Speech models' phoneme representations are more phonetic than distributional

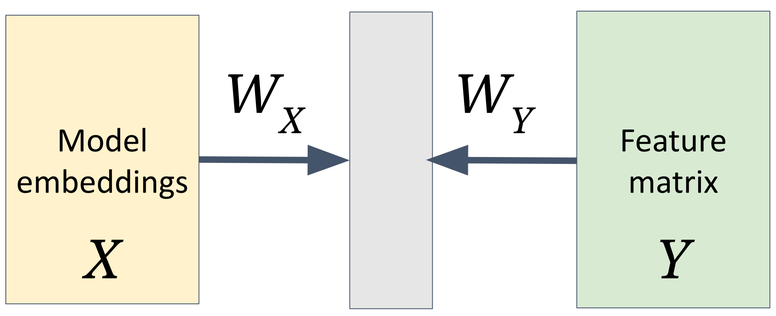
INTRODUCTION

Do self-supervised speech models (S3Ms) represent phonemes like humans do?

Substantive [1,2]	Distributional [3,4]
Speakers use phonetic information in their representations of phonemes	Speakers represent phonemes distributionally, without reference to what they sound like
/ŋ/, /g/, /m/ phonetically similar	/ŋ/, /ʒ/, /v/ distributed similarly

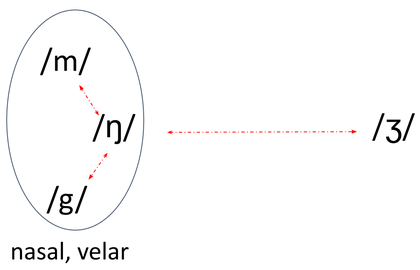
Which type of feature system **better corresponds to models' representations?**

APPROACH

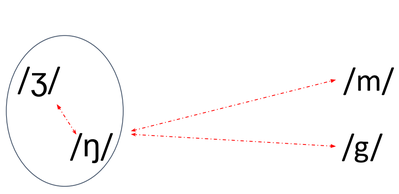


Align S3M **embeddings** of speech sounds with different **feature systems** using Canonical Correlation Analysis (CCA):

- **Substantive feature system [5]** (17 dim):
  → what do phonemes sound like?



- **Distributional feature system** [6] (34 dim):
  → how are phonemes distributed in words?



Data

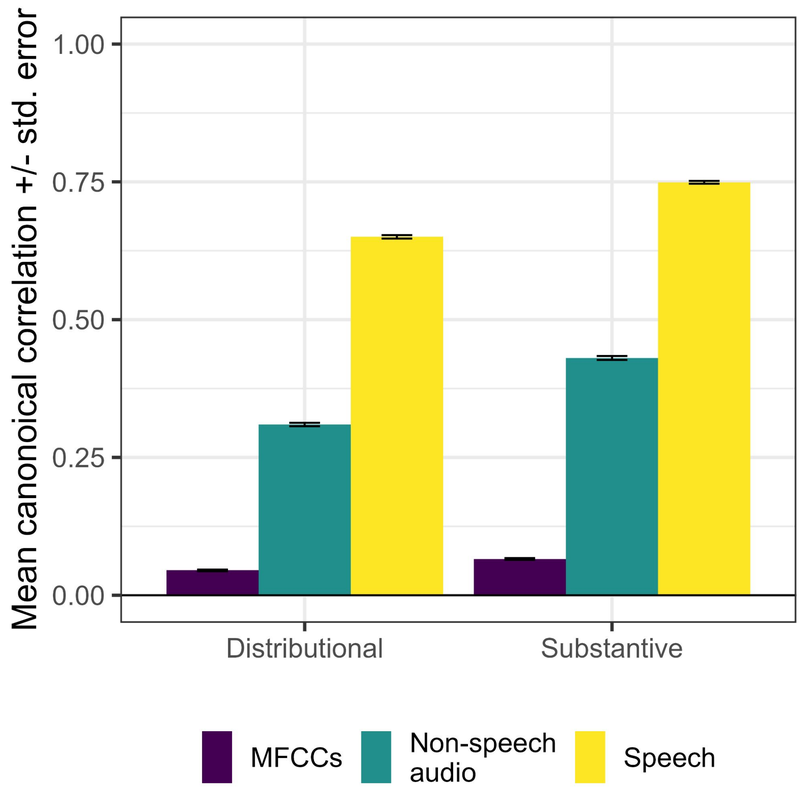
39 English phonemes, extracted from CV and VC sequences of English, synthesized by 10 TTS voices.

MODELS TESTED

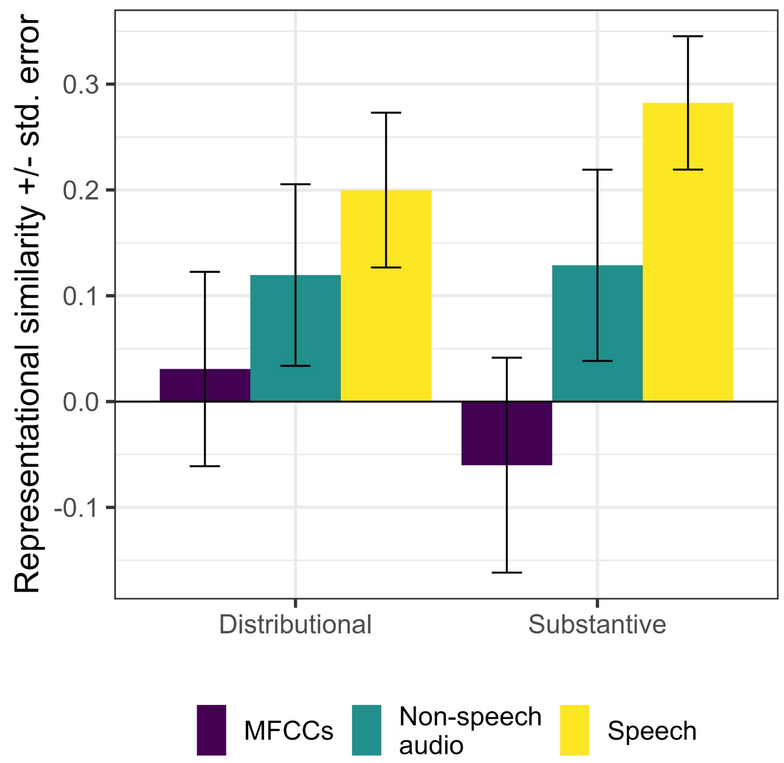
Hubert-Large (trained on English); Hubert-Large (trained on non-speech ambient sounds); Wav2Vec2-Large (not shown)

RESULTS

CCA model-feature alignment:



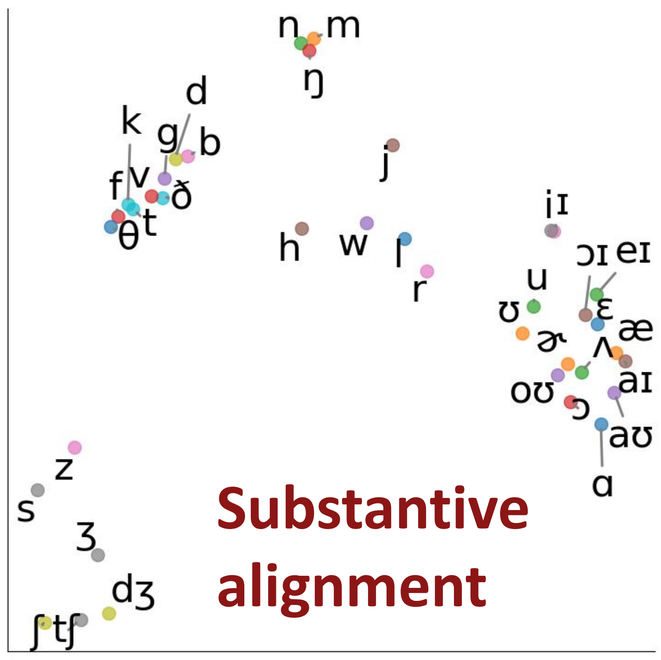
Second-order representational similarity:

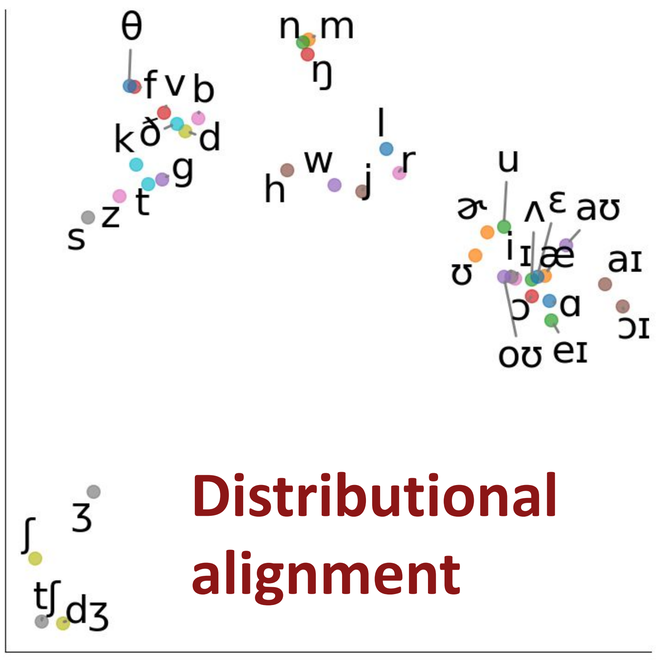


Plots derived from 12th transformer layer.

ERROR ANALYSIS

Model embeddings struggle to capture the [ŋ] ~ [ʒ] relationship in the distributional feature system, even with supervision through CCA.





TAKEAWAYS

S3M's phoneme embeddings are primarily **substantive**: they best encode phonetic (acoustic–articulatory) differences between sounds, while poorly encoding abstract distributional properties.

[1] Chomsky, N., & Halle, M. (1968). The sound pattern of English. **[5]** Hayes, B. P. (2011). *Introductory phonology*. John Wiley & Sons. [6] Mayer, C., & Daland, R. (2020). A method for projecting features from observed sets of phonological classes. *Linguistic Inquiry*, *51*(4), 725-763. **[3]** Mielke, J. (2008). *The emergence of distinctive features*. OUP. **[2]** Pierrehumbert, J. (2000). The phonetic grounding of phonology. *Bulletin de la communication parlée*, *5*, 7-23. **[4]** Silfverberg, M. P. et al., (2018). Sound analogies with phoneme embeddings. *SCiL*.

Canaan Breiss & Jon Gauthier

USC University of Southern California

UCSF University of California San Francisco

Take a picture to **download** the **poster**