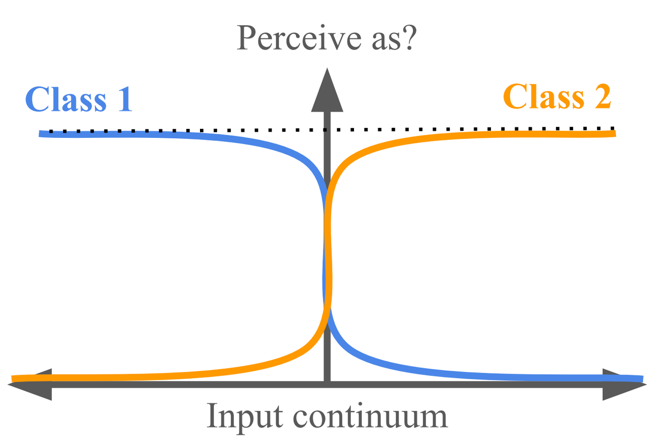
Self-supervised Speech Models Rediscover Phonemes

Kwanghee Choi kwanghee@cmu.edu Eunjung Yeo Kalvin Chang
William Chen Shinji Watanabe David R. Mortensen
Carnegie Mellon University, Language Technologies Institute

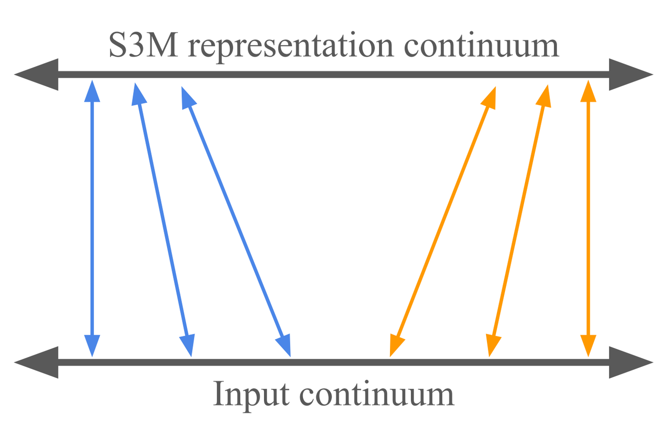
Disclaimer

This paper is in the initial brainstorming stage. We're here to discuss ideas and move this further!

Q1. Do S3Ms perceive sound categorically?



(a) Humans



(b) S3Ms

- Humans perceive stimuli categorically, rather than in a continuous manner.
- Even if it is continuous in the signal domain, category boundary is clear.
- Will S3Ms also have category boundaries? Are those boundaries similar to humans?

Toy experiments using sine signals

- **Idea**: Vowels are dictated by formants F1 and F2 (sometimes F3). In other words, the bare-bone version of vowels are the sum of 3 sine signals.
- **Advantage:** We can easily synthesize an input grid.

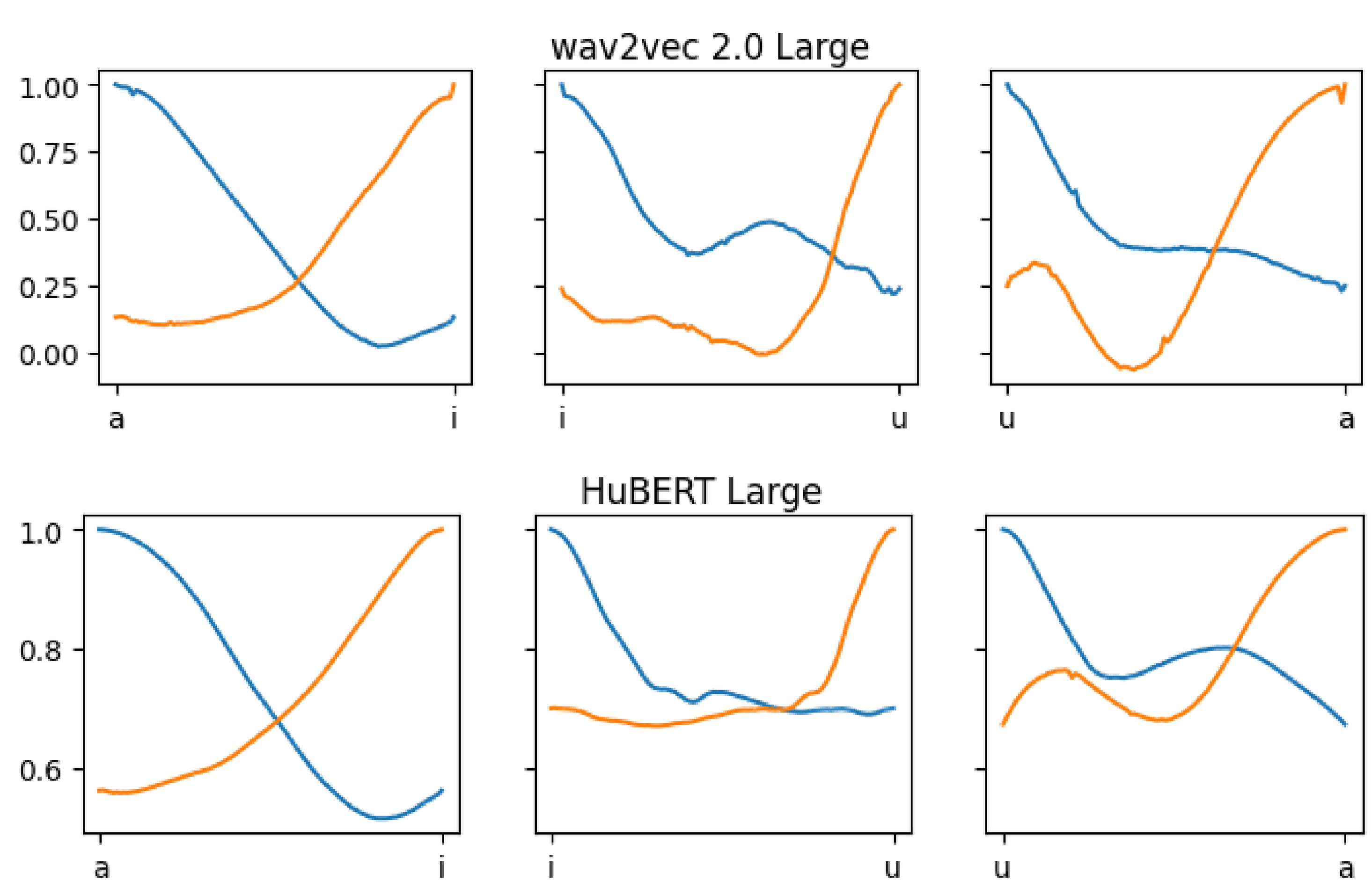
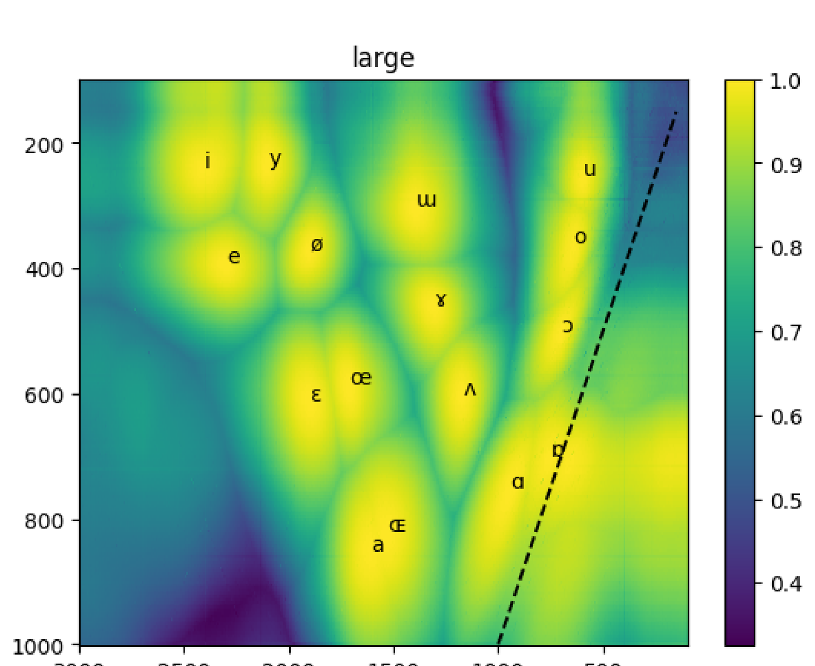
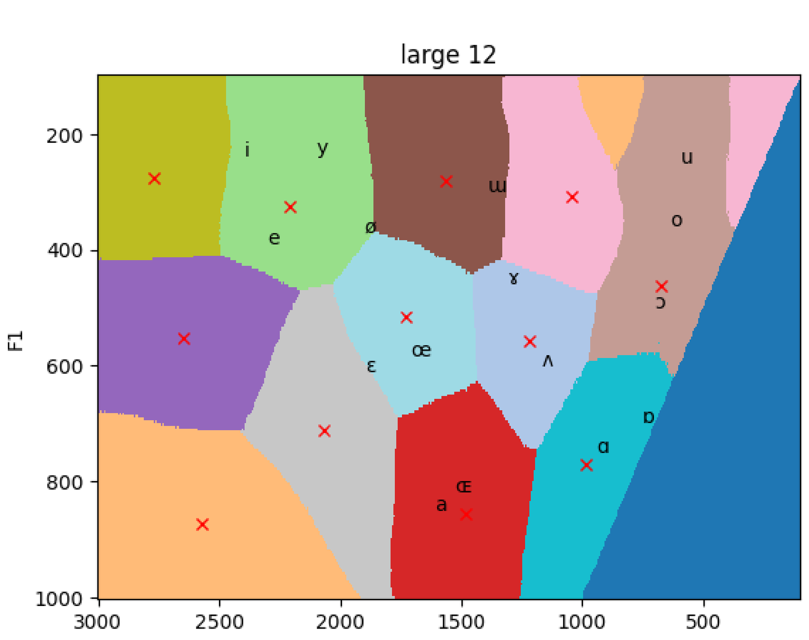


Figure 1. Categorical perception of S3Ms on corner vowels



(a) Similarity Heatmap



(b) K-means result (K=15)

Figure 2. wav2vec 2.0 Large representation on vowel space

Q2. Do S3Ms' similarities correlate with phonological feature distances?

Settings

- **Input signals**: Use the sum of 3 sine signals.
- **Phonological features** (+/-): High, Low, Back
- **Evaluation**: Measure Spearman's corr. between phonological feature distance and S3M representation similarities
- **S3Ms**: wav2vec 2.0-base, large, XLS-R 300m

Results

Phonological feature	base	large	XLS-R
High	0.2652	0.2922	0.3639
Low	0.2136	0.2283	0.1129
Back	0.1456	0.1967	0.2146
High + Low + Back	**0.3523**	**0.3777**	**0.3677**

- Using all the features results in the highest correlation.
- Base and large model focuses less on vowel backness, unlike XLS-R.

Future work

- **More realistic signals**: Sine signals are easy yet less convincing. We are now preparing existing signal continuums.
- **Consonants**: Compared to vowels, consonants are more dynamic. How will place and manner of articulation encoded in S3Ms? Will S3Ms be consistent with perturbation theory?
- **Training dynamics**: Do these characteristics emerge during training, or is it due to inductive bias of neural net architecture?
- **Compositionality**: Do S3Ms encode phonemes in a compositional manner? For example, will the S3M feature of /i/ be the summation of vowel, closeness, frontness, and unroundness feature? Do S3Ms have to see all the languages to handle unseen phonemes?
- **Downstream tasks**: Can we recognize/synthesize phones in a zero-shot manner? Can we conduct data augmentation in the S3M representation domain?
- **Acoustic understanding**: Can we extract fundamental units for non-speech signals?

Work in progress.