Detecting Pause Anomalies in Read Japanese L2 Speech

Stephen McIntosh, Daisuke Saito, Nobuaki Minematsu

The University of Tokyo

Background and Objectives

Pronunciation feedback has been shown to be effective in second-language pronunciation training, but it can be difficult to obtain. Software-based mispronunciation feedback can fill this need.

Here we focus on two types of errors in speakers' pause patterns:

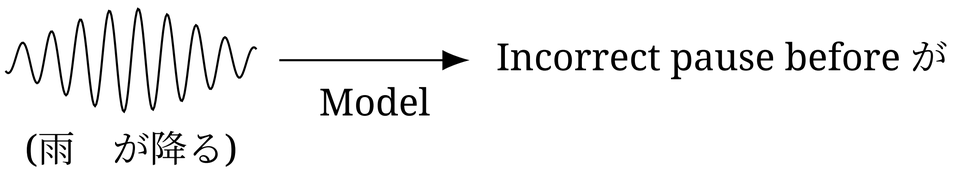
Incorrect Pause

A word boundary where the speaker paused, but L1 (native) speakers would normally not pause (e.g. 雨|が降る)

Missing Pause

A word boundary where the speaker did not pause, but most L1 speakers would pause (e.g. 美しい∧本当に 今日の景色は。)

We'd like to build the "Model" in the following figure:



Method

We use two main components:

- **Forced aligner**: detects pauses in the input
- **Pause model**: predicts where native speakers would pause, outputting a classification PAUSE / NO_PAUSE for every word boundary

We use an off-the-shelf forced aligner, but for the pause model, we tried two approaches:

Full Cascade

We apply ASR to the input utterance and then feed the result to a separate pause model that predicts the word boundaries in the transcription where a native speaker would have paused.

$$\text{AUDIO} \xrightarrow{\text{ASR}} \text{TEXT} \xrightarrow{\text{PM}} \text{PAUSE PATTERN}$$

The pause pattern is a sequence of inferred pause probabilities, one for each word boundary.

Fused Model

We use a single model that transcribes the utterance, interspersing the transcript with pause markers.

$$\text{AUDIO} \xrightarrow{\text{ASR\_PM}} \text{TEXT W/ PAUSE MARKERS}$$

For example, if the input audio contains the phrase つ ま り 彼 は 来 な い, then we should get つまり [PAUSE] 彼は来ない, regardless of the observed pause pattern.

To get a final output, we compare the *predicted* typical pause locations and the *actual* pauses found by the forced aligner:

Pred	うん	|	大学	を	やめて	ね	|	会社	を	作る	らしい	よ
Actual	うん		大学	を	やめて	ね	|	会社	を	| 作る	らしい	よ
	うん	∧	大学	を	やめて	ね	|	会社	を	| 作る	らしい	よ

Experiment

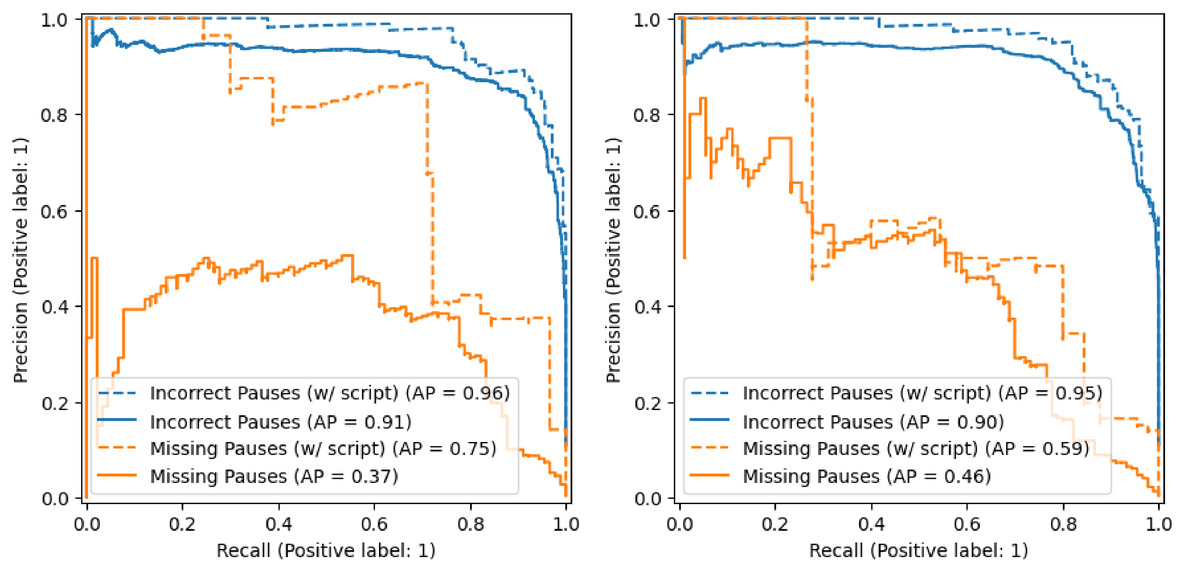
We use reazonspeech-espnet-v2 for ASR and the Montreal Forced Aligner.

For the full-cascade pause model, we fine-tune XLM-RoBERTa based on a subset of the JVS, JSUT, and UME-JRF corpuses.

We build two fused pause model+FA models, trained only on UME-JRF:

- a w2v-BERT 2.0-based model trained with CTC loss
- a Whisper (small)-based model. In this case, we additionally experiment with forcing Whisper to produce the correct transcription when the script is available (denoted *w/ script* in the results).

Results



Precision-recall curves for the **fully-cascaded** model in two training conditions

(*left*: average pause pattern, *right*: one utterance, one sample)

Model Name	Condition	Precision	Recall	F1-Score
Whisper small	Incorrect Pauses	0.64	0.96	**0.77**
	Missing Pauses	0.15	0.50	0.24
	Incorrect Pauses (w/ script)	0.66	0.93	0.78
	Missing Pauses (w/ script)	0.17	0.66	0.27
w2v-BERT 2.0	Incorrect Pauses	0.57	0.96	0.71
	Missing Pauses	0.18	0.62	**0.28**

Results for **pause model+FA** models

To provide a flavor for the kinds of sentences that are difficult, here are some patterns where the text-based pause model struggled:

- **Overindexing on particles**: our models tended to incorrectly predict a pause after an early particle, making mistakes like "それは|たいてい一時間にも及ぶ" or "イランに|天気予報はない". We saw a similar phenomenon with words like もちろん that appeared in the training set often and were usually followed by a pause.
- **Missing semantic pauses**: some words like 何げなく and まるで break up a sentence, but these happened to show up only in our validation set, so the model did not do well on such examples.
- **Lists**: when multiple nouns are listed (e.g. 生活扶助医療扶助住宅扶助など) the model struggles to break them up properly.

Discussion

Since we only need to catch some pausing mistakes, the **full-cascade** model, when operating on the left side of the curve, is **usable in practice**.

The performance of the pause model+FA approach is **less exciting**, but our models were trained on less data. Our evaluation was also a bit unfair to these models, since the outputs had to be aligned with the reference script, where in practice this wouldn't be the case.

Our **evaluation method** has room for improvement. By averaging pause probabilities over multiple utterances, **we assume that pauses** at one word boundary **are independent** of pauses at others. For a simple counter-example, notice that pauses are unlikely to occur in succession. To test whether this is a significant effect, we compared the accuracy of the two training conditions of the fully cascaded models on predicting the pause patterns of held-out native utterances. Both have similar performance, even though only the *one utterance, one sample* model can capture pauses that are not independent of each other.