INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Show & Tell: Voice Activity Projection and Turn-taking

Erik Ekstedt, Gabriel Skantze

KTH, Royal Institute of Technology, Sweden

erikekst@kth.se, skantze@kth.se

Abstract

We present Voice Activity Projection (VAP), a model trained on spontaneous spoken dialog with the objective to incrementally predict future voice activity. Similar to a language model, it is trained through self-supervised learning and outputs a probability distribution over discrete states that corresponds to the joint future voice activity of the dialog interlocutors. The model is well-defined over overlapping speech regions, resilient towards microphone "bleed-over" and considers the speech of both speakers (e.g., a user and an agent) to provide the most likely next speaker. VAP is a general turn-taking model which can serve as the base for turn-taking decisions in spoken dialog systems, an automatic tool useful for linguistics and conversational analysis, an automatic evaluation metric for conversational text-to-speech models, and possibly many other tasks related to spoken dialog interaction.

Index Terms: turn-taking, spoken dialog, text-to-speech

1. Introduction

Turn-taking is one of the most fundamental aspects of human spoken interaction and can be thought of as the coordination of turns, or the organization of speech activity, that promotes efficient incremental exchange of information and alleviates the need for interlocutors to listen and speak at the same time [1]. This fundamental role arguably makes a turn-taking focused perspective useful for a wide range of tasks and research topics concerning spoken dialog systems (SDS) or the modeling of turn-taking in general.

However, analyses of turn-taking have largely been constrained to areas like conversational analysis and psycholinguistics where the focus is to understand human-human turn coordination and the signals used to this effect. SDSs have historically been constrained by technology and researchers have focused on aspects such as the ability to generate speech (TTS), automating responses (NLP) and understanding the content of human speech (ASR). Considering that these technologies have been rapidly maturing, we now enter an era where human-like psychological and conversational aspects of speech, like turn-taking, are of increasing importance.

In this show-and-tell, we present Voice Activity Projection (VAP), a self-supervised, general computational model of turn-taking, that can be used to facilitate progress on a wide range of tasks grounded in spoken interaction. It can be used to control the turn-taking decisions in SDSs, classify suitable locations for the generation of backchannels, resolving overlapping speech by discriminating user backchannels from interruptions, or vice versa. It can be used together with TTS systems as an automatic evaluation metric or directly during training to guide the speech generation to contain appropriate prosodic turn-taking signals.

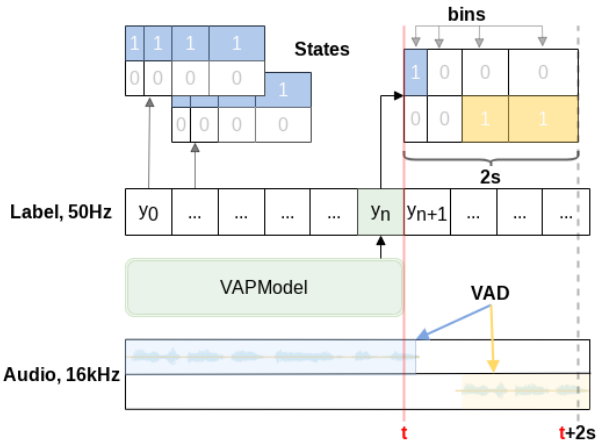


Figure 1: *The VAP model receives a stereo channel audio input and predicts a discrete label, y, at frame, n, corresponding to time, t. Each label represents a state that consists of 8 binary bins, 4 for each speaker, spanning the next 2s of dialog. There are 256 possible states, i.e. the size of the VAP vocabulary*

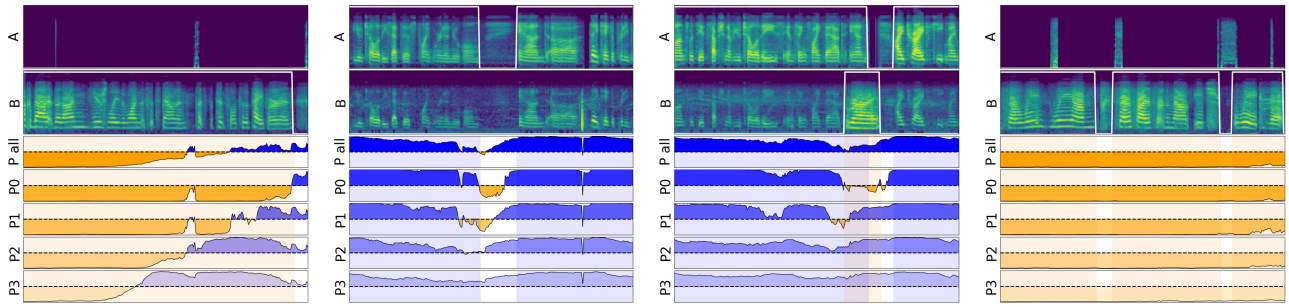
Furthermore, it may be useful as a tool for linguists to label large sets of conversational corpora, find unlikely turn-shifts or interruptions, or as an automatic approach to investigate the role of fillers, breaths, laughter, etc., and their relation to turn-taking.

2. Model

The Voice Activity Projection [2] (VAP) objective is defined as the incremental prediction of discrete future voice activity (VA) states over spoken dialog. The VA is defined in binary terms (speech/no-speech), and the two speakers' future activities are jointly encoded into a discrete state that represents the upcoming 2s of dialog. The states are defined by discretizing the 2s windows of activity into eight smaller sub-state-bins, four for each speaker, of increasing duration (0.2s, 0.4s, 0.6s, 0.8s) and are considered active if they contain a majority of VA. This discretization step produce $2^8 = 256$ possible discrete states (labels) to predict during training, see Figure 1. The model operates on two channel waveforms (one for each speaker) that are processed by a pre-trained audio encoder followed by a GPT-like transformer, using shared weights and cross-attention between the channels. It is trained on the Switchboard and Fisher corpora with code and pre-trained weights publicly available[1].

During inference we scale each sub-state-bin, with their as-

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



(a) *A shift-prediction from the yellow speaker to the blue speaker.*

(b) *A backchannel prediction. Note how p1, p2 differ from p3, p4.*

(c) *Winner at overlapping speech? Note how p1, p2, p3 favors blue.*

(d) *A likely turn-hold without predicted backchannels, notice the .*

Figure 2: *(Top) Stereo channel mel-spectrogram with VA outlines (white) and the presence of natural "bleed-over" between channels ((b), (c)). Decreasing order from the top, the next speaker probability of the entire projection window (p-all), the first, second, third and last bin. The VA is shown as shaded areas (blue/yellow) and the next speaker predictions as black curves.*

sociated label probability, and combine all contributions to a single aggregate state representation, exemplified in Figure 3. The probabilities are normalized across speakers to produce a value between 0 and 1 that represent the prediction probability of the speakers being active in the corresponding region. We consider the contribution over all bins as the *p-all* probability and the individual bins as *p0, p1, p2, p3* for the increasingly distant bins.

To highlight the output provided be the VAP model we show 4 different 5 s examples of spontaneous spoken dialog in Figure 2. In Figure 2a, we show a turn-shift prediction where the probabilities associated with different bins provide predictions over different time resolutions. Figure 2b shows a backchannel prediction to highlight the models ability to predict activity for different speakers over different prediction times. In Figure 2c we visualize the output of overlapping speech (a backchannel from yellow) during which the model favors blue for the later bins. If the blue speaker was a SDS this can provide a way to avoid being "interrupted" and continue on with the planned utterance. Finally, we highlight a clear turn-hold prediction, shown in Figure 2d. These examples are meant to illustrate how one can interpret the model predictions and provide a basis for thinking about other types of problem that could benefit from using a trained VAP model.

3. Ongoing and Future Work

Using the VAP model to control turn-taking decisions for a SDS opens up a variety of behaviors that requires further research. It enables the automation of interruptions, backchannel generation and finer grained control of the different types of turn-taking behavior to imbue the system. The predictions can be thresholded to produce binary actions or using the exact state predictions to control for the amount of silence before taking the turn.

Inspired by linguistic research on human-human turn-taking we have analyzed the models sensitivity to various prosodic cues and shown that the VAP training objective converge towards similar sensitivity to cues as humans [3]. Additionally, we have used the model [4] as an automatic tool to gain insight into the role of fillers ("uh", "um") and their effect on turn-taking. Our work showed that, contrary to some linguistic literature, that the two fillers have the same effect on turn-taking and that it is the prosodic realization rather than semantics which provides a statistically significant effect.

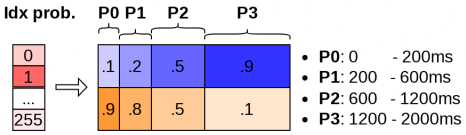


Figure 3: *Illustration of aggregate state probabilities used during inference.*

Furthermore, we applied the model to evaluate the ability to produce coherent turn-taking signals in commercial TTS service providers and open-source architectures . Our research indicate that while these systems are both natural and intelligible, they still lack the ability to produce coherent turn-taking signals. Ongoing research also focuses on using the VAP model to label commonly used TTS corpora, enabling training of style controllable systems, to facilitate the generation of turn-holding/yielding signals without the need to collect additional data. Finally, we believe that the VAP model could be incorporated directly into the training of TTS systems to learn specific turn-taking cues simply by optimizing the predictions of the VAP model applied to the generated speech.

In conclusion, we believe that turn-taking is becoming more important as other SDS based technologies continue to improve and that VAP is a useful model that can provide much benefit to a wide range of researchers focused on improving the capabilities of SDS and/or interesting in further our understanding of human spoken interaction in general.

4. References

[1] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, 1974.

[2] E. Ekstedt and G. Skantze, "Voice Activity Projection: Self-supervised Learning of Turn-taking Events," in *Interspeech*, 2022.

[3] ——, "How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models," in *Proc. 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2022.

[4] B. Jiang, E. Ekstedt, and G. Skantze, "What makes a good pause? Investigating the turn-holding effects of fillers," in *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023.