BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Endpointing for spoken dialogue

Sathvik Udupa, Petr Schwarz

Abstract
Accurate, low-latency endpointing is an important part of spoken dialogue systems. Traditionally, endpointers rely on spectrum-based features to represent audio. Building upon recent works with neural audio codecs, we propose real-time speech endpointing for multi-turn spoken dialogue, using streaming low-bit rate neural audio codec features. Further, to reduce the cutoff error rate, we introduce label delay training. This technique achieves a 31% relative reduction for spectrum-based and 12% for codec-based endpointing at 200 ms median latency. Moreover, with label delay training, codec-based endpointing demonstrates a 32% relative reduction in cutoff error rate. Finally, we demonstrate efficient integration with a codec-based speech large language model, improving response time by 900 ms median latency and cutoff error by 30%.

*udupa@fit.vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Advancements in spoken dialogue systems [1] are leading to widespread adoption of speech technologies. Consequently, speech is a primary way users interact with applications ranging from voice assistants [2] and speech Large Language Models (LLM) [3] to systems like customer support, emergency services. For such applications, it is beneficial to know when a user has stopped speaking so that further processing can begin. This process is referred to as speech *endpointing*. Developing an effective endpointer requires balancing latency and accuracy. Errors can cause incomplete data and hurt the user experience, while accurate, low-latency endpoints greatly improve satisfaction.

We propose to use neural audio codecs (NAC) [4] as speech feature representation for the endpointing task. Neural audio codecs, originally developed for audio compression, provide discrete audio representations. Compared to traditional speech features such as Mel-Spectrograms, neural audio codecs can represent audio at low bit-rates [5] while maintaining high reconstruction quality. Unlike non-streaming, higher bit-rate self-supervised learning (SSL) features, streaming NACs [5, 4] are uniquely suited for real-time applications. Furthermore, utilising neural codec-based endpointer offers multi-task capability, enabling integration of codec-based endpointer with codec-based ASRs and codec-based speech LLMs.

While using NAC for endpointing has benefits, using it can lead to a sub-optimal performance [6] mainly due to the presence of short pauses and filler words. Utilization of speech features only may prove insufficient to prevent the premature triggering of the endpointer due to silences in the middle of a turn. Several approaches have been explored to enhance endpointer accuracy, including the use of ASR features [6], multi-task training with ASR [7, 8]. Further, many works have used additional pause labels [9].

We propose label delay training, a novel approach to reduce errors in standalone speech endpointers without additional features or pause labels. Recent multi-stream speech language models have used delayed tokens [10, 4] to reduce dependencies between parallel output streams. In contrast, we apply label delay to a single output stream, specifically to optimize the latency-accuracy trade-off and enhance endpointer performance. By shifting target labels, we encourage delayed, high-confidence predictions, which leads to inherently fewer errors.

The contributions of our work are as follows -

- Using streaming neural codec features for endpointing in multi-turn spoken dialogue.
- Label delay training for improved endpointing.
- Integrating codec-based endpointer with codec-based speech LLM to achieve low response latency and cutoff error.

2. Proposed methodology

2.1 Baseline endpointer

For streaming endpointing, we use a unidirectional Long Short-Term Memory (LSTM). The LSTM endpointer, operating at the input frame rate, predicts probabilities for four labels per frame - <user>, <system>, <user-end>, <system-end>. Endpoint triggering is based on a threshold applied to the turn-end probability.

2.2 Neural audio codecs

We propose using features from NACs [5]. We extract code vectors by indexing the NAC codebook with the encoded codes and use these pre-trained code vectors as the input features for the LSTM endpointer.

2.3 Label delay

To encourage the endpointer to delay predictions, we introduce a delay to the labels during training. This induces an implicit latency learned from the training objective, without requiring explicit latency during inference. Given the label sequence $\mathbf{Y} = (y_1, y_2, \ldots, y_T)$ with T frames, we generate the label sequence with delay τ, $\mathbf{Y}_\emptyset$ as shown in Poster figure 3.

3. Experimental setup

3.1 Dataset

We train and evaluate our model using the the SpokenWoZ corpus [11]. The dataset consists of many multi-turn spoken dialogues between two speakers.

3.2 Training

All models are trained using *PyTorch*. and the Mimi NACs are used from *transformers*[1] library. All models were trained for 50 epochs using frame-level cross-entropy loss with Adam optimizer.

3.3 Evaluation

We use three evaluation metrics commonly used in endpointing [12]. We use *ep50* and *ep90* to measure latency, which corresponds to the time difference between the endpoint timestamp and the true turn-end timestamp. *ep50* corresponds to the median latency, and *ep90* represents the worst case - the tail latency at the 90th percentile over all <user> turns. The cutoff error rate is captured by *ep-cutoff* - the proportion of <user> turns where the endpointer is triggered before the true turn-end.

[1] https://github.com/huggingface/transformers

3.4 Neural codec endpointer with speech LLM

We integrate Mimi-based endpointer with an open-sourced speech LLM, Moshi [4], which uses the Mimi NAC. We simulate interaction with the LLM using SpokenWoz test sentences and measure Moshi latency and cutoff rate.

4. Results and discussions

4.1 Neural audio codec performance

Poster figure 5 shows the metrics using baseline versus proposed NACs as input. Mimi NAC achieves lower cutoff rates across both median and worst-case latencies, outperforming the baseline. This is likely due to the larger size of Mimi NAC, and the semantic context distillation in the Mimi pre-training [4].

4.2 Label delay

Label delay training achieves substantial error reduction at fixed latency of 200 ms, with relative gains of 31% and 12% for Mel-Spectrogram and Mimi NAC features, respectively (Poster figure 6). Further, Mimi-based endpointer consistently outperforms the baseline with label delay training. At 200 ms median latency, Mimi reduces the cutoff rate to 5.28% compared to baseline's 7.76% – a significant 32% relative error reduction – while incurring an 80 ms worst-case latency increase. Meanwhile, if we consider fixed worst-case latency of 800 ms, we observe that Mimi based endpointer achieves obtains a lower error rate of 5.28%, when compared to 6.05% from baseline, along with 80 ms reduction in median latency.

4.3 Error analysis

We analyze the locations where the endpointer cut off the user's speech and consider the mid-silence duration at these points. The results (Poster figure 8), show that label delay training effectively teaches the endpointer to avoid triggering short pauses.

4.4 Using endpointer with Moshi speech LLM

Poster figure 7 shows the response latency and cutoff rate when interacting with the Moshi speech LLM as explained. Integrating the Mimi-based endpointer with Moshi significantly improves the latency and cutoff rate metrics across different inference configurations. Using an endpointer makes it possible to explicitly control the response timing.

Acknowledgements

I would like to thank my supervisor Petr Schwarz, along with Shinji Watanabe and Jan Černocký for their help.

References

[1] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. WavChat: A Survey of Spoken Dialogue Models. *arXiv preprint arXiv:2411.13577*, 2024.

[2] Lawal Ibrahim Dutsinma Faruk, Mohammad Dawood Babakerkhell, Pornchai Mongkolnam, Vithida Chongsuphajaisiddhi, Suree Funilkul, and Debajyoti Pal. A review of subjective scales measuring the user experience of voice assistants. *IEEE Access*, 2024.

[3] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

[5] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

[6] Roland Maas, Ariya Rastrow, Chengyuan Ma, Guitang Lan, Kyle Goehner, Gautam Tiwari, Shaun Joseph, and Björn Hoffmeister. Combining acoustic embeddings and decoding features for end-of-utterance detection in real-time far-field speech recognition systems. In *ICASSP*, pages 5544–5548. IEEE, 2018.

[7] Shuo-Yiin Chang, Rohit Prabhavalkar, Yanzhang He, Tara N Sainath, and Gabor Simko. Joint endpointing and decoding with end-to-end models. In *ICASSP*, pages 5626–5630. IEEE, 2019.

[8] Shuo-Yiin Chang, Bo Li, Tara Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. Turn-Taking Prediction for Natural Conversational Speech. In *Interspeech*, pages 1821–1825, 2022.

[9] Dawei Liang, Hang Su, Tarun Singh, Jay Mahadeokar, Shanil Puri, Jiedan Zhu, Edison Thomaz, and Mike Seltzer. Dynamic speech endpoint detection with regression targets. In *ICASSP*, pages 1–5. IEEE, 2023.

[10] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *NeurIPS*, 36, 2024.

[11] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. SpokenWOZ: a large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *NeurIPS*, 2024.

[12] Matt Shannon, Gabor Simko, Shuo-Yiin Chang, and Carolina Parada. Improved End-of-Query Detection for Streaming Speech Recognition. In *Interspeech*, page 1909–1913. ISCA, August 2017.