Probing articulatory representation learning for phonological distinctions

Sean Foley, Louis Goldstein

University of Southern California

Full poster PDF

Background

- A range of previous studies have probed the representations of self-supervised acoustic models for phonetic and phonological information, suggesting such information is encoded in these models [3, 2, 12, 11]
- Can models trained on articulatory data also derive representations that encode phonological information?

Articulatory representation learning

- A number of studies have applied self-supervised learning approaches to articulatory data, termed *articulatory representation learning* [9, 4, 10]
- No previous study has probed the extent to which these representations may capture meaningful *phonological* distinctions

This study: trains a predictive learning model on roughly one hour of articulatory data from a single speaker collected via real-time MRI and probes the learned representations of this model for crucial phonological distinctions

Method

Model: Contrastive predictive coding (CPC) [13]; convolution-based encoder (1D conv x3) and LSTM-based autoregressive module (LSTM x3)

Analysis:

1. Multinomial logistic regression probes – phoneme classification
2. ABX [15, 5] probes – constriction degree (CD) and constriction location (CL); raw latents vs. k-means codes ($k = 100$)

Dataset: single speaker real-time MRI speech corpus; ∼1 hour of speech; combination of read and spontaneous speech; midsagittal orientation, 99 frames/sec

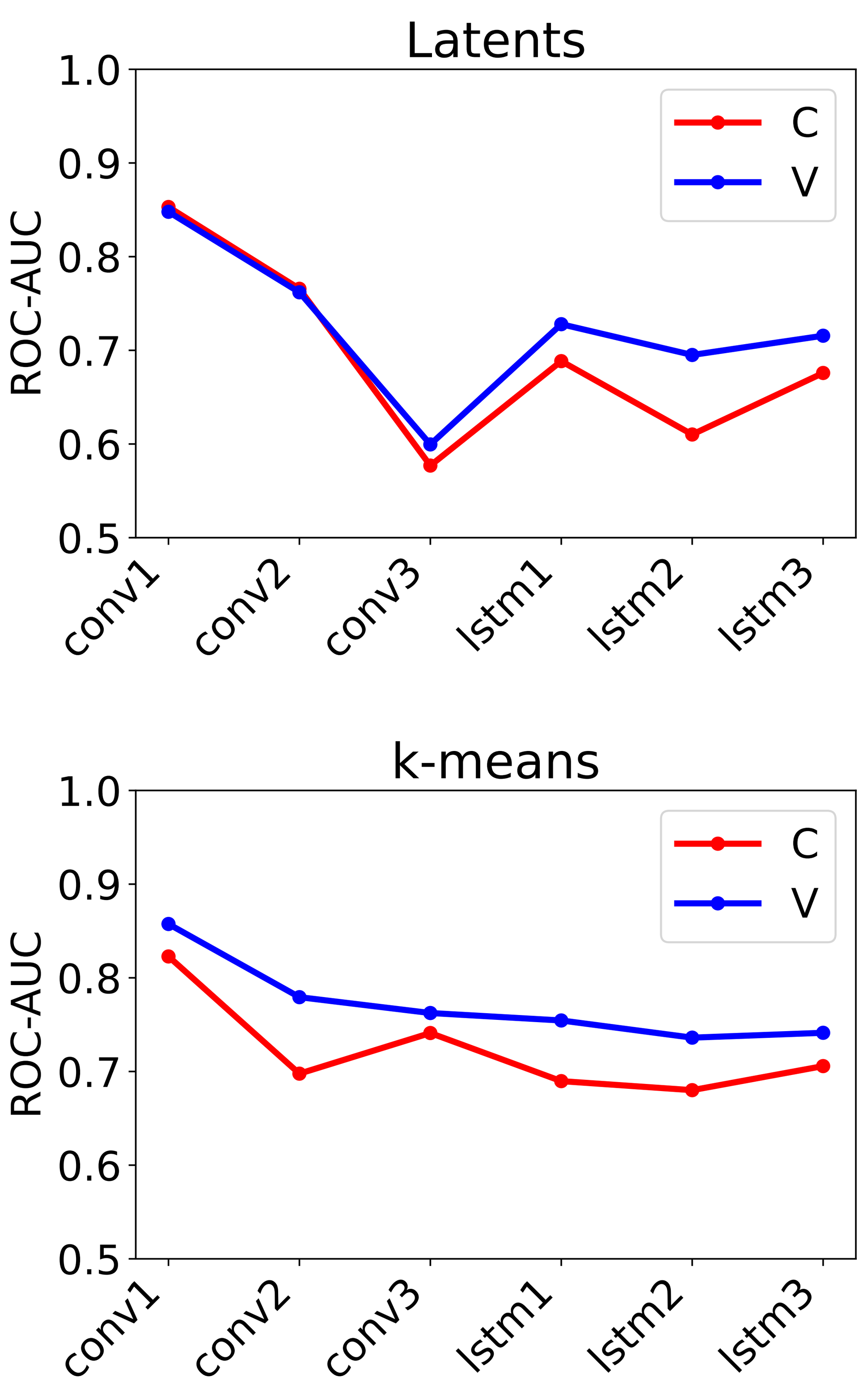
Preprocessing: video chunked up to 1s in length using pretrained VAD [1] and hand-annotated phoneme alignment; all video frames were z-scored and rescaled to 128 × 128

Implementation details:

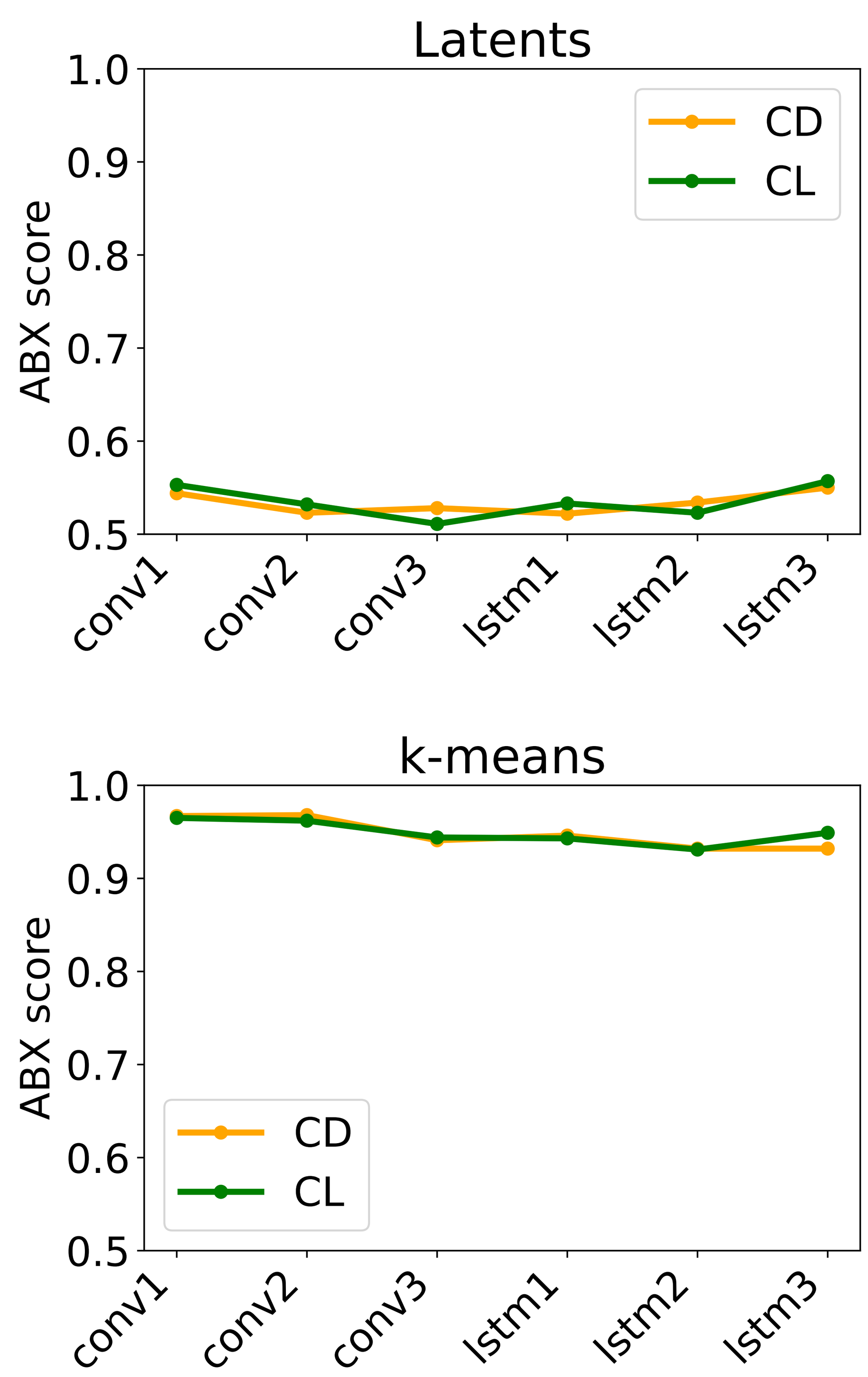
Parameter	Value
Learning rate	1e-3
Batch size	32
Conv dims	(2048, 1024, 521)
Conv kernels	(1, 3, 3)
LSTM dim	512
Prediction horizon	12

Example MRI frame from the data set.

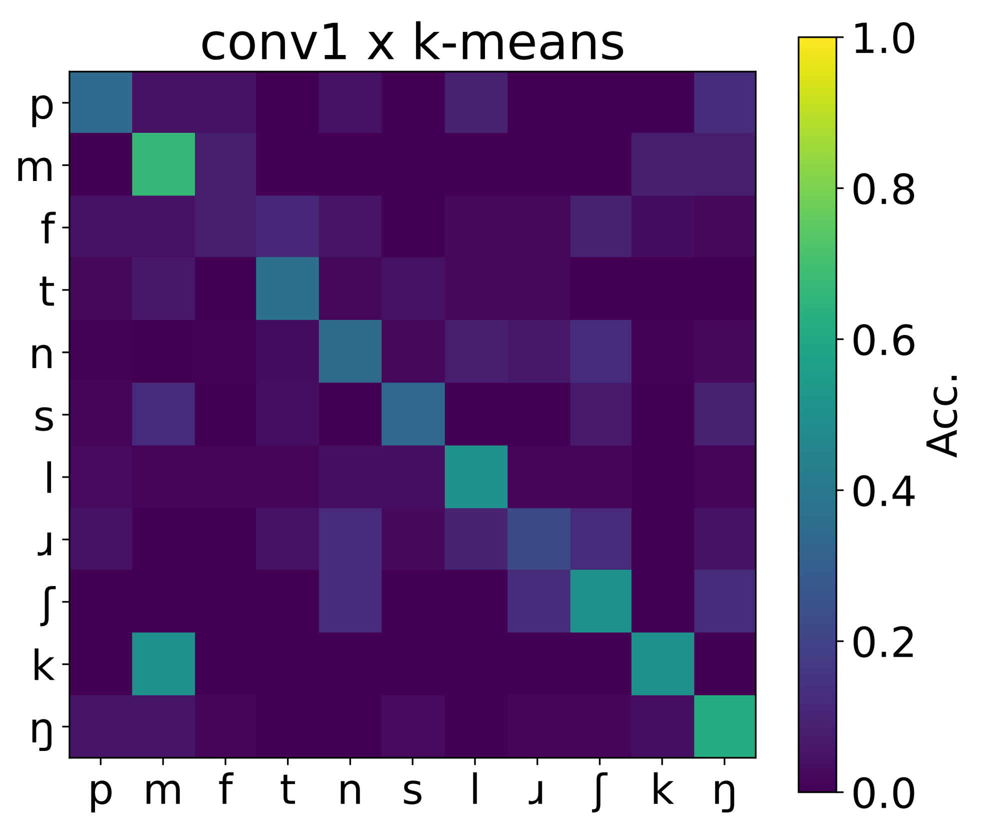
Results

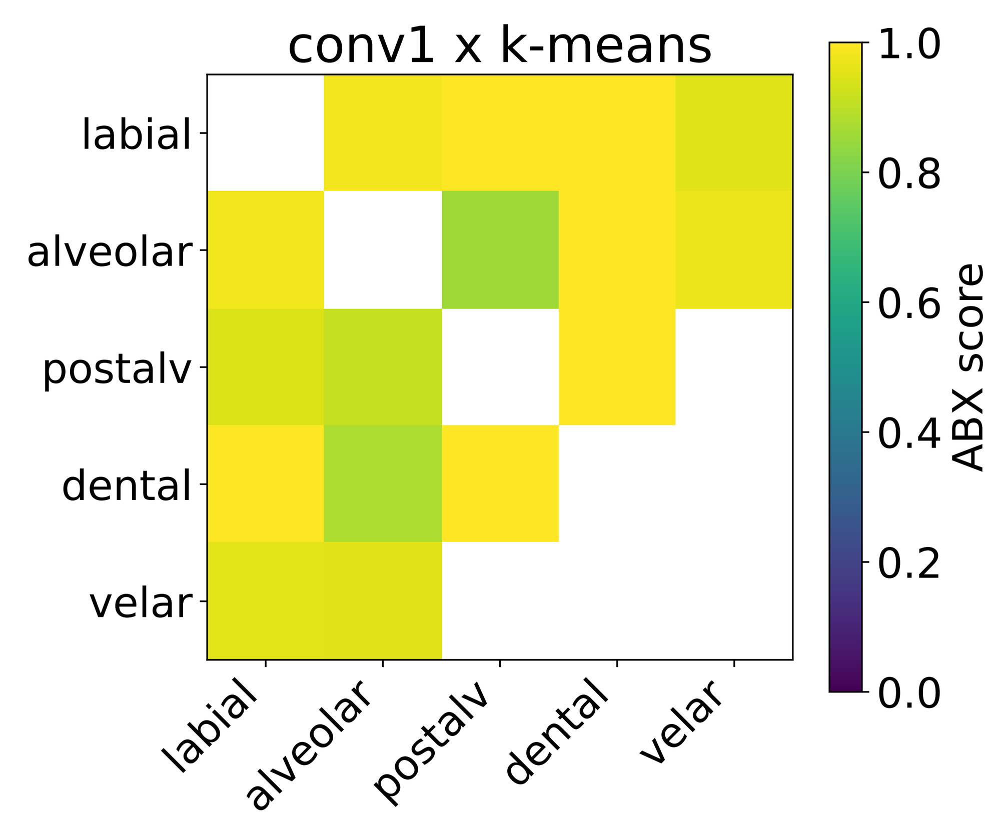


Both the raw latents (top) and k-means codes (bottom) perform well above chance in phoneme classification. The k-means codes show generally better performance across the layers.



While the latents show poor performance in ABX discrimination of CD/CL distinctions (top), the k-means codes score near ceiling across all layers (bottom).





As expected, consonant errors are most common among coronals in phoneme classification (top). Dental, alveolar, and post-alveolar CLs were most often confused in the CL ABX tests (bottom).

Findings

1. While the latents perform decently in phoneme classification, their performance is poor in making CD/CL distinctions
2. Only once the latent space is discretized via k-means do the CD/CL probes perform well
3. Representations from the convolution-based encoder generally outperformed those of the LSTM-based autoregressive module
4. As expected, distinctions among coronal consonants are the least well-separated in the model's representations

Interpretation

1. The raw latents likely encode more phonetic information and are susceptible to contextual effects, while the discretized space has less noise
2. While the discretized CD/CL probes score near ceiling, differences unrelated to local CD/CL may explain this performance, e.g. different lingual postures in /s/ and /t/
3. Confusion between labials and velars may be due to CV coarticulation - the model encodes all dorsal constrictions (C or V) similarly

Future Directions

- Future work could try other forms of predictive learning, e.g. masked prediction
- Given the evidence for the role of sensory prediction in the control of speech production [8, 7, 6, 14], multimodal models may be more insightful
- Extending this work to multi-speaker corpora could allow for more robust representation learning

Acknowledgments

This work was supported by NIH grant T32 DC009975.

References

[1] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 7124–7128. IEEE, 2020.

[2] Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K Anumanchipalli. Evidence of vocal tract articulation in self-supervised learning of speech. *arXiv preprint arXiv:2210.11723*, 2022.

[3] Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. Probing phoneme, language and speaker information in unsupervised speech representations. *arXiv preprint arXiv:2203.16193*, 2022.

[4] Marc-Antoine Georges, Jean-Luc Schwartz, and Thomas Hueber. Self-supervised speech unit discovery from articulatory and acoustic features using vq-vae. *arXiv preprint arXiv:2206.08790*, 2022.

[5] Marc-Antoine Georges, Marvin Lavechin, Jean-Luc Schwartz, and Thomas Hueber. Decode, move and speak! self-supervised learning of speech units, gestures, and sound relationships using vocal imitation. *Computational Linguistics*, 50 (4):1345–1373, 2024.

[6] Frank H Guenther. *Neural control of speech*. Mit Press, 2016.

[7] Gregory Hickok. Computational neuroanatomy of speech production. *Nature reviews neuroscience*, 13(2):135–145, 2012.

[8] John F Houde and Srikantan S Nagarajan. Speech production as state feedback control. *Frontiers in human neuroscience*, 5:82, 2011.

[9] Jiachen Lian, Alan W Black, Louis Goldstein, and Gopala Krishna Anumanchipalli. Deep neural convolutive matrix factorization for articulatory representation decomposition. *arXiv preprint arXiv:2204.00465*, 2022.

[10] Jiachen Lian, Alan W Black, Yijing Lu, Louis Goldstein, Shinji Watanabe, and Gopala K Anumanchipalli. Articulatory representation learning via joint factor analysis and neural matrix factorization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[11] Oli Danyi Liu, Hao Tang, Naomi Feldman, and Sharon Goldwater. A predictive learning model can simulate temporal dynamics and context effects found in neural representations of continuous speech. *arXiv preprint arXiv:2405.08237*, 2024.

[12] Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Levy. Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration. *arXiv preprint arXiv:2306.06232*, 2023.

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[14] Benjamin Parrell, Vikram Ramanarayanan, Srikantan Nagarajan, and John Houde. The facts model of speech motor control: Fusing state estimation and task-based control. *PLoS computational biology*, 15(9):e1007321, 2019.

[15] Thomas Schatz. *ABX-discriminability measures and applications*. PhD thesis, Université Paris 6 (UPMC), 2016.