Real-time and Continuous Turn-taking Prediction Using Voice Activity Projection

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara and Gabriel Skantze

Abstract A demonstration of a real-time and continuous turn-taking prediction system is presented. The system is based on a voice activity projection (VAP) model, which directly maps dialogue stereo audio to future voice activities. The VAP model includes contrastive predictive coding (CPC) and self-attention transformers, followed by a cross-attention transformer. We examine the effect of the input context audio length and demonstrate that the proposed system can operate in real-time with CPU settings, with minimal performance degradation.

1 Introduction

Turn-taking is a crucial aspect of human spoken interaction and therefore an important function to model in spoken dialogue systems (SDSs) [1]. The problem of turn-taking in SDSs involves predicting the end of the user's turn and smoothly initiating the system's turn. Despite recent progress in large language models (LLMs) that enable the generation of sophisticated responses in SDSs, turn-taking is still typically handled in a simplistic manner. In practical SDSs, turn-taking is commonly implemented using a simple silence timeout threshold, usually around 1 or 2 seconds,

Koji Inoue
Kyoto University, Japan, e-mail: inoue.koji.3x@kyoto-u.ac.jp

Bing'er Jiang
KTH Royal Institute of Technology, Sweden, e-mail: binger@kth.se

Erik Ekstedt
KTH Royal Institute of Technology, Sweden e-mail: erikekst@kth.se

Tatsuya Kawahara
Kyoto University, Japan, e-mail: kawahara@i.kyoto-u.ac.jp

Gabriel Skantze
KTH Royal Institute of Technology, Sweden e-mail: skantze@kth.se

to indicate the end of a turn. However, silence is not a reliable indicator, as pauses within turns are usually longer than pauses between turns in human-human interaction [2, 3]. As a result, SDSs often suffer from long response delays or frequent interruptions during pauses.

To address this problem, many proposals have been made for end-of-turn prediction models that consider verbal and non-verbal cues (such as linguistic and prosodic features) of preceding user utterances, in order to predict whether the user is just pausing (a *hold*), or whether the turn is yielded (a *shift*). The models used for this prediction range from recurrent neural networks (RNNs) [4, 5] to transformers [6, 7, 8, 9]. However, in general, the problem setting mostly involves binary prediction of whether the user's turn ends, which is done at the utterance level. When implementing this in a spoken dialogue system, it is necessary to determine the appropriate waiting time when the system takes a turn [10, 11, 12, 7]. Therefore, it is preferable to always perform turn-taking prediction in continuous time frames, rather than at the utterance level.

We have proposed a continuous turn-taking prediction model called voice activity projection (VAP) [13]. The VAP model utilizes multi-layer transformers to predict the near future voice activities of dialogue participants by processing the raw audio signals from the two speakers in a dyadic dialogue. However, the processing time of the transformer depends on the length of the input context. It is unclear whether the VAP model can function effectively in real-time environments for SDSs. In this demonstration, we showcase the real-time processing of VAP for SDSs in CPU environments, investigating the impact of the input context audio length on performance, ensuring minimal degradation.

2 Voice activity projection

As stated above, the main objective for the VAP model is to predict future voice activity of two speakers in a dialogue, based on raw audio input. A detailed explanation of the VAP model can be found in its original paper [14]. The source codes for the VAP model are publicly available[1].

Fig. 1 illustrates the VAP model architecture including a pre-trained contrastive predictive coding (CPC) model for encoding the input audio signal [15], followed by separate one-layer self-attention transformer for each channel. The outputs from the two channels are further processed by a cross-attention transformer, which captures interactive information between the channels [16]. The final output is obtained by concatenating the outputs of the two transformers and passed through linear layers for multitask learning, including the VAP objective and voice activity detection (VAD) subtask.

The VAP model predicts voice activities for two speakers within a two-second time window by predicting the joint activity of both speakers over four binary bins. The

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

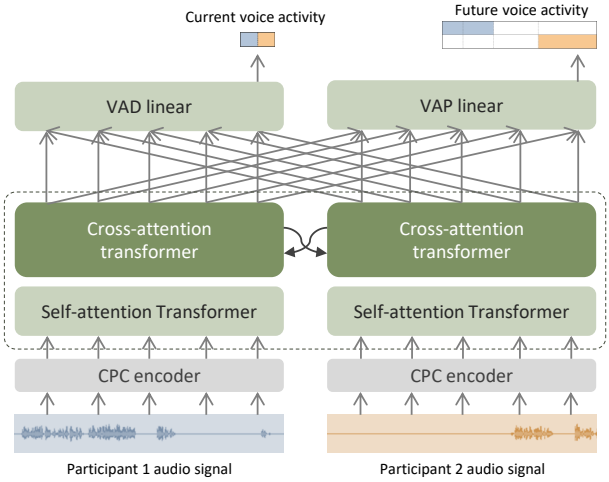

Fig. 1 Architecture of the VAP model

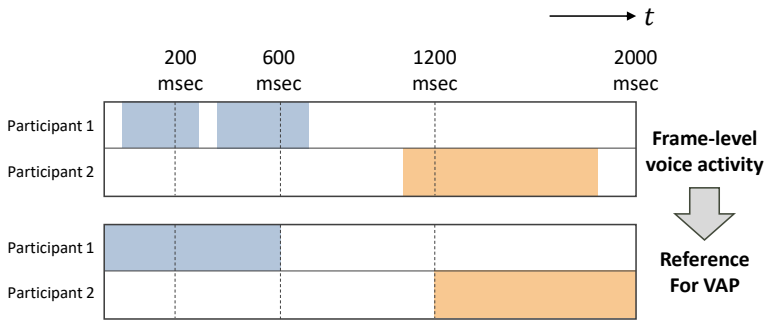

Fig. 2 Discretizing bins for the VAP model

objective is to classify the future two seconds into one of 256 possible activations. The bins are discretized as 'voiced' or 'unvoiced' based on the ratio of voiced frames, as depicted in Fig. 2.

The probability distribution over the possible VAP states predicts the turn-taking dynamics. However, it is complex to use and interpret directly. To simplify, we can sum up the probability values of each participant's bins in the 0-200 msec and 200-600 msec regions. Then, apply softmax to both sums to obtain $p_{now}(s)$. This represents a short-term future voice activity prediction for participant s (i.e., "how likely is the participant to speak in the next 600 msec"). Similarly, for the 600-1200 msec and 1200-2000 msec bins, we use $p_{future}(s)$ as a slightly longer-term future voice activity prediction. Note that this is just one example of how the VAP output can be used.

In Figure 3, an example of a GUI for output based on the VAP model is shown. This example depicts the alternating transition of speaking turn from a yellow participant

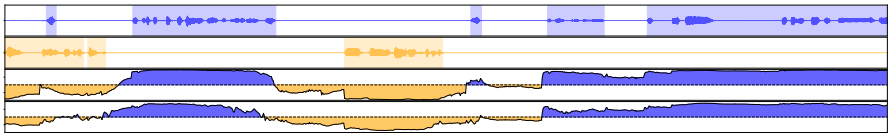

Fig. 3 Output example of multilingual VAP - Each graph consists of, from top to bottom, input waveforms of both participants, near future voiced probability (p_{now}), and future voiced probability (p_{future}) among participants.

to a blue participant. At each time frame, the values of $p_{now}(s)$ and $p_{future}(s)$ are indicated. During the initial transition, the value of $p_{future}(s)$ for the blue participant becomes higher just before the end of the yellow participant's turn. Immediately after the end of the speech, $p_{now}(s)$ also predicts the transition of turns. Similarly, the subsequent transition from blue to yellow can be successfully predicted. There is also a longer mutual pause in the transition from yellow to blue, indicating an ambiguous prediction where $p_{now}(s)$ and $p_{future}(s)$ are kept around even among the two participants. This suggests that the next speaker is not clear in this case, so it can be said that the model correctly predicts this kind of ambiguous place. In the subsequent scene where the blue participant continues to hold the speaking turn, even with pauses, the values related to blue remain high, correctly predicting the turn-holding. In this demonstration, a GUI application is presented to display input waveforms and output model values in real-time.

3 Performance

We conducted an investigation into the accuracy and processing speed of the VAP model. The VAP model utilizes transformers, which leads to slower inference speed as the input data length increases. In order to mitigate this issue, we performed a study on the accuracy and inference speed of turn-taking prediction by reducing the input sequence length of the transformer. It is worth noting that a gated recurrent unit (GRU) inside the audio encoder CPC is an auto-regressive model, so we inputted all the audio sequences to the GRU, for durations of up to 20 seconds.

The model in this experiment was trained using the Japanese Travel Agency Task Dialogue dataset [17]. The training data consisted of 92.5 hours, with 11.5 hours allocated for both the validation and test sets. The evaluation scheme is based on previous research [13], which focuses on predicting the next speaker during periods of mutual silence lasting longer than 0.25 seconds. The model predicts whether there will be a turn transition or holding by calculating and averaging the value of $p_{now}(s)$ over time during mutual silence. The higher value between the two participants is then considered the prediction result. In the test data used for this study, there were 1,023 instances of turn transition and 1,371 instances of turn hold. The evaluation metric used is balanced accuracy which calculates the accuracies for both positive and negative examples and then takes the average of them. Note that the random

Table 1 Performance of turn-taking prediction with different input context length

Input length [sec.]	Balanced accuracy [%]	Inference time / frame [msec] (real-time factor)
20.0	74.20	273.84 (13.69)
10.0	75.73	94.93 (4.75)
5.0	75.01	33.66 (1.68)
3.0	75.75	30.54 (1.53)
1.0	76.16	14.61 (0.73)
0.5	75.41	13.11 (0.66)
0.3	71.50	12.19 (0.61)
0.1	62.81	12.45 (0.62)

prediction score of this metric would be 0.5. The VAP model parameters consist of a self-attention transformer with 1 layer for each channel, and a cross-channel transformer with 3 layers. Both have 4 attention heads and a unit size of 256. The inference was performed on an Intel Xeon Gold 6128 CPU with a clock speed of 3.40 GHz.

Table 1 demonstrates the changes in prediction performance and inference time. Even when the input sequence length is limited to approximately 1 second in the transformer, there is no decrease in prediction performance. This indicates that the trained VAP model, assuming 50 frames per second in the input, can process in real-time with sufficient inference time. This result also suggests that the GRU of CPC has likely acquired a certain level of representation regarding the information in the input sequence. It is important to note that when the input sequence is shortened to less than 0.3 seconds, the prediction performance starts to degrade. In summary, these results indicate that by restricting the input sequence to around 1 second in the transformer, real-time processing becomes feasible without compromising accuracy.

Furthermore, models for English and Mandarin Chinese have also been developed, making this demonstration multi-lingual. The English model was trained using the Switchboard dataset [18], while the Mandarin Chinese model was trained using the HKUST Mandarin telephone speech corpus [19]. Both models yielded similar results as mentioned above.

4 Conclusion

We presented a real-time demonstration of a continuous turn-taking prediction model called voice activity projection. Conducting investigations on the impact of the sequence length of input to the transformer, we discovered that even with a 1-second input sequence, the prediction accuracy remains unaffected. This allows the model to operate in real-time on a CPU environment. In the future, we plan to integrate this system into spoken dialogue systems and evaluate the effectiveness of this turn-taking prediction model through dialogue experiments.

Acknowledgement

This work was supported by JST ACT-X (JPMJAX2103), JST Moonshot R&D (JPMJPS2011), and JSPS KAKENHI (JP19H05691 and JP23K16901).

References

1. Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178, 2021.
2. Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.
3. Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731):1–17, 2015.
4. Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 220–230, 2017.
5. Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka. Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. In *INTERSPEECH*, pages 1661–1665, 2017.
6. Erik Ekstedt and Gabriel Skantze. TurnGPT: A Transformer-based language model for predicting turn-taking in spoken dialog. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2981–2990, 2020.
7. Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. Response timing estimation for spoken dialog systems based on syntactic completeness prediction. In *Spoken Language Technology Workshop (SLT)*, pages 369–374, 2023.
8. Toshiki Muromachi and Yoshinobu Kano. Estimation of Listening Response Timing by Generative Model and Parameter Control of Response Substantialness Using Dynamic-Prompt-Tune. In *INTERSPEECH*, pages 2638–2642, 2023.
9. Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems. In *INTERSPEECH*, pages 2658–2662, 2023.
10. Antoine Raux and Maxine Eskenazi. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing*, 9(1):1–23, 2012.
11. Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 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, 2017.
12. Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In *International Conference on Multimodal Interaction (ICMI)*, pages 78–86, 2018.
13. Erik Ekstedt and Gabriel Skantze. Voice Activity Projection: Self-supervised learning of turn-taking events. In *INTERSPEECH*, pages 5190–5194, 2022.
14. Erik Ekstedt. *Predictive Modeling of Turn-Taking in Spoken Dialogue: Computational Approaches for the Analysis of Turn-Taking in Humans and Spoken Dialogue Systems*. PhD thesis, KTH Royal Institute of Technology, 2023.
15. Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418, 2020.

16. Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023.
17. Michimasa Inaba, Yuya Chiba, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. Collection and analysis of travel agency task dialogues with age-diverse speakers. In *Language Resources and Evaluation Conference (LREC)*, pages 5759–5767, 2022.
18. John J Godfrey, Edward C Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 517–520, 1992.
19. Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff. HKUST/MTS: A very large scale mandarin telephone speech corpus. In *International Symposium Chinese Spoken Language Processing (ISCSLP)*, pages 724–735, 2006.