Probing Speech Encoders for Phonetic Detail in /s/

Angelique Charles-Davis
Department of Computer Science
angelcd@stanford.edu

Abstract

/s/ frontness or backness is a sociophonetically rich variable that can contribute significantly to speaker identification and speaker demographic classification tasks. However, the extent to which place of articulation in /s/, as measured by center of gravity, is implicitly encoded in self-supervised speech encoders remains unclear. We probe layer representations for wav2vec and HuBERT models of different sizes and with different finetuning strategies in order to identify which models provide the most informative and interpretable representations of center of gravity in /s/, and find that this subphonemic information peaks in early transformer layers. We also conclude that model size and self-training versus regular supervised finetuning are the strongest mediators of representation robustness, and that speaker identification fine-tuning does not lead to enhanced detail in /s/.

1 Introduction

In pursuing more inclusive and personalized technologies, demographic feature extraction becomes of particular relevance in speech encoding while bringing forward important ethical questions. Modern encoders have achieved staggering levels of accuracy in classification tasks, able to predict age, gender, country of origin, race, and education level by leveraging differences in fundamental frequency, vowel height, amplitude perturbation, and other tonal, lexical or syntactic qualities (Yang et al., 2025).

The place of articulation for a given production of /s/ can be measured by the center of gravity of the resulting spectral moment. Frontness in /s/ is a sociophonetically rich and complex variable. Calder & King found that in addition to well-documented gender patterning in productions of /s/, in which women produce on average a much fronter /s/ than men, black men in Bakersfield produce a fronted /s/ that correlates more closely with that of white women, with an increase in frontness in apparent time (Calder and King, 2022). A fronted /s/ can also index sexual orientation, town vs. country orientation, and certain street styles or local variants (Mack and Munson, 2012; Pharao et al., 2014). In the TIMIT data used in this study, center of gravity in /s/ is a statistically significant predictor of sex across all participants and of race in male participants.

Given that models have total access to the entirety of an acoustic signal, those that are sensitive to subphonemic information may be implicitly aware of these social categories at a level that matches or surpasses human perception. Coupled with the predictive strength of /s/, this intuition suggests that without additional finetuning, encoded demographic information may be rich even in samples in which voicing is not present.

We can further partition self-supervised model based on training pretext. Wav2vec style models predict a masked target that comes from the quantized local features and use contrastive loss, while HuBERT style models predict cluster assignments for features coming from intermediate transformer layers, and use cross-entropy loss (Baevski et al., 2020; Hsu et al., 2021). In this work, we probe wav2vec and HuBERT configurations for center of gravity in /s/, which can help reveal learning differences between discrete and quantized feature learning.

2 Related Work

Previous work has sought to explain the mechanisms underlying the encoding of low-level and high-level linguistic information across model layers in self-supervised models.

Pasad et.al found that wav2vec models belong to a class of autoencoder-style models, which generate representations that bear the highest similarity to their input features in the initial and final layers, with lower similarity in intermediate layers.

Figure 1: COG Distribution

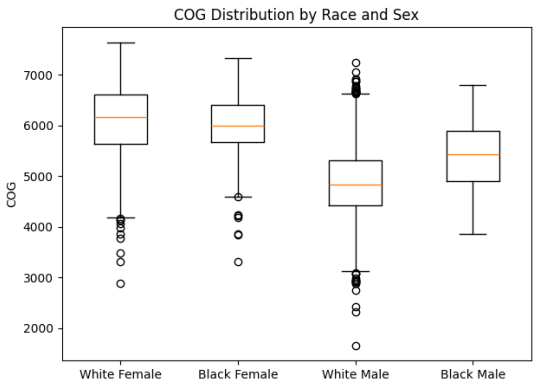

Figure 2: Demographic Distribution

HuBERT models retain less of this input information over time. Comparing model features and spectrogram features through canonical correlation analysis (CCA) reveals the highest correlation in late CNN layers and early transformer layers. For phonetic and word-level information, autoencoder-style models peak in intermediate transformer layers before dropping in later layers, while HuBERT-style models concentrate phonetic and word-level information in higher layers. Pasad et al. also found that CCA was a strong predictor of downstream task performance, meaning that while it is not a measure of mutual information, it can provide valuable insight into what speech encoders know and represent (Pasad et al., 2023).

Martin et al.'s findings in HuBERT corroborate that much of the work of phonemic discrimination work is concentrated within the CNN layers. They also found that phonetic distinctions were encoded early in HuBERT transformer layers, as well as in randomly initialized and non-speech models, suggesting that the transformer architecture itself, combined with model pretraining allows HuBERT to surpass the representational capacity of the log-mel baseline. In the case of aspiration, the model was able to discriminate between members of the same phonemic class (aspirated and non-aspirated /p/) as well as members of the same phonetic class, (neutralized /p/ and /b/) illustrating HuBERT's ability to draw high-level phonological distinctions without losing lower-level phonetic information, a phenomenon which appears early in the processing system (Martin et al., 2023).

Abdullah et al. take an information theoretic approach to explaining wav2vec's discretized features. By modeling phonemes as distributions over individual phones, they find that the distribution-entropy of the computed wav2vec features aligns with these larger phonemic categories. This shows that wav2vec features reliably capture sub-phonemic detail, which can encompass place of articulation, and their findings reveal that /s/ and /z/ are the most closely distributed in feature space, mirroring their phonetic adjacency as voiceless and voiced fricatives(Abdullah et al., 2023).

This work presents a compelling account of where and how phonetic and sub-phonemic work happens in self-supervised models and attests to the fact that these low-level features can be enhanced and augmented throughout the encoding process, surpassing the informativeness of the model input. We can therefore apply similar probing techniques to understand the perception of subphonemic information in /s/, which is not critical to overall model performance. Furthermore, because fine-tuned models implicitly achieve phonetic normalization by suppressing task-irrelevant information (Wang et al., 2025), investigating to what extent models fine-tuned for speaker identification models suppress information about within-/s/ variation can tell us more about its utility in this task.

3 Approach

For this study, we used the TIMIT Acoustic-Phonetic Continuous Speech Corpus, which contains 6300 sentences spoken by 630 speakers that are aligned phonetically and on the word level. To obtain the target center of gravity labels, we isolated /s/ segments only and generated spectrograms and their corresponding time-averaged spectra, from which the center of gravity was computed. The data was split along the original TIMIT lines, resulting in a training set of 7475 tokens and a test set of 2639 tokens.

In order to understand the role of phonological context on /s/ encoding, we took two probing approaches. In the first, indvidual /s/ sequences were passed into the relevant model as raw audio vectors sampled at 16kHz, and the resulting layer outputs

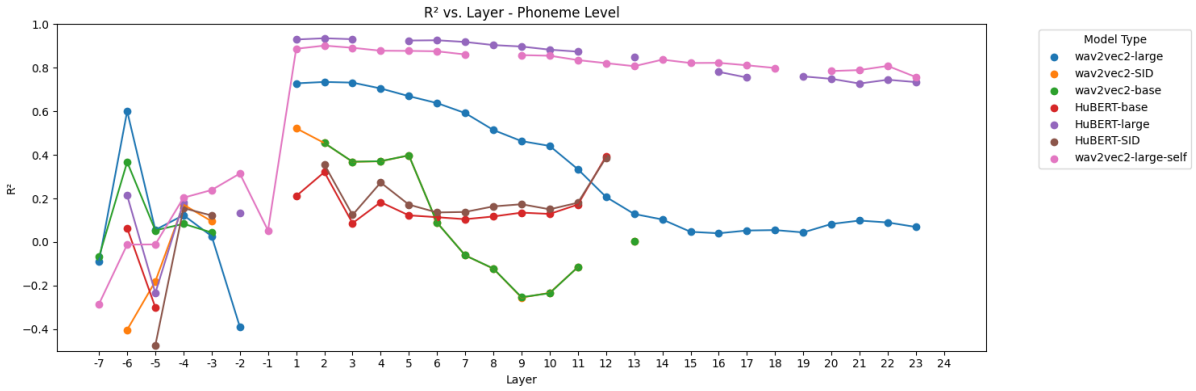

Figure 3: Cross-model Performance - Phoneme-level Encoding

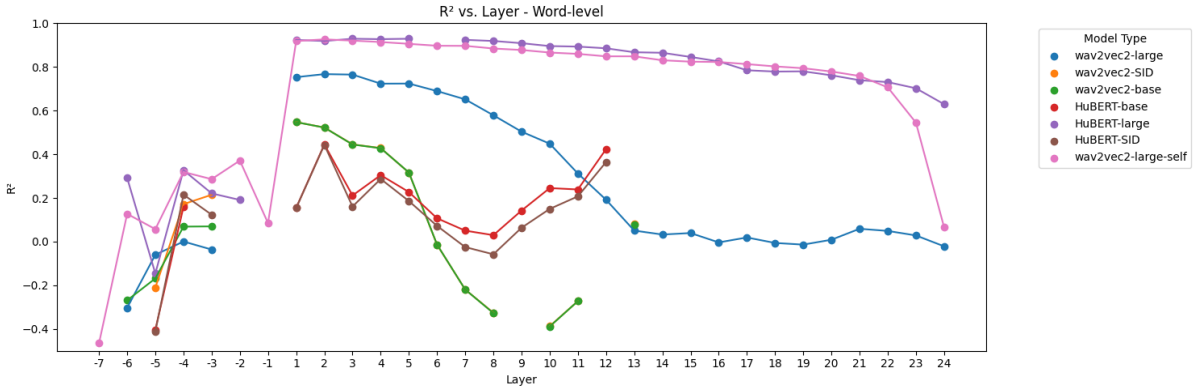

Figure 4: Cross-model Performance - Word-level Encoding

were extracted and mean-pooled over time. In the second approach, word-level encodings, we passed entire words into the model, extracted the time-aligned /s/ segment of the encoding for each layer, and mean-pooled this over time.

For each model-layer pair, a single linear layer regression model was trained on the extracted features and subsequently evaluated on the test set. We take the R^2 value of this evaluation to be the measure of feature performance for that model-layer pair. The base, speaker identification fine-tuned, and large models explored in this analysis are, respectively: `wav2vec-base`, `wav2vec2-base-superb-sid`, `wav2vec2-large-ls-960h` and `wav2vec2-large-960h-lv60-self` , which was fine-tuned on a small amount of labeled data along with a larger amount of self-labeled, on the wav2vec side, and `hubert-base-ls960`, `hubert-base-superb-sid`, `hubert-large-ls-960-ft` on the HuBERT side. For each model, we also trained control probes, which shuffled the features prior to evaluation. Taking the average of these randomized probes yielded no results above 0.2, indicating that any model performance above this threshold can be attributed to the information present in the embedding, rather than the intrinsic strength of the regression model.

4 Experiments

Figures 3 and 4 show the regression results for each of the model-layer pairs, with absent points representing values below the -0.4 R^2 threshold. Each model has 7 convolutional layers. Base models have an additional 13 layers, while large models have an additional 24. We see that none of the models achieve their peak performance within the convolutional layers, with most peaking in early transformer layers before declining in accuracy. Convolution layer representations are stronger in phoneme-level encodings, likely due to their proximity to this more localized and temporally precise input.

In both phoneme level and word-level encoding contexts, model size is the greatest discriminator of encoding strength, with HuBERT-large demonstrat-

ing the strongest R^2 peak, followed by self-trained wav2vec-large and finetuned wav2vec-large. This indicates that models with more parameters encode correspondingly richer phonetic representations. In the case of Hubert-large and wav2vec-large-self, this informativeness remains relatively stable across model layers, with faster degrading in the word-level encodings, suggesting that at the word-level, phonemic detail is more likely to be sacrificed for higher-order representations in later layers. Informativeness in the fine-tuned wav2vec-large model, on the other hand, is steadily degraded after the first transformer layer. The difference in the two large wav2vec configurations demonstrates the role of supervised finetuning in promoting implicit phonetic normalization and by extension /s/ variation suppression, and that this may be mitigated by self-training and the corresponding increase in overall data as well as the noisier data quality. Pseudo-labeled training data can reinforce previously learned features while also preventing overfitting and subsequent phonetic normalization. A small probing experiment of `openai/whisper-small` similarly revealed incredibly low-levels of learning in the supervised model, due to the coarser learning objectives and lack of contextual learner 5.

SID-finetuned models demonstrate almost identical performance to their non-finetuned base counterparts in the transformer layers, particularly in wav2vec. In HuBERT models, SID fine-tuned informativeness is slightly stronger in phoneme-level encodings and slightly weaker in word-level encodings, which supports the account of an enhanced process of phonetic normalization taking place in finetuned models.

Conclusion

Our analysis reveals patterns that both align with and complicate previous work on subphonemic information encoding. When present, /s/ information clusters in early transformer layers, but is notably lacking in convolution layers, despite the expectation that these will align most closely with spectrogram features (Pasad et al., 2023). Furthermore, we find that increased model size is primarily what allows for /s/ information to be richly encoded, with self-training in wav2vec helping to reduce the effects of phonetic normalization that take place in supervised fine-tuning. HuBERT-large performs the best, with near-perfect prediction in its earliest transformer layers. Future work may examine the extent to which increasing or decreasing the amount of fine-tuning performed on the HuBERT-large impacts this performance. Overarching our analysis is the finding that base models do not tend to encode /s/ variation very strongly and that this is not significantly altered by finetuning for a speaker identification task. While the predictive power of /s/ varies greatly between target demographics and speaker populations (Calder and King, 2022; Pharao et al., 2014), future work can examine whether enhancing the robustness of /s/ representations can benefit speaker-sensitive models in cases where voicing and tonal cues may be absent or unreliable, such as in whispered speech, noisy or low-quality audio, various speech pathologies or situations in which overt voicing cues may be consciously suppressed by the speaker. In models with powerful representations, the ethical implications of implicitly encoding such a perceptively subtle linguistic variable should be considered. Overall, our findings highlight the importance of incorporating sociophonetic and sociolinguistic scholarship into interpretability research in speech and language.

References

Badr M. Abdullah, Mohammed Maqsood Shaik, Bernd Möbius, and Dietrich Klakow. 2023. An information-theoretic analysis of self-supervised discrete representations of speech. In *Interspeech 2023*, pages 2883–2887.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

J. Calder and Sharese King. 2022. Whose gendered voices matter?: Race and gender in the articulation of /s/ in bakersfield, california. *Journal of Sociolinguistics*, 26(5):604–623.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Sara Mack and Benjamin Munson. 2012. The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype. *Journal of Phonetics*, 40(1):198–212.

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

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Nicolai Pharao, Marie Maegaard, Janus Spindler Møller, and Tore Kristiansen. 2014. Indexical meanings of [s+] among copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society*, 43(1):1–31.

Yiming Wang, Yi Yang, and Jiahong Yuan. 2025. Normalization through fine-tuning: Understanding wav2vec 2.0 embeddings for phonetic analysis.

Yuchen Yang, Thomas Thebaud, and Najim Dehak. 2025. Demographic attributes prediction from speech using wavlm embeddings.

A Appendix

Full project code and files are here.

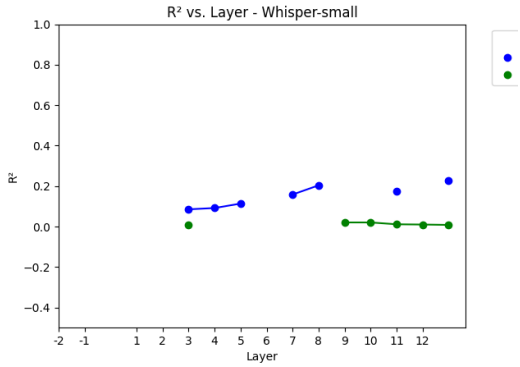

Figure 5: Whisper Layer Performance