Leveraging Allophony in Self-Supervised Speech Models for Atypical Pronunciation Assessment

Kwanghee Choi kwanghee@cmu.edu Eunjung Yeo Kalvin Chang
Shinji Watanabe David R. Mortensen

Carnegie Mellon University, Language Technologies Institute

Problem Settings

Atypical (Dysarthric/Nonnative) speech assessment

- Entails pinpointing atypical phonemes (for clinical training, education, etc.).
- Datasets are often low-resource (<10k utterances).
- In our datasets, each utterance x is sliced into N segments: $\{(s_1, p_1), (s_2, p_2), \cdots (s_N, p_N)\}$.

Goodness of Pronunciation (GoP)

- `GoP` aims to measure the level of typicality of a segment given phoneme.
- `GoP` aims to have a high correlation with the level of dysfluency/disfluency.
- Often, pre-existing phoneme classifier is used to define GoP: $\texttt{GoP}_p(s) = \log P(p|s)$.
- Pronunciation score: $\texttt{Pro}(x) = \frac{1}{N}\sum_{i=1}^{N} \texttt{GoP}_p(s)$.

Our method: MixGoP

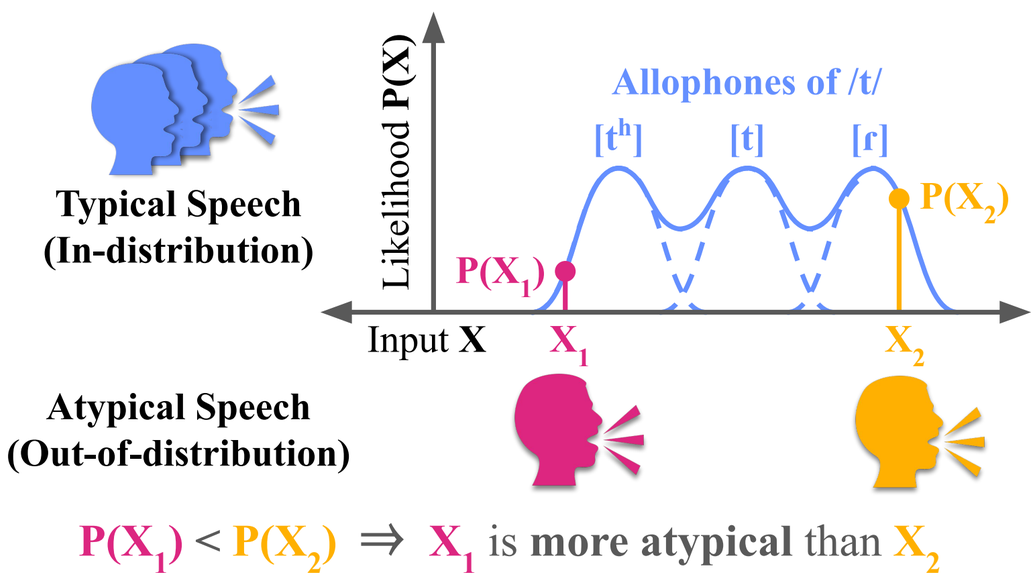


Figure 1. MixGoP models the likelihood of each phoneme using a Gaussian mixture, trained on typical speech (in-distribution), to capture allophonic variations. We then evaluate on atypical speech (out-of-distribution).

Problems of using a Phoneme Classifier for GoP

- Final linear layer enforces a single phoneme to be a *single cluster*, disregarding allophony.
- Using softmax assumes *in-distribution* categories, which is not adequate for out-of-distribution speech.

Gaussian Mixture Models solve both!

- Gaussian mixtures can capture multiple clusters of allophones.
- We can use the GMM likelihood directly, *i.e.*, $\texttt{MixGoP}_p(s) = \log P(s|p)$.

Experimental Results

Settings

- **Features**: Traditional features (Mel spec, MFCCs), Existing SotA (TDNN-F), Frozen self-supervised features (WavLM-Large, XLS-R-300M)
- **Datasets**: Dysarthria (TORGO, UASpeech, SSNCE), Nonnative (speechocean762, L2-ARCTIC)
- **Baselines**: Phoneme classifier-based (Four GoP calculation methods), Phoneme density modeling-based (3 OOD methods and our MixGoP)
- **Evaluation**: Kendall-tau corr. between pronunciation scores and dysfluency/disfluency labels

Results

- MixGoP achieves SotA on 4 out of 5 datasets.
- MixGoP works much better on dysarthria datasets, but phoneme-classification-based works well on nonnative datasets.
- XLS-R performance peaks in the middle, WavLM performance generally increases to the last layer.

Allophony of S3M features

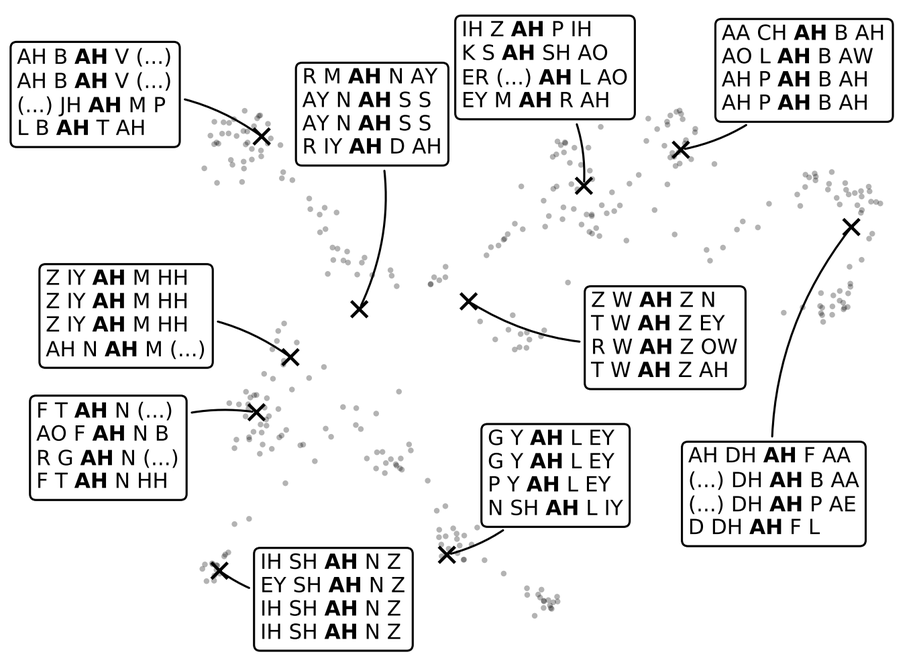
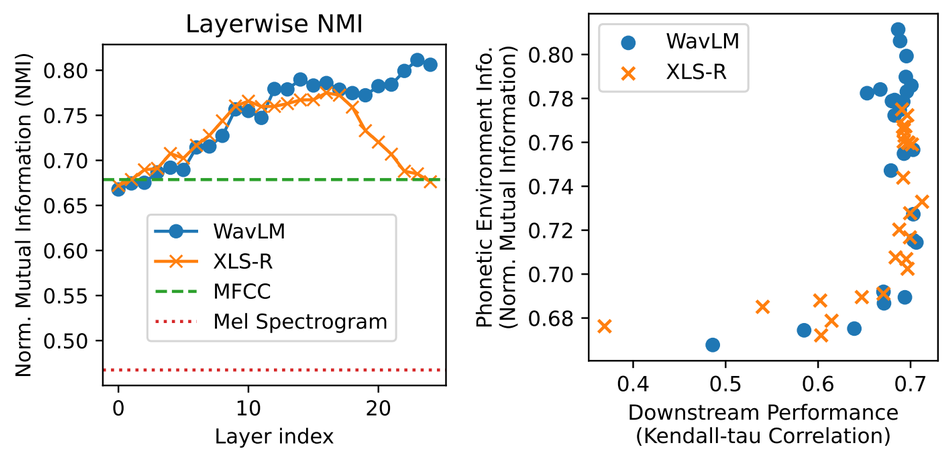


Figure 2. WavLM-Large feature visualization.



(a) Allophony Info. (b) Allophony vs. Downstream

Figure 3. We measure allophony via normalized MI between k-means cluster indices and surrounding acoustic environment.

Submitted to NAACL 2024