leave for future work. Second, the dataset is small-scale and limited to Japanese university students, leaving the effects of cultural background and age unexplored. Future work should collect larger and more diverse data, both to assess generalizability and to establish statistical robustness.

Ethical Considerations

This study was approved by the Shiba Palace Clinic Ethics Review Committee, and all participants provided written informed consent prior to the experiment. Participation was entirely voluntary. Participants were informed of their right to withdraw at any time without penalty. Any data collected prior to withdrawal would be deleted.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23H00476. This work was supported by JST SPRING, Japan Grant Number JPMJSP2123.

References

Yuya Chiba and Ryuichiro Higashinaka. 2025. Investigating the impact of incremental processing and voice activity projection on spoken dialogue systems. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 3687–3696.

Samantha Gordon Danner, Jelena Krivokapić, and Dani Byrd. 2021. Co-speech movement in conversational turn-taking. *Frontiers in Communication*, 6:779814.

Soumia Dermouche and Catherine Pelachaud. 2019. Generative model of agent's behaviors in human-agent interaction. In *Proceedings of 2019 International Conference on Multimodal Interaction (ICMI)*, pages 375–384.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? Investigations using voice activity projection models. In *Proceedings of the 23rd 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 *Proceedings of Interspeech 2022*, pages 5190–5194.

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

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

Ryo Ishii, Shinichiro Eitoku, Ryota Yokoyama, and Junichi Sawase. 2025. Predicting end-of-turn and backchannel based on multimodal voice activity prediction model. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI)*, pages 446–455.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems. In *Proceedings of Interspeech 2023*, pages 2658–2662.

Saki Mizuno, Nobukatsu Hojo, Satoshi Kobashikawa, and Ryo Masumura. 2023. Next-speaker prediction based on non-verbal information in multi-party video conversation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-body pose estimation using IMUs in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2024. Multimodal voice activity projection for turn-taking and effects on speaker adaptation. *IEICE Transactions on Information and Systems*, pages 445–453.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

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

Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with earables: A systematic literature review and taxonomy of phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 6(3):1–57.

Sam O'Connor Russell and Naomi Harte. 2025a. Visual cues support robust turn-taking prediction in noise. In *Proceedings of Interspeech 2025*, pages 1073–1077.

Sam O'Connor Russell and Naomi Harte. 2025b. Visual cues enhance predictive turn-taking for two-party human interaction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 209–221.

Takeshi Saga and Catherine Pelachaud. 2025. Voice activity projection model with multimodal encoders. *arXiv preprint arXiv:2506.03980*.

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

A Dialogue Topics

Prior to the experiment, participants reported their stance (agree/disagree) on each of the following eight topics. The following eight topics were chosen to facilitate dialogue among Japanese university students. The topics were originally prepared in Japanese, and English translations are provided here for reference.

1. Whether job-hunting should begin as early as possible
2. Whether extracurricular activities are as important as academic work
3. Whether classes should be conducted primarily online
4. Whether campus life should prioritize urban over suburban environments
5. Whether attendance should be prioritized over academic outcomes
6. Whether English-only courses and study abroad should be mandatory
7. Whether curricula should shift toward interdisciplinary breadth over specialization
8. Whether generative AI will make employment harder for students majored in Computer Science