Integrating Respiration into Voice Activity Projection for Enhancing Turn-taking Performance

Takao Obi
Institute of Science Tokyo / Tokyo
smalltail@lr.pi.titech.ac.jp

Kotaro Funakoshi
Institute of Science Tokyo / Tokyo
funakoshi@lr.pi.titech.ac.jp

Abstract

Voice Activity Projection (VAP) models predict upcoming voice activities on a continuous timescale, enabling more nuanced turn-taking behaviors in spoken dialogue systems. Although previous studies have shown robust performance with audio-based VAP, the potential of incorporating additional physiological information, such as respiration, remains relatively unexplored. In this paper, we investigate whether respiratory information can enhance VAP performance in turn-taking. To this end, we collected Japanese dialogue data with synchronized audio and respiratory waveforms, and then we integrated the respiratory information into the VAP model. Our results showed that the VAP model combining audio and respiratory information had better performance than the audio-only model. This finding underscores the potential for improving the turn-taking performance of VAP by incorporating respiration.

1 Introduction

In conversational systems designed for emotional support and customer assistance, it is crucial for the user and the system to engage in smooth and natural dialogues. A key factor in achieving such smooth communication is effective turn-taking, wherein each participant can seamlessly begin and end speaking without awkward interruptions or prolonged silences. In this context, there has been a growing body of research aimed at predicting turn-taking behaviors in spoken dialogue between the user and the system (Skantze, 2017; Roddy et al., 2018).

Recently, Voice Activity Projection (VAP) has been proposed as a method for more natural turn-taking in spoken dialogue (Ekstedt and Skantze, 2022). VAP dynamically models voice activities in dyadic interactions by processing the raw audio signals from both speakers, predicting future voice activity in a series of short time windows (at window lengths of 200 ms, 400 ms, 600 ms, and 800 ms within a 2-second horizon). This approach yields a 256-class prediction representing binary voice activity in each of the four time windows for each speaker. In addition, VAP defines four evaluation tasks, SHIFT/HOLD, SHORT/LONG, SHIFT-prediction, and Backchannel-prediction, to assess how effectively the model can predict turn-shifts and backchannels. Specifically, SHIFT/HOLD tests the model's ability to predict which speaker will take the next turn during mutual silence; SHORT/LONG tests the ability to predict at its onset whether a speaker's utterance will be a short backchannel or a longer utterance; SHIFT-prediction tests the ability to predict whether a turn-shift will occur during active speech; Backchannel-prediction tests the ability to predict future backchannels. Various extensions of VAP have been explored, including the incorporation of prosodic information, gaze, and gestures (Onishi et al., 2023), the extension of multilingual data (Inoue et al., 2024a), and real-time predictions (Inoue et al., 2024b).

In this work, we aim to further enhance VAP by integrating respiratory information, which is a nonverbal cue closely tied to speech production. Prior research about respiration has observed the synchronization of respiratory patterns during turn-taking (Rochet-Capellan and Fuchs, 2014), as well as behaviors such as speakers taking a quick breath when they wish to continue speaking and next speakers inhaling when the previous speaker finishes speaking (Rochet-Capellan and Fuchs, 2014; Torreira et al., 2015; Ishii et al., 2016). These observations have motivated attempts to predict turn continuations, endings, and next-speaker transitions using respiratory signals (Ishii et al., 2016; Włodarczak and Heldner, 2019). In human-system spoken dialogues, respiration has also been investigated to predict a user's speech onset (Włodarczak et al., 2017; Obi and Funakoshi, 2023), indicating that

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 272–276
May 27–30, 2025. ©2025 Association for Computational Linguistics

respiratory information can facilitate smoother turn management.

We focus on the turn-taking performance of VAP and investigate how integrating respiratory information affects the model's performance in SHIFT/HOLD, SHORT/LONG, and SHIFT-prediction tasks. We collected Japanese dialogue data containing both audio and respiratory waveforms, and then we integrated the respiratory information into the VAP model. Our results showed that the VAP model combining audio and respiratory information had better performance than the audio-only model. This finding underscores the usefulness of respiratory information for VAP turn-taking tasks.

2 Data Collection

Because no publicly available dataset for integrating respiration into VAP was available, we collected spoken dialogue data.

2.1 Participants

Thirty-six pairs (72 in total; 32 male and 40 female; ranging in age from 20 to 60) who are native speakers of Japanese were recruited through an agency. Written informed consent was obtained from each participant before data collection. The data collection was pre-approved by the authors' institutional ethical committee.

2.2 Equipment

We employed two main components for data recording.

Audio Recorder: The audio data were recorded using a Kinect v2 microphone made by Microsoft.

Respiration Sensor: The respiratory waveforms were recorded using a device that combines a Biopac TSD201 sensor and a homemade signal amplifier. We used two identical units of this device to record data from two participants in parallel.

2.3 Recording

Because VAP uses separate speaker inputs, we recorded audio and respiration data for each participant in each pair separately. During each recording session, both audio and respiratory waveforms were captured with millisecond-level synchronization by our own recording software, which also logged the start time in milliseconds. This mechanism allowed us to align the data between the two participants in each pair.

Audio Recording: The audio was recorded at 16 kHz with 16-bit PCM (pulse code modulation) encoding.

Respiration Recording: Expansion and contraction of the torso during respiration were recorded using sensor belts around the thorax. The respiration stream was sampled at approximately 90 Hz and stored with corresponding timestamps.

2.4 Procedure

The two participants of each pair were placed in hard-wired soundproof rooms individually and interacted remotely. First, they attached the respiration sensor belts around their thoraxes and sat in front of a screen displaying the other participant. They were then given a discussion topic (e.g., choosing items for survival in a desert) and engaged in a 15-minute discussion. If any time remained after finishing the discussion, they were allowed to talk freely. After a short break, they performed another 15-minute dialogue session on a different discussion topic. We adopted this two-session design to minimize participant fatigue and ensure sufficient dialogue content.

3 Experiments

We investigated whether respiratory information can help improve VAP performance in turn-taking.

3.1 Preprocessing

Data Alignment: Because each participant's data was recorded separately, we aligned the start times of the paired recordings based on the later start time. Specifically, we cut the beginning of the earlier recording to match the start of the later one.

Audio Data: We normalized audio waveforms by amplitude and detected voice activities using Silero-VAD[1]. After that, using the VAP dataset creation scripts[2], we created audio splits and corresponding voice activity labels.

Respiratory Waveform: We first removed drift to mitigate environmental noise. Because the respiration stream was not sampled at perfectly uniform intervals, we applied cubic spline interpolation to resample at 90 Hz. We applied a low-pass filter to remove frequencies above 1 Hz (reflecting the typical human respiratory rate of 0.15–0.40 Hz (Beda et al., 2007)). Finally, because amplitude ranges varied across the two devices, we applied z-score

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

[2] https://github.com/ErikEkstedt/VoiceActivityProjection

normalization to the waveforms. After preprocessing, participants' respiratory rates ranged from 11.8 to 24.3 breaths per minute (BPM), with an average of 16.9 BPM (SD = 2.43).

Data Splitting: For model training and evaluation, we split the data into 80%/15%/5% for training, validation, and test sets, respectively. To properly evaluate the model performance, we split the sets so that they did not contain the same participant pairs.

3.2 VAP Model

We used the public VAP model[2]. The model consists of four main components:

Contrast Predictive Coding (CPC) Encoder: A 5-layer CNN followed by a 1-layer GRU, pretrained on the LibriSpeech Dataset (Panayotov et al., 2015). This encoder is frozen during training.

Self-attention Transformer: A single Transformer layer with 256 dimensions to model each speaker's audio stream separately.

Cross-attention Transformer: Three Transformer layers with 256 dimensions that perform cross-attention between both speakers' encoded audio streams.

Linear layer: Two separate linear layers for multitask learning output probabilities for a 256-class VAP state $p_{\text{vap}}(y)$ and per-speaker VAD $p_{\text{vad}}(s)$.

The model losses are defined as $L = L_{\text{vap}} + L_{\text{vad}}$, where

$$\begin{aligned} L_{\text{vap}} &= -\log p_{\text{vap}}(y), \\ L_{\text{vad}} &= -\sum_{s=1}^{2} \Big[v_s \log p_{\text{vad}}(s) \\ &\quad + (1 - v_s) \log\big(1 - p_{\text{vad}}(s)\big) \Big], \end{aligned}$$

$y \in \{1, \ldots, 256\}$ is the reference VAP index, and $v_s \in \{0, 1\}$ indicates whether participant s is speaking. For brevity, the time frame indexing is omitted, but these calculations apply to all input frames.

3.3 Evaluation

We focused on three VAP tasks for evaluating turn-taking: SHIFT/HOLD, SHORT/LONG, and SHIFT-prediction. We set the input signal segment to 20 seconds, following the findings in (Inoue et al., 2024b), which reported high performance for Japanese with a 20-second segment. To evaluate model performance, we used weighted F1-scores based on the original VAP study (Ekstedt and Skantze, 2022). The training was repeated with random seeds from 1 to 10.

Table 1: Means and variances of weighted F1-scores for turn-taking performance of VAP in evaluation settings. Values marked with * are significantly higher ($p < 0.01$) than the corresponding audio-only baseline based on bootstrap tests.

Evaluation setting	SHIFT/ HOLD	SHORT/ LONG	SHIFT-prediction
Audio-only	0.608 (0.000)	0.794 (0.000)	0.635 (0.001)
Resp-only	0.514 (0.001)	0.574 (0.000)	0.455 (0.001)
Combination	**0.635*** (0.001)	**0.796** (0.000)	**0.648*** (0.002)

We evaluated the model's performance in three settings:

Audio Only: For the baseline audio-only VAP model, we used the original training configuration, including a batch size of 8, a learning rate of 3.63×10^{-4}, a weight decay of 0.001, and the AdamW optimizer. We trained for 20 epochs and used the model checkpoint that yielded the lowest validation loss for testing.

Respiration Only: We replaced the encoder with a similarly structured one modified to handle respiratory waveforms. Unlike the CPC encoder (which was frozen for audio), we trained the respiratory encoder along with the other layers. We increased the total epochs to 30 based on validation loss trends, keeping all other hyperparameters the same.

Combination: To explore a straightforward way of combining respiratory information with audio, we used separate encoders and attention transformers for each modality. We then concatenated the outputs from each cross-attention before passing them to the linear layers. Training settings were identical to the audio-only.

4 Results

The experimental results are shown in Table 1. As shown in Table 1, the highest performance was achieved when voice and respiratory waveforms were used together. The combination model achieved significantly higher SHIFT/HOLD and SHIFT-prediction F1-scores ($p < 0.01$) than the audio-only baseline, using bootstrap resampling methods[3].

[3] https://github.com/fpgdubost/bstrap

5 Discussion

Our results showed that combining respiratory information with audio improves VAP performance in turn-taking, especially SHIFT/HOLD and SHIFT-prediction tasks (Table 1). This enhancement likely arises because respiratory information provides additional cues about a speaker's readiness or intention to speak, helping reduce uncertainty around turn boundaries. This finding indicates that respiration is valuable supplementary information for VAP turn-taking prediction.

6 Limitations and Future Work

Although our experiments demonstrated the potential benefits of integrating respiratory information into VAP, several limitations remain.

First, the amount of data used in this study was relatively small, and participants took part in remote dialogues. To further validate the effectiveness of respiratory information for VAP, we plan to collect additional data in more diverse conversational settings.

Second, we used contact-based respiration sensors to record respiratory waveforms. However, for real-world spoken dialogue systems, it is preferable to measure a user's respiration in a non-contact manner. By combining our approach with non-contact respiratory estimation methods (Obi and Funakoshi, 2023; Matheus et al., 2023), which capture users' respiratory information using only an RGB camera, we can eliminate the need for wearable sensors. We will adopt this combined approach to implement VAP with integrated respiration in real-world dialogues.

Third, the method of combining audio and respiratory information in our model was quite simplistic, relying on a straightforward concatenation of features. By improving the model architecture or employing more advanced fusion strategies, it may be possible to more accurately integrate voice and respiratory signals. We will explore these more sophisticated approaches to better leverage respiratory information for VAP.

Finally, although the original VAP includes Backchannel-prediction, we focused on SHIFT/HOLD, SHORT/LONG, and SHIFT-prediction in this study. Evaluating the effectiveness of respiratory information on Backchannel-prediction remains an important direction for future work and may further clarify the potential of respiratory information.

7 Conclusion

In this work, we explored how respiratory information can be combined with audio to improve Voice Activity Projection (VAP). We collected Japanese dialogue data with synchronized audio and respiratory waveforms to investigate the efficacy of combining this information for VAP. Our results indicate that combining audio and respiratory information can improve VAP performance in turn-taking. This finding underscores the potential value of leveraging respiratory information to enhance the turn-taking performance of VAP.

We will explore more sophisticated fusion mechanisms that might better integrate respiratory information into VAP.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP22H04859 and JST BOOST, Japan Grant Number JPMJBS2430.

References

Alessandro Beda, Frederico C. Jandre, David I.W. Phillips, Antonio Giannella-Neto, and David M. Simpson. 2007. Heart-rate and blood-pressure variability during psychophysiological tasks involving speech: Influence of respiration. *Psychophysiology*, 44(5):767–778.

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

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. 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 2024)*, pages 11873–11883, Torino, Italia. ELRA and ICCL.

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)*.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Using respiration to predict who will speak next and when in multiparty meetings. *ACM Trans. Interact. Intell. Syst.*, 6(2).

Kayla Matheus, Ellie Mamantov, Marynel Vázquez, and Brian Scassellati. 2023. Deep breathing phase classification with a social robot for mental health. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 153–162, New York, NY, USA. Association for Computing Machinery.

Takao Obi and Kotaro Funakoshi. 2023. Video-based respiratory waveform estimation in dialogue: A novel task and dataset for human-machine interaction. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 649–660. Association for Computing Machinery.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, HAI '23, page 13–21, New York, NY, USA. Association for Computing Machinery.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Amélie Rochet-Capellan and Susanne Fuchs. 2014. Take a breath and take the turn: How breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B*, 369.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale rnns. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 186–190, New York, NY, USA. Association for Computing Machinery.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany. Association for Computational Linguistics.

Francisco Torreira, Sara Bögels, and Stephen C. Levinson. 2015. Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology*, 6.

Marcin Włodarczak and Mattias Heldner. 2019. Breathing in conversation — what we've learned. In *1st International Seminar on the Foundations of Speech : BREATHING, PAUSING, AND THE VOICE, 1st –3rd December 2019 in Sønderborg, Denmark : Conference Proceedings*, pages 13–15.

Marcin Włodarczak, Kornel Laskowski, Mattias Heldner, and Kätlin Aare. 2017. Improving Prediction of Speech Activity Using Multi-Participant Respiratory State. In *Proc. Interspeech 2017*, pages 1666–1670.