Dynamic Speech Endpoint Detection with Regression Targets

Dawei Liang^, Hang Su*, Tarun Singh*, Jay Mahadeokar*, Shanil Puri*, Jiedan Zhu*, Edison Thomaz^, Mike Seltzer*

^ The University of Texas at Austin, * Meta AI

Abstract

Traditionally, speech end-pointing is based on classification along with arbitrary binary targets. This paper proposes a novel regression-based speech end-pointing model, which enables an end-pointer to adjust its detection behavior based on the context of user queries. Specifically, we present a pause modeling method and show its effectiveness for dynamic end-pointing.

Problem Statement

End-point detection (end-pointing) is the process to automatically detect when a user of a voice assistant has finished a query.

Difference from voice activity detection (VAD): Common VAD systems do not aim to differentiate end-of-sentence pauses (endpoints) and within-sentence pauses.

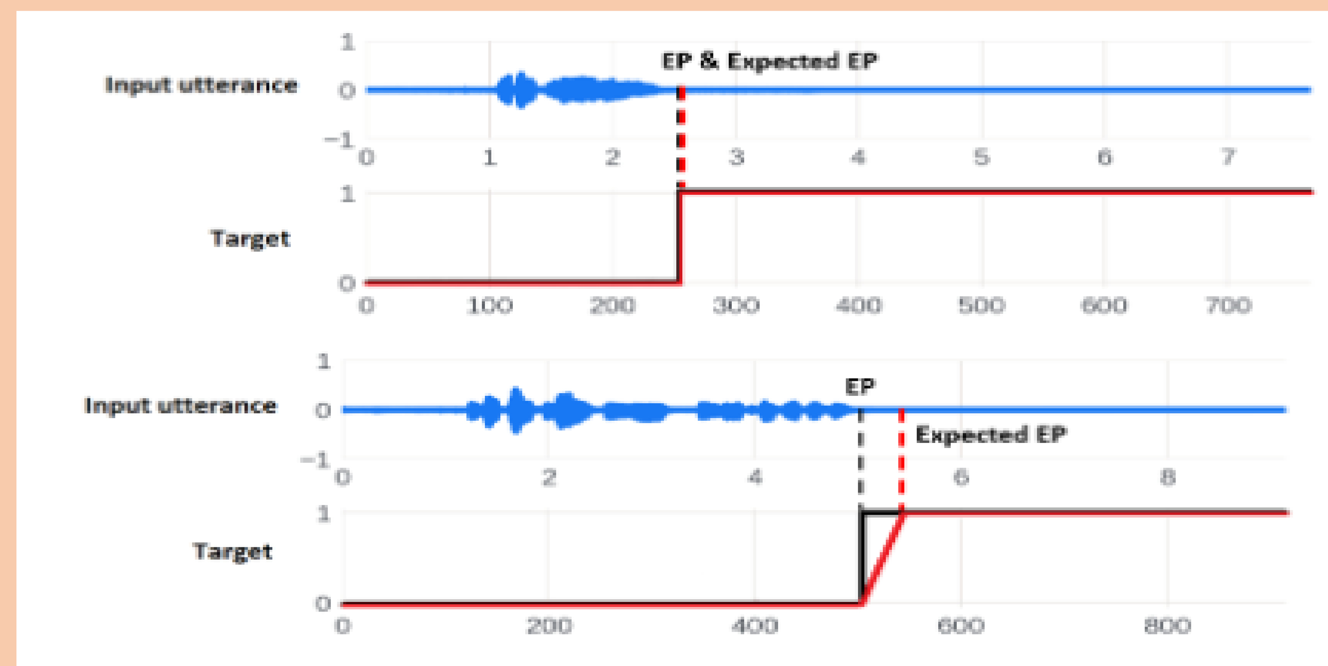
Task Formulation

Problem: On the one hand, the model should be less aggressive for fewer early-cuts; on the other hand, it should be aggressive enough for faster response.

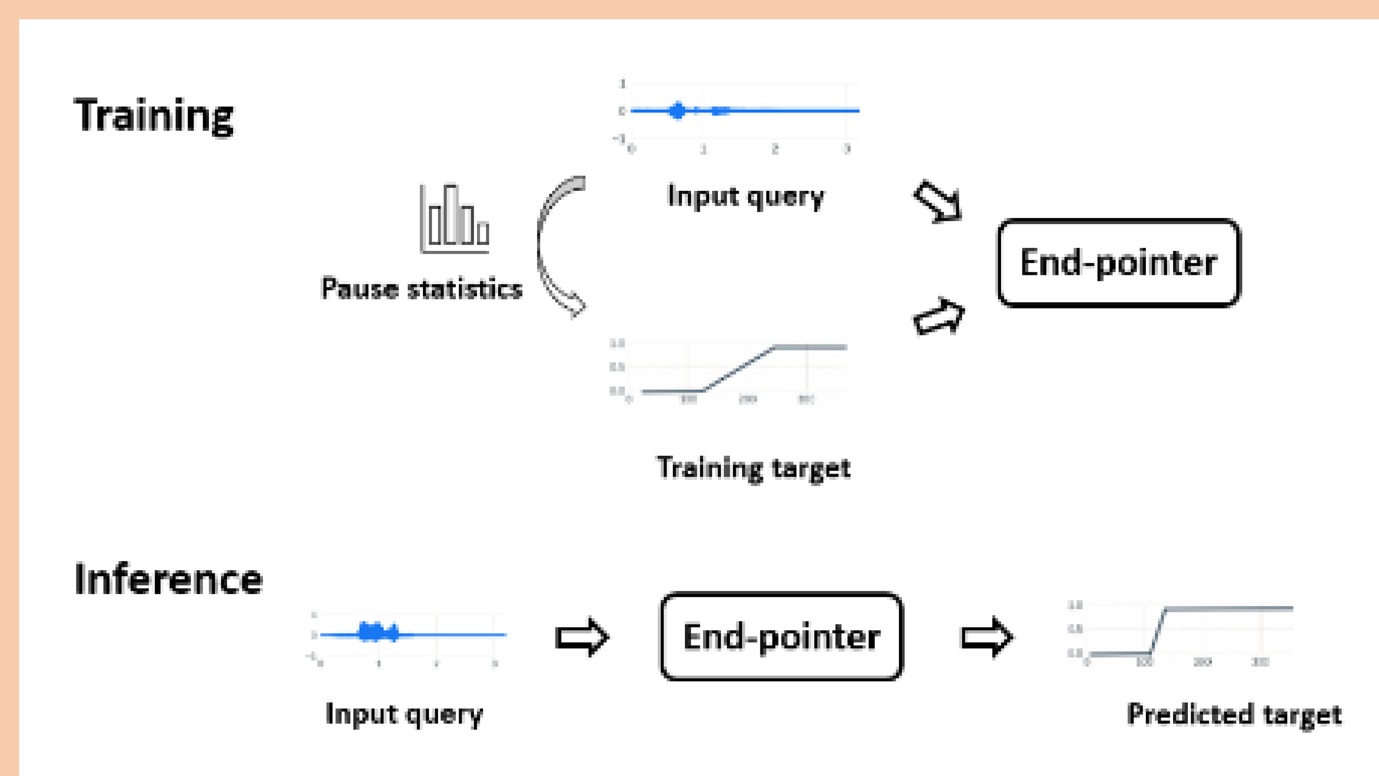
Model: A neural end-pointer with unidirectional LSTMs + linear output layer.

Methodology

We modify the training targets for an end-pointer from hard-coded binary values (target in black) to soft-coded float values (target in red):



Overall setup of our method:



How we adjust the training targets for the input queries:

- We calculate the expected pause duration for each text utterance (T_T), scaled by the user's speaking rate (R):
$$T_S = T_T \times R$$
- The expected pause duration (T_T) can be obtained based on context / prefix of the text utterance (C), and by aggregating the pause statistics of the context in the training set:
$$T_T | C \sim N(\mu, \sigma^2)$$

Experiment

Datasets

- **Smartphone data**: ~ 14M clean utterances from data vendor and live traffic
- **Wearables data**: 440k real-user utterances collected by smart glasses

Metrics

- **Latency**: P50, P75, P90, P99
- **Accuracy**: early-cut rate

Result

Smartphone data

Threshold	Early-cut rate (%)	P50 (ms)	P75 (ms)	P90 (ms)	P99 (ms)
0.50 / 0.63	3.39 / 3.38	160 / 120	180 / 150	230 / 210	480 / 530
0.60 / 0.74	2.45 / 2.38	170 / 120	200 / 160	250 / 240	530 / 600
0.70 / 0.82	1.74 / 1.67	180 / 130	210 / 170	280 / 260	590 / 660

Wearables data

Threshold	Early-cut rate (%)	P50 (ms)	P75 (ms)	P90 (ms)	P99 (ms)
0.56 / 0.50	14.83 / 14.81	420 / 350	590 / 450	860 / 810	1990 / 2500
0.60 / 0.59	12.82 / 12.75	470 / 430	670 / 510	970 / 870	2050 / 2500
0.70 / 0.67	10.53 / 10.37	580 / 500	780 / 650	1100 / 1130	2120 / 2500

- The end-pointer performs better on both datasets by applying our method.
- Our method shows further advantages for tiny models (details in the paper).

TEXAS The University of Texas at Austin Meta AI