Identifying and typifying demographic unfairness in phoneme-level embeddings of self-supervised speech recognition models

Felix Herron[1,2] and **Solange Rossato**[2] and **Alexandre Allauzen**[1] and **François Portet**[2]
[1]MILES Team, LAMSADE, Université Paris Dauphine-PSL (1)
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
felix.herron@univ-grenoble-alpes.fr

Abstract

Modern automatic speech recognition (ASR) systems have been observed to function better for certain speaker groups (SGs) than others in myriad contexts. One potential impediment to progress towards fairer ASR is a more nuanced understanding of the types of modelling error that speech encoder models make, and in particular the difference between the structure of embeddings for high-performance and low-performance SGs. This paper proposes a framework for explicitly modelling two types of geometric error that can occur in modelling phonemes in ASR systems: random error/high variance in phoneme embedding, vs systematic error/embedding bias. We find that training phoneme classification probes only on a single, typically disadvantaged SG, sometimes improves performance for that SG, which is evidence for the existence of SG-level bias in phoneme embeddings. On the other hand, we find that speakers and SGs with higher levels of phoneme variance are the same as those with worse phoneme prediction accuracy. We conclude that both types of error are present in phoneme embeddings and both are candidate causes for SG-level unfairness in ASR, though random error is a bigger problem than systematic error. Furthermore, we find that finetuning encoder models using a fairness-enhancing algorithm (domain enhancing and adversarial training) has impact neither on in-domain probe training benefit nor on measured levels of random embedding error.

1 Introduction

Automatic Speech Recognition (ASR) systems have been repeatedly observed to perform better for certain speaker groups (SGs) than for others (Vipperla et al., 2008; Aman et al., 2013; Feng et al., 2024; Sekkat et al., 2024; Veliche et al., 2024; Attanasio et al., 2024). Inasmuch there has been ample research in trying to improve ASR fairness (Das et al., 2021; Maison and Estève, 2023; Alonzo-Canul et al., 2025; Rai et al., 2025); however, resulting fairer models leave ample room for improvement. One potential drawback to existing fairness enhancing studies is that they tend to treat ASR systems like black boxes - it is challenging to design a precise intervention to reduce demographic unfairness without better understanding the mechanisms involved within the inner workings of the model. This paper provides a framework for understanding two sources of modelling error in tems with respect to unfair SG-level treatment (see Figure 1): embedding bias/systematic error, in which phonemes are modeled around different modes depending on SG; vs unequal variance/random error, in which phonemes are modeled around the same modes but with more noise for one SG than another.

We perform probing experiments on intermediate representations of deep attention-based speech encoder models, in particular on self-supervised speech processing models (S3Ms), to evaluate the bias and variance of phoneme embeddings across SGs. First, we evaluate whether phoneme recognition (PR) probes trained on a single SG perform better on unseen speakers from that same SG and worse for unseen speakers from unseen SGs. First, we find that the same SGs which achieve better ASR performance also achieve better PR accuracy on probes trained on balanced data. For probes trained only on a single SG, we find statistically significant improvement for in-domain training phoneme probes for some but not all SGs, at some but not all layers; however, certain phonemes benefit much more than others. Second, we directly calculate the variance in phoneme embeddings using a k-nearest neighbors (KNN) distance heuristic, a proxy for phoneme embedding variance, which we find correlates strongly for every model with phoneme prediction accuracy.

Finally, we investigate the difference between how phonemes are embedded for different SGs when we finetune for ASR using the fairness en-

Findings of the Association for Computational Linguistics: ACL 2026, pages 33603–33620

hancing algorithm Domain Enhancing/Adversarial Training (DET/DAT). We find that this changes nothing about the relative KNN distance between phonemes, nor does it change the benefit achieved from training PR probes using in-domain data. Thus, we conclude that to the extent that KNN distance is a useful indicator of precise phoneme modelling, that DET/DAT, at least as constituted in our study, is not useful at rendering the phoneme embedding space fairer.

2 Related work in interpretability of speech processing systems

Pasad et al. (2022, 2023) show that certain types of information, such as speaker identity or word meaning, contained within speech are maximized at different model layers of S3Ms. Masson and Carson-berndsen (2023) find that artificially generated non-native speech is represented in a similar manner to real non-native speech in tems. This suggests that tems model non-native speech in a systematically different and reproducible manner, and anecdotally supports the systematic error/embedding bias hypothesis (see Figure 1). Feng et al. (2024, 2021) find that many of the same phonemes tend to be the hardest to understand in tems regardless of SG or ASR architecture. This would suggest that certain phonemes are harder to recognize overall, and that those same phonemes are simply modeled in a less recognizable manner by any tem, providing anecdotal support of the random error/variance hypothesis.

Liu et al. (2023); Mohamed et al. (2024) show that phoneme information and speaker information are modeled in orthogonal subspaces of S3M embeddings. This would suggest that lack of demographic parity in ASR is unlikely due to speaker-specific phoneme modelling differences. Herron et al. (2026a) shows that layers of pretrained self-supervised speech processing models (S3Ms) that are the best capable of modelling ASR are also the least fair, and that this effect is not attenuated after ASR finetuning. Furthermore, they show that even S3Ms pretrained on non-English corpora display the same biases as S3Ms pretrained only on English. These findings motivate the notion that certain speakers are inherently harder to model, regardless of training data and model architecture. Finally, Herron et al. (2025) shows that pretrained S3Ms model certain SGs in a differentiable manner on an utterance scale, though this decreases considerably for most SGs when finetuned for ASR; Prasad and Jyothi (2020) shows similar results for accent encoding.

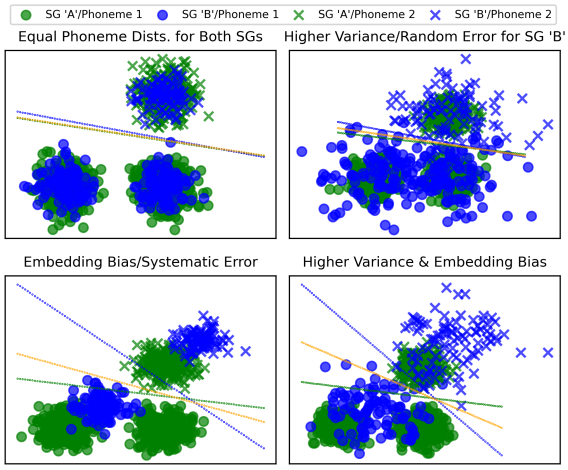

Figure 1: Toy visualization of high variance vs embedding bias in by-SG modelling of phonemes, along with corresponding linear separators trained on the SG of the corresponding cover (orange trained on both SGs). If, compared to an equal-error world (upper left), the S3M embeds phonemes from SG 'B' with higher variance than SG 'A' (upper right), the learned ASR head won't necessarily learn a different transformation (thought it might due to the training data being less noisy) but it will return higher error rate for SG 'B'. However, if phoneme 1 is modeled about different modes depending on SG, representing embedding bias (lower left), this will bias the PR head towards the more heavily represented SG in the training set.

Other studies have delved into ASR errors due to hallucination in decoding (Lin et al., 2025; Frieske and Shi, 2024; Barański et al., 2025; Koenecke et al., 2024). Our study doesn't touch on this - we consider only speech encoders, and perform any decoding using a single linear layer to map to phoneme space. Furthermore, we simplify the task by avoiding the challenges of sequence-to-sequence mapping, instead focusing on single isolated phoneme embeddings. However, it is important to note that the two types of encoding error which we study are not the only potential sources of error in ASR.

3 Motivation

The objective of this paper is to examine the differences between how S3Ms model phonemes of various SGs. In an ideal world for ASR, each phoneme embedding should unambiguously be represented by a vector that can be linearly mapped to its corresponding phoneme label, regardless of speaker or

SG. However, sometimes phoneme embeddings are misclassified, which we hypothesize is due to one of two effects, based on a classical error dichotomy in classification error of random vs systematic error (Taylor, 2022), as illustrated in Figure 1 on toy data. **1) random error**, in which a phoneme is embedded about a correct mode in embedding space, but with sufficient noise as to result in a misclassification in phoneme prediction; or **2) systematic error**, in which a phoneme is embedded about a different mode than the linear classifier has been trained on, due to particularities of a SG to which the speaker belongs. Or **3) both**: in all likelihood, a mixture of both types of errors. We seek to measure whether one error type better explains unfairness for some SGs than for others.

4 Experimental setup

4.1 Data preparation

We use the Sonos Voice Control Bias Assessment Dataset for our analyses (Sekkat et al., 2024). Its recordings are clean and of similar fidelity across speakers, thus reducing bias deriving from varied access to high-quality recording equipment and environments. Furthermore, Sonos provides SG-level metadata for speakers' gender, age, dialect, and ethnicity.

We start data preparation by aggregating some of the SG categories to avoid accidentally biasing our results due to imbalance in dataset construction (Herron et al., 2026b). To wit, we aggregate all native speakers of different American dialects (of which there are 6 in Sonos) into a "Native" category; we do the same for middle-aged adults (of which there are several arbitrarily delineated categories). Our aggregations are based on 1) a lack of expected modelling difference based on theoretical linguistic understanding of how voices age (Rojas et al., 2020), and 2) empirical results of previous studies which have analyzed this dataset and found little difference between adult age groups or Native speakers of American English (Sekkat et al., 2024; Herron et al., 2026a). We are thus left with 1038 speakers distributed over gender (male, female), dialect (Native, Latino, Asian), age (children, young adults, older adults), and ethnicity (Caucasian, African-American)[1].

When defining SG partitions for our experiments (i.e. defining the group of speakers who fall under "men" vs "women"), it is important to consider the multifaceted identities of individual speakers (Herron et al., 2026b). Sonos provides SG labels for four demographic variables - we take advantage of these to isolate *single* demographic variables as much as possible by removing influence of other confounding variables. For example, if most men in the dataset are African American and most women are Caucasian, if we sample an equal number of men and women, we are likely to at least partially inadvertently measure the effect of ethnicity in addition to gender (Veliche et al., 2024). To counteract this, in our analyses we avoid speakers from non-mode SGs (with respect to the dataset's construction) SGs. That is to say, when testing the effect of gender, we include neither children, old people, non-native speakers, or speakers with labeled ethnicity[2]; when testing the effect of age, we include no non-native speakers and speakers with ethnicity labeled; when testing the effect of dialect, we include no children. See Appendix Section B for further technical details about phoneme embedding extraction.

Retaining only mode speakers has the advantage of avoiding certain sources of dataset-driven bias, but adds a fresh layer of experimental bias by not taking uncommon SG combinations into account. By ignoring non-native children, for example, we are unable to measure the potentially confounding effect of the intersectionality of those SGs together (Wang et al., 2022).

4.2 Backbone Encoder Models

This paper studies speech encoder models, and we focus in particular on self-supervised speech processing models (S3Ms) (Baevski et al., 2020). S3Ms use only audio during pretraining, isolating them from potential bias that comes from finetuning for ASR. Furthermore, pretrained S3Ms have been shown to model phonetic rather than semantic aspects of speech (Choi et al., 2024). This insulates them from bias related to semantics, which facilitates our error analysis. S3Ms are also useful starting points for transfer learning and can be finetuned using little labeled training data to solve diverse downstream tasks, from ASR to SID

[1] Ethnicity labels are only available for a small subset of the dataset (98 speakers).

[2] Sonos labels ethnicity for less than 10% of all speakers, so we have no way of knowing the ethnicities of the other speakers. By using only speakers with no ethnicity tag, at least we avoid measurably oversampling one ethnicity over another. However, this is a potential source of bias in our analysis, given that we show that ethnicity does impact phoneme embeddings.

(Baevski et al., 2020). However, we also study the Whisper encoder, which was trained end-to-end on ASR. Whisper achieves state of the art ASR performance, so it is particularly informative to understand its behavior (Radford et al., 2022).

The models we include in our study include WavLM-base-plus (WavLM-base+), WavLM-large, Wav2vec 2.0-large-ls (W2V2-lg-ls), Wav2vec 2.0-large-xlsr-53 (XLS-R), DeCoAR 2.0 (DeCoAR2), and Whisper-medium (Whisper-med) - one for each of the three families of S3Ms according to the typology defined in Mohamed et al. (2022). We choose WavLM models (Chen et al., 2022) as they have been empirically shown to create the best universal speech embeddings for English of any S3M based on the SUPERB benchmark (Yang et al., 2021). We compare between the base and large (100M and 300M parameters respectively) model sizes, both of which were trained on the 60k hour librivox corpus. We also include two Wav2vec 2.0 models (Baevski et al., 2020), as it remains the most popular S3M according to Huggingface downloads. Furthermore, we can compare the effect of pretraining on a small, constrained, and regular corpus like LibriSpeech in Wav2vec 2.0-large-ls vs a more varied, multilingual corpus in Wav2vec 2.0-large-xlsr-53 (Babu et al., 2021). Finally, we consider the encoder module of Whisper-medium (Radford et al., 2022), which has the same number of parameters as the large S3Ms. The pretraining data for Whisper are not public which makes it more challenging to compare with its peers. We used Speechbrain to extract phoneme embeddings and train our PR probes (Ravanelli et al., 2021).

We finetune each of the pretrained S3Ms for ASR using the same data and training loop to ensure optimal comparability. All of our adaptations are trained based on Speechbrain recipes using a subset of 1500 speakers sampled from CommonVoice 16 (Ardila et al., 2020). We finetune using using CTC loss and a 3-layer MLP decoder. We start by freezing the S3M and train the decoder until convergence, then unfreeze the S3M and train for 30k steps.

4.2.1 Domain Enhancing/Adversarial Finetuning One common technique for improving fairness in speech processing is domain enhancing or adversarial training (DET/DAT), in which the model is forced to maximize or minimize some type of information at specific layers (Zhou et al., 2023; Tanaka et al., 2022; Na and Park, 2021). Typically, this involves learning a classifier for speaker attributes (e.g. accent or speaker ID), and using its gradient to force the backbone encoder model to create embeddings that maximize speaker information in middle layers and minimize it in final layers (using a gradient reversal layer for adversarial training). This is typically applied as a multi-task objective during ASR finetuning (CTC + DET/DAT).

Empirical results using CTC + DET/DAT have not been exemplary. Recent work has shown that this could be due to models being already irreparably biased during pretraining, and that fairness interventions ought to come *before* ASR finetuning (Herron et al., 2026a), as well as that ASR finetuning by itself tends to cause encoder models to discard most speaker information by the final layer (Herron et al., 2025). However, we are interested to see the effects of DET/DAT on the phoneme-embedding level. We follow Zhou et al. (2023), learning enhancing and adversarial speaker ID classifiers on layers 5 and 10 for base models, and 10 and 21 for large models respectively. Following Das et al. (2021), we first warm up the model to conversion on both classifiers before unfreezing its weights. Following Herron et al. (2025), we use an x-vector as our adversarial classifier to avoid learning a too narrow classifier and retaining speaker information in subspaces orthogonal to the linear classifier (Snyder et al., 2018). Following Pasad et al. (2022), we re-initialized the final two layers of Wav2vec 2.0 models, as they are overfit to their initial contrastive objective and do not provide good starting points for other speech modelling tasks.

5 Methodology

We perform three types of experiments to measure both the variance and bias in phoneme embeddings for each SG in the Sonos corpus.

5.1 Indirectly measuring embedding variance vs bias via PR probes

For our first test, we train phoneme classification probes on varying subsets of the extracted (and frozen) phoneme embeddings using a **linear** probe. We chose the simplest possible architecture in order to avoid as much potential downstream bias infiltration as possible into our probes. (The one obvious downside of this choice is that real life ASR use cases rarely use raw S3M embeddings without an more complex transformation to an output space). For each demographic variable (gender,

age, dialect, ethnicity), we have **two probe training settings**:

1. Train the linear probe using data from only a single SG (e.g. men). We use the same number of speakers no matter which SG[3] - (e.g. the same number of men in the men-only as the number of women in the women-only).

2. Train the linear probe using data from a balanced number of speakers from each SG (i.e. the same number of men as women). Critically, we use the same number of speakers overall as in the previous setting to ensure direct comparability.

We then measure the macro F1 accuracy for each classifier in predicting each phoneme for each SG under each setting, at each layer of each encoder model. We replicate each experiment five times using different speakers for each replication (apart for the SG with the maximum number of speakers for its corresponding demographic variable, thereby defining the number of speakers used for all settings of that demographic variable). Such training-set balancing is a common strategy in fairness research (ElGhazaly et al., 2025b,a; Garnerin et al., 2021), though it is often used in training or finetuning entire networks. Our intervention has the potential to impact far fewer parameters (only those of the phoneme probe), so we are likely to see more consistent results than those observed in aforementioned studies.

5.2 Random error as KNN distance

We can also directly measure the random error of phoneme embeddings for each speaker of each SG (Figure 1 upper right). To this end, we measure the KNN distance between phoneme embeddings for individual speakers to determine how closely clustered they are. See Appendix Section A for an explanation as to why KNN distance is an appropriate metric for measuring random error in phoneme embeddings.

Our procedure is as follows: fixing a speaker s and layer M_ℓ, we first normalize over all phoneme embeddings to ensure that the phoneme embeddings are in a unit-mean/standard deviation space.

[3]Previous studies have shown the number of speakers to be a critical factor in generalizability of tems (Maison and Estève, 2023; Berrebbi et al., 2023; Whetten et al., 2026). Even though this is just a probe, we are fastidious in equally balancing numbers of speakers to avoid conclusions based on random dataset imbalance.

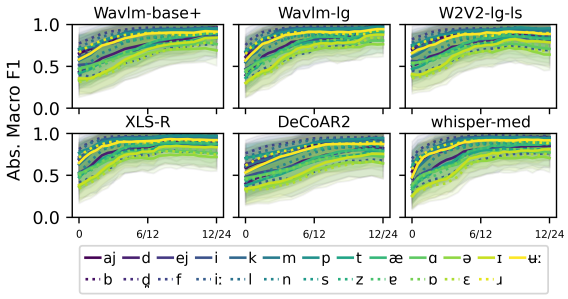

Figure 2: Absolute F1 score at every layer for each phoneme of each ASR-finetuned encoder model.

Then, we reduce dimensionality by retaining the dimensions which account for 95% of all variance between phoneme embeddings for all phonemes, using principal component analysis (PCA). This greatly reduces the size of embeddings and potentially winnows out noise. Then, for each phoneme p, we calculate the KNN distance as the squared L2 distance from each embedded sample of p (i.e. p_i^{s,M_ℓ} for $i \in [1..30]$) to its k-nearest neighbours of the same phoneme (i.e. $p_{\{j,k,l\}}^{s,M_\ell}$ for $i \notin \{j,k,l\}$). We use the same number of samples N=30 and k=3 for each speaker/phoneme to ensure comparability between phonemes, speakers, and layers. We then aggregate over all speakers for a SG to determine a distribution of KNN distance over each SG. If an S3M produces embeddings with higher KNN distance for one SG than another, that is evidence of higher random error in the way that SG is processed by the encoder.

6 PR probing results

Table 1: Overall F1 phoneme classification rate for best overall layer of each ASR-finetuned encoder model, for PR probes trained on balanced data for each demographic variable (train setting 2). Gap for each demographic variable is the percent difference between the worst and best performing SGs (Dheram et al., 2022).

	Wavlm-base+	Wavlm-lg	W2V2-lg	XLS-R	DeCoAR2	whisper-med
Macro F1	0.88	0.91	0.87	0.90	0.84	0.88
Gender gap	0.08	0.64	0.21	0.21	0.86	0.03
Age gap	3.85	2.74	4.18	3.02	3.95	3.51
Dialect gap	6.26	4.48	7.70	6.23	7.28	6.33
Ethnicity gap	2.39	2.43	3.46	2.34	3.01	1.90

Over all of the following experiments, the one overarching pattern is that all the encoder models we test behave roughly equivalently with respect to fairness in PR. While some models have higher overall performance (WavLM-large is the best, see row 1 of Table 1), the relative performances between SGs are similar between all models (see rows

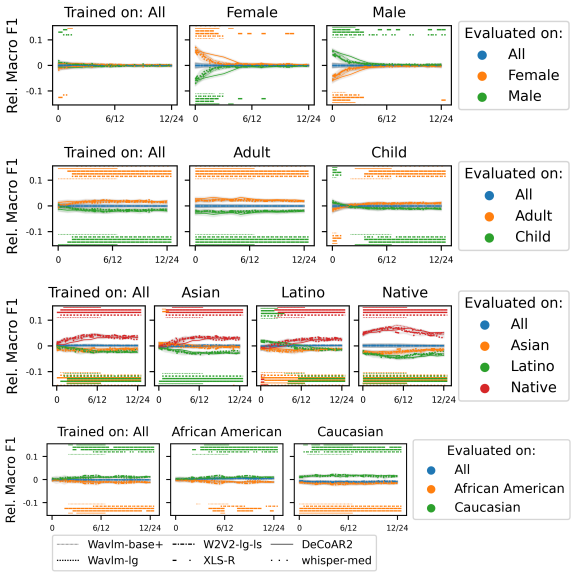

Figure 3: Macro F1 phoneme classification scores, **relative to the macro average over all SGs** (e.g. men and women) for the corresponding demographic variable (e.g. gender) for ASR-finetuned S3Ms. Values > 0 indicate that SG has a better-than-average macro F1 (e.g. top left: when trained on all data, females have above-average macro F1 performance in layer 0). Horizontal lines on top and bottom of each figure denote statistical significance (for $p < 0.05$) for relative F1 $< / > 0$ respectively on 1-sided 1-sample t-test.

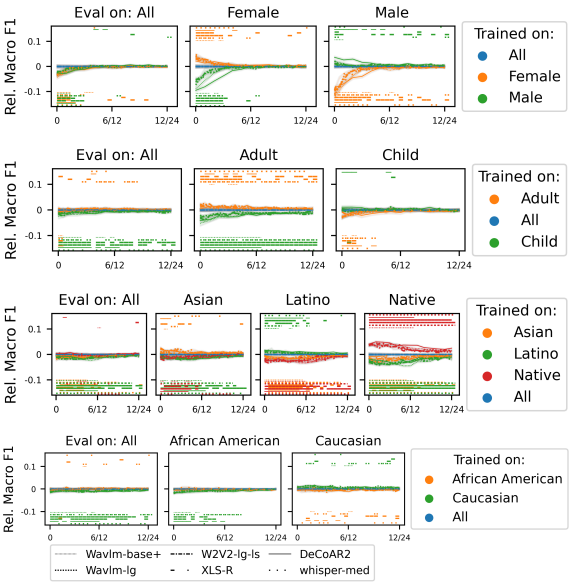

Figure 4: Macro F1 phoneme classification scores, **relative to probe training on a balanced dataset** for SGs from the corresponding demographic variable for ASR-finetuned S3Ms. Values > 0 indicate that training on that SG results in better performance than for a balanced dataset (e.g. top left: when phoneme classifiers are trained on only men, men have a higher PR than for probe training on balanced data). Horizontal lines on top and bottom of each figure denote statistical significance (for $p < 0.05$) for relative F1 $< / > 0$ respectively on 1-sided 1-sample t-test.

2-5 of 1). Furthermore as Figure 2 shows, while some phonemes are harder to identify overall, all models tend to struggle with the same phonemes. We also replicate these experiments on pretrained only S3Ms and find similar results (see Appendix Section C) - thus we can rule out this being an artifact of our ASR finetuning data.

We also note an overall lack of evidence supporting embedding bias in finetuned S3Ms. When compared to pretrained models (see Appendix Section C), finetuning for ASR seems to reduce any bias that might have persisted. This is not terribly surprising - Herron et al. (2025) shows that SG information is dropped during ASR finetuning, thus eliminating a source of embedding bias.

6.1 Analysis of in-domain training overall

Figures 3 and 4 show the relative macro F1 score for S3Ms (and `Whisper`) finetuned for ASR. The first figure shows the effect on all SGs when training the probe a single SG (or all SGs); the latter shows the effect for a single SG of training the probe on a single SG (or all SGs). We compute 1-sample 1-sided t-tests for each model, layer, and SG, to compare it with the baseline - for Figure 3, the baseline is average F1 over all SGs; for Figure 4, the baseline is by-SG performance for probes trained on balanced training data. Following Herron et al. (2026b), we first compute the mean over all speakers, then calculate the mean for each SG to avoid biasing our calculations towards speakers who were randomly selected more frequently when sampling speakers for the train/test splits.

6.1.1 Does SG-specific probe training change relative performance among SGs? First, when examining the left-most columns of Figure 3, for probes trained on a balanced group of all SGs, our results overall are in line with previous studies analyzing `Sonos` (Sekkat et al., 2024; Herron et al., 2026a) - native speakers have the best performance for all models, while adults and Caucasians do also (though not significantly for any model on our tests, likely due to small numbers of speakers per SG). For gender, there is most discrepancy in the first several layers, while for dialect, there is greatest discrepancy in middle to later layers.

Furthermore, when regarding the other columns

(phoneme probes trained on speakers from a single SG) we see that in-domain training has little effect on the pecking order of relative PR performance. There is the greatest divergence for the first several layers of each model, where the polarity sometimes flips (for gender and dialect, for example). However, those layers have poor overall PR performance (see Figure 2). For later layers with better overall PR, the same SGs get the best PR performance **regardless of which SG is used in training the PR probes**. When viewed from a general perspective (in calculating macro F1), this result sets a low ceiling for the amount of embedding bias/systematic error in phoneme embeddings for any of the models we tested. It also corroborates the finding in Liu et al. (2023) that speaker and phoneme information are modeled orthogonally - if there were more mixture of the two, then we would expect performance improvements for in-domain phoneme probe training.

6.1.2 Does in-domain training help individual SGs? Figure 4 shows the effect on individual SGs when training phoneme probes on speakers from a single SG. For overall performance, we observe that training on a balanced dataset is almost always best, apart for the first few layers of models for dialect. For gender, we observe that in-domain probe training is beneficial for both females and males in early layers but that they experience less boost in later layers (though still statistically significantly so for some S3Ms). At the same time, the out-of-domain gender experiences degraded performance.

For age, we observe that training on only children benefits no one, not even children, while training on only adults is beneficial for adults. We interpret this as emblematic of greater variance in children's phoneme embeddings - in-domain training on noisier data results in a worse calibrated classifier, and testing on children's data is inherently harder due to it being noisier. If there were significant embedding bias, we would observe children getting some boost due to in-domain training.

The most interesting category is dialect, where we observe a clearer example of in-domain training being beneficial, albeit only barely (and not significantly for all models), likely indicating embedding bias. Asian, Latino, and Native speakers (row 3, columns 2, 3, and 4 respectively) achieve their best performance at many layers when trained on only similar speakers. For ethnicity, we observe minimal advantage gained from in-domain training for

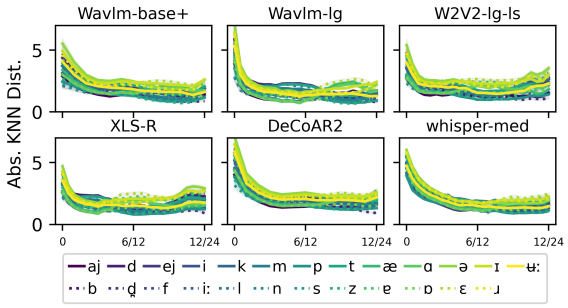

Figure 5: Absolute KNN distance for each phoneme, averaged over all speakers. Higher values for KNN distance represent greater variance in phoneme embedding.

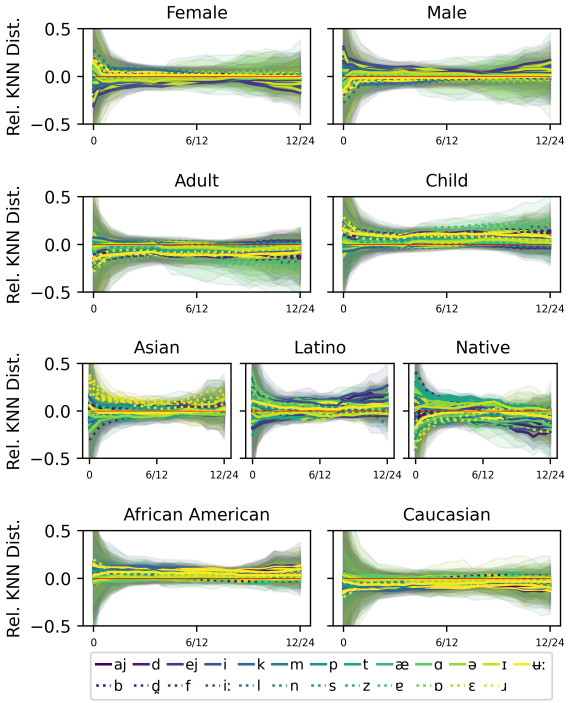

Figure 6: Relative KNN distance between embeddings of the same phoneme for the same speaker at each layer.

Caucasian speakers but none for African-American speakers.

It is important to note that these are macro F1 scores over all phonemes, and might not highlight bias in individual phonemes if their effect is small and they are few in number. Appendix Section D shows that some phonemes benefit more than others from in-domain probe training.

7 KNN distance results

Figure 5 shows the absolute KNN distance for all phonemes across all layers for ASR finetuned encoder models. Note how the KNN distance decreases layer by layer over the first several model

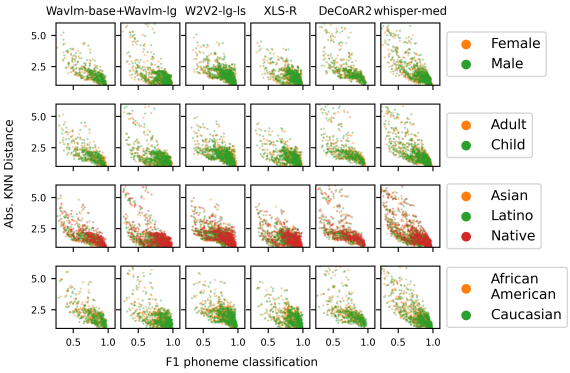

Figure 7: Relationship between KNN distance and phoneme classification rate. Each datapoint is a single phoneme at a single layer for a single SG. Statistically significant pearson's r with $p < 0.001$ for every model and every demographic variable.

layers for every phoneme and model - this is logical, given that PR accuracy tends to increase over these layers, thus distance between embeddings should decrease (Pasad et al., 2022).

Figure 6 shows the relative KNN distance between closest pairs of phoneme embeddings for the same phoneme and speaker at each layer, with each S3M superimposed. (See Appendix Figure 15 for a cleaner view with a single S3M). As before, we note that all models exhibit similar behavior. Most have high KNN distances for most phonemes in earlier layers, which is logical as phonemes are not as well defined in earlier layers and PR is poorer.

Our KNN distance results resemble those for phoneme classification towards later layers. Females and males have similar KNN distances for all phonemes; adults have much lower KNN distance for all phonemes than children, as do Caucasians. Native speakers likewise have lower KNN distance for most phonemes than Latinos and Asians by the final layers, though middle layers are more varied. This is logical - dialect was the demographic variable in which we saw the greatest evidence for embedding bias, as opposed to age and ethnicity.

To further establish the connection between phoneme classification and KNN distance, we examine the direct relationship between the two of them for every layer of every model, divided by SG. Figure 7 shows that phonemes that achieve higher classification accuracy also have lower KNN variance. This supports the utility of KNN distance as a metric for understanding phoneme modelling error. While we have no evidence for a causal relationship between the two, this observation does suggest that learning strategies that minimize KNN distance could in turn help minimize phoneme classification error.

8 Finetuning using CTC + DET/DAT

We investigate the effect that using DET/DAT during ASR finetuning has both on the effect of training PR probes on single SGs, as well as KNN distance. Figure 8 shows the difference between relative macro F1 scores on S3Ms finetuned on ASR alone vs CTC + DET/DAT. If DET/DAT were to reduce embedding bias, we would see curves below 0 for SGs which are disadvantaged due to embedding bias in the normal ASR setting, and above 0 for SGs which are advantaged. However, we see that there is no statistically significant difference between embeddings trained using DET/DAT for any SG in any demographic variable apart from the earliest layers where variance is high and PR is poor. This shows that if there is any embedding bias (for which we didn't find strong evidence to begin with, at least in later layers), DET/DAT is not reducing it.

Figure 9 shows the difference in relative KNN distances for models finetuned for ASR vs CTC + DET/DAT. If DET/DAT were to reduce relative KNN variance for a SG with previously higher-than-average KNN variance, we would see curves below the zero line. However, as before, we find negligible impact on KNN distance when finetuning using DET/DAT. DET/DAT is not helping reduce KNN variance for any SG relatively, and as a consequence no SG at all in the absolute.

These findings complement Herron et al. (2025) and Herron et al. (2026a) which show in more abstract terms that ASR finetuning with DET/DAT does not have a tangible effect on fairness. Our findings provide a clearer answer as to why: we see little by way of embedding bias in ASR-finetuned S3Ms to begin with. DET/DAT is meant to enhance speaker information in middle layers and erase it in later layers; however, unfairness in phoneme prediction doesn't appear to be an embedding bias issue, and more of a variance issue. Thus, it is not totally surprising that this method is ineffective at bringing forth fairer phoneme embeddings.

9 Discussion and outlook

The purpose of this study is to provide tools for better understanding types of error in phoneme embeddings. The more insights we have into model

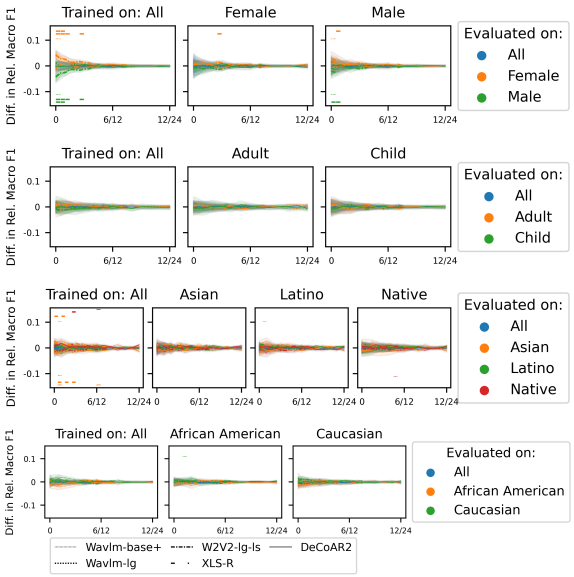

Figure 8: Difference in relative (to balanced training data) macro F1 phoneme prediction results between S3Ms finetuned on ASR and S3Ms finetuned on ASR *and* DET/DAT. Values < 0 imply that a SG performs worse relative to an average of all SGs for probes trained on CTC + DET/DAT than on just ASR. Horizontal lines on top and bottom of each figure denote statistical significance (for $p < 0.05$) for a difference in relative F1 $< / > 0$ respectively on 2-sided 2-sample t-test.

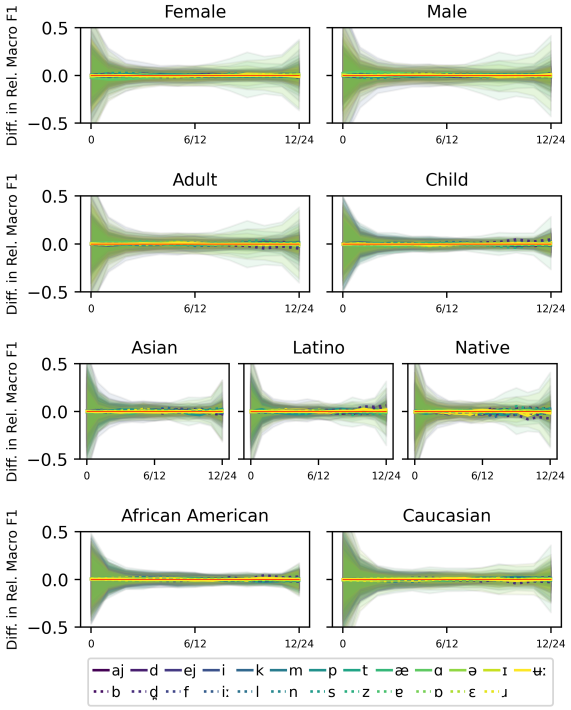

Figure 9: Difference in relative KNN distance between embeddings of the same phoneme for the same speaker at each layer. Values below zero indicate that a SG has lower average relative KNN distance for that phoneme on CTC + DET/DAT than on ASR only; however, we observe negligible deviation from zero.

bias, the better we can be at making precise interventions to improve fairness. We show that varied models all produce phoneme embeddings for speakers of particular SGs that can have both an embedding bias and/or unequal variance. Furthermore, we show that finetuning for ASR did next to nothing to change this, even when finetuning using the fairness enhancing DET/DAT. This shows the limitations of current speech encoder systems in fairly modelling phonemes from different speaker groups. In demonstrating that high variance is strongly associated with poor phoneme recognition results (as opposed to embedding bias), we conclude that current strategies for fairness enhancement are potentially ill-equipped to improve fairness, as they are designed to treat bias more than variance issues.

Reducing the problem of demographic fairness to a geometric modelling problem could provide a cleaner angle of tractability for future work in designing fairer ASR systems. In particular, we encourage further research into methodology that could stabilize variance/KNN distance in phoneme embeddings, both overall as well as for particularly SGs. There is ample work in other domains of self-supervised learning working on solving this type of problem, such as using Siamese networks (Chen and He, 2020) or maximizing the effective entropy of embeddings (Chakraborty et al., 2025). Potentially methods such like supervised contrastive learning could be used to reduce KNN distance for disadvantaged SGs (Khosla et al., 2021). Indeed, some work has already used similar methodology in speech processing applications, though not with respect to fairness (Li et al., 2025).

We also encourage further research into error analysis from a causal perspective. Our results are a useful first step into better understanding unfairness in ASR but we are still lacking answers as to what exactly is responsible for either the variance or bias we observe in phoneme embeddings. With the amelioration of voice generation models, causal intervention on speech is increasingly viable and effective (Masson and Carson-berndsen, 2023; Baas and Kamper, 2022). Future work could thus make precise manipulations to an input signal corresponding to demographic characteristics associated with any protected SG, to discern where and how along the layerwise computation phoneme representations begin to shift in response.

Limitations

We conducted our analysis on the Sonos dataset primarily due to its metadata, allowing for fairness analysis. However, this limits its generalizability to speech processing in general due to the controlled nature of recordings. Speech corpora with high quality annotations are few and far between - we had previously worked with Fair-speech (Veliche et al., 2024) but it lacks speaker IDs, so we cannot control for the number of speakers per SG, rendering our analyses much less useful.

Furthermore, we used 100- and 300-million parameter models, much smaller than those used to achieve state of the art ASR performance, such as Whisper-large. It is possible, due to increased capabilities of larger models, that the effects we note in our study would be different on such models.

Acknowledgements

This research has been funded by the French National Research Agency (ANR), project "E-SSL" (ANR-22-CE23-0013). It was also partially supported by ANR through the MIAI "AI & Language" chair and the MIAI "Socialization and Language at School" chair (ANR-23-IACL-0006). This work was performed using HPC resources from GENCI at IDRIS under the allocations 2023-A0151014633, 2024-A0171014633 on the Jean Zay supercomputers.

References

Laura Alonzo-Canul, Benjamin Lecouteux, and François Portet. 2025. Vers l'entraînement de modèles de reconnaissance automatique de la parole autosupervisés équitables sans étiquettes démographiques. In *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 780–790, Marseille, France. ATALA \textbackslash\textbackslash& ARIA.

Frederic Aman, Michel Vacher, Solange Rossato, and Francois Portet. 2013. Speech recognition of aged voice in the AAL context: Detection of distress sentences. In *2013 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD)*, pages 1–8, Cluj-Napoca, Romania. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. *Preprint*, arXiv:1912.06670.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps. *Preprint*, arXiv:2402.17954.

Matthew Baas and Herman Kamper. 2022. Voice Conversion Can Improve ASR in Very Low-Resource Settings. *Preprint*, arXiv:2111.02674.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *Preprint*, arXiv:2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Preprint*, arXiv:2006.11477.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, asdf, pages 1–5.

Dan Berrebbi, Ronan Collobert, Navdeep Jaitly, and Tatiana Likhomanenko. 2023. More Speaking or More Speakers? In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Deep Chakraborty, Yann LeCun, Tim G. J. Rudner, and Erik Learned-Miller. 2025. Improving Pre-trained Self-Supervised Embeddings Through Effective Entropy Maximization. *Preprint*, arXiv:2411.15931.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Xinlei Chen and Kaiming He. 2020. Exploring Simple Siamese Representation Learning. *Preprint*, arXiv:2011.10566.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-Supervised Speech Representations are More Phonetic than Semantic. *Preprint*, arXiv:2406.08619.

Nilaksh Das, Sravan Bodapati, Monica Sunkara, Sundararajan Srinivasan, and Duen Horng Chau. 2021. Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning. *Preprint*, arXiv:2103.05834.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I.-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities. In *Interspeech 2022*, pages 1268–1272.

Hend ElGhazaly, Bahman Mirheidari, Heidi Christensen, and Nafise Sadat Moosavi. 2025a. Fairness in Automatic Speech Recognition Isn't a One-Size-Fits-All. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19169–19178, Suzhou, China. Association for Computational Linguistics.

Hend ElGhazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025b. Exploring Gender Disparities in Automatic Speech Recognition Technology. *Preprint*, arXiv:2502.18434.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying Bias in Automatic Speech Recognition. *Preprint*, arXiv:2103.15122.

Rita Frieske and Bertram E. Shi. 2024. Hallucinations in Neural Automatic Speech Recognition: Identifying Errors and Hallucinatory Models. *Preprint*, arXiv:2401.01572.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the Impact of Gender Representation in ASR Training Data: A Case Study on Librispeech. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92, Online. Association for Computational Linguistics.

Felix Herron, Maja Hjuler, Solange Rossato, Alexandre Allauzen, and François Portet. 2026a. Where do self-supervised speech models become unfair?

Felix Herron, Ange Richard, François Portet, Alexandre Allauzen, and Solange Rossato. 2026b. Responsible benchmarking of fairness for automatic speech recognition. In *Proceedings of the Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis workshop (SPEAKABLE) @ LREC 2024*, Palma, Spain. ELRA.

Felix Herron, Solange Rossato, Alexandre Allauzen, Benoit Favre, and François Portet. 2025. Speaker Group Encoding in Self-supervised Speech Recognition Models. In *Text, Speech, and Dialogue*, pages 121–132, Cham. Springer Nature Switzerland.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised Contrastive Learning. *Preprint*, arXiv:2004.11362.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Kewei Li, Hengshun Zhou, Kai Shen, Yusheng Dai, and Jun Du. 2025. Phoneme-Level Contrastive Learning for User-Defined Keyword Spotting with Flexible Enrollment. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Hyderabad, India. IEEE.

Yu-Xiang Lin, Chih-Kai Yang, Wei-Chih Chen, Chen-An Li, Chien-yu Huang, Xuanjun Chen, and Hung-yi Lee. 2025. A Preliminary Exploration with GPT-4o Voice Mode. *Preprint*, arXiv:2502.09940.

Oli Liu, Hao Tang, and Sharon Goldwater. 2023. Self-supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces. *Preprint*, arXiv:2305.12464.

Lucas Maison and Yannick Estève. 2023. Some Voices are Too Common: Building Fair Speech Recognition Systems Using the CommonVoice Dataset. In *INTERSPEECH 2023*, pages 4428–4432. ISCA.

Margot Masson and Julie Carson-berndsen. 2023. Investigating Phoneme Similarity with Artificially Accented Speech. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 49–57, Toronto, Canada. Association for Computational Linguistics.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502. ISCA.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Mukhtar Mohamed, Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2024. Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations. *Preprint*, arXiv:2406.09200.

Hyeong-Ju Na and Jeong-Sik Park. 2021. Accented Speech Recognition Based on End-to-End Domain Adversarial Training of Neural Networks. *Applied Sciences*, 11:8412.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2022. Layer-wise Analysis of a Self-supervised Speech Representation Model. *Preprint*, arXiv:2107.04734.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. *Preprint*, arXiv:2211.03929.

Archiki Prasad and Preethi Jyothi. 2020. How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3739–3753, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Anand Rai, Satyam Rahangdale, Utkarsh Anand, and Animesh Mukherjee. 2025. ASR-FAIRBENCH: Measuring and Benchmarking Equity Across Speech Recognition Systems. *Preprint*, arXiv:2505.11572.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A General-Purpose Speech Toolkit. *Preprint*, arXiv:2106.04624.

Sandra Rojas, Elaina Kefalianos, and Adam Vogel. 2020. How Does Our Voice Change as We Age? A Systematic Review and Meta-Analysis of Acoustic and Perceptual Voice Data From Healthy Adults Over 50 Years of Age. *Journal of Speech, Language, and Hearing Research*, 63(2):533–551.

Chloe Sekkat, Fanny Leroy, Salima Mdhaffar, Blake Perry Smith, Yannick Estève, Joseph Dureau, and Alice Coucke. 2024. Sonos Voice Control Bias Assessment Dataset: A Methodology for Demographic Bias Assessment in Voice Assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15056–15075, Torino, Italia. ELRA and ICCL.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Tomohiro Tanaka, Ryo Masumura, Hiroshi Sato, Mana Ihori, Kohei Matsuura, Takanori Ashihara, and Takafumi Moriya. 2022. Domain Adversarial Self-Supervised Speech Representation Learning for Improving Unknown Domain Downstream Tasks. In *Interspeech 2022*, pages 1066–1070. ISCA.

John R. Taylor. 2022. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. MIT Press.

Benjamin van Niekerk, Leanne Nortje, Matthew Baas, and Herman Kamper. 2021. Analyzing Speaker Information in Self-Supervised Models to Improve Zero-Resource Speech Processing. *Preprint*, arXiv:2108.00917.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-Speech dataset. *Preprint*, arXiv:2408.12734.

Ravichander Vipperla, Steve Renals, and Joe Frankel. 2008. Longitudinal study of ASR performance on ageing voices. In *Interspeech 2008*, pages 2550–2553. ISCA.

Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 336–349.

Ryan Whetten, Titouan Parcollet, Marco Dinarelli, and Yannick Estève. 2026. A Study of Data Selection Strategies for Pre-training Self-Supervised Speech Models. *Preprint*, arXiv:2601.20896.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I. Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing Universal PERformance Benchmark. *Preprint*, arXiv:2105.01051.

Wei Zhou, Haotian Wu, Jingjing Xu, Mohammad Zeineldeen, Christoph Lüscher, Ralf Schlüter, and Hermann Ney. 2023. Enhancing and Adversarial: Improve ASR with Speaker Labels. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

A Motivation for KNN distance metric

Our first intuition when considering calculating random error in phoneme embeddings was to simply calculate the Euclidean distance from each phoneme embedding to the mean for each phoneme and speaker, i.e. the variance. However, we realized that was a potentially foolish proposition, as it assumes a distribution with a single mode or mean. There is no reason to assume this to be the case for phoneme embeddings, and indeed, Figure 10 provides a visualization showing it to be otherwise for some phonemes. At each layer, we train a 2-dimensional PCA reduction on embeddings for all phonemes, then visualize several phonemes individually. Note that some layers' embeddings for some phonemes are clearly bimodal starting in later layers (e.g. 'r' in particular), while none is a 2-dimensional Gaussian, apart from the earliest layers. Indeed, this is a mere 2-dimensional reproduction; there may be further multidimensional means that are tamped down by PCA.

With that in mind, Figure 10 provides a motivating example for why KNN-distance is a more suitable metric for measuring phoneme distribution variance. It remains low for a multimodal distribution where each mode has low variance. The red curve (Euclidean distance from the mean) jumps up in later layers, despite the phonemes being being increasingly well defined; on the other hand, the green curve (KNN distance) often decreases monotonically and certainly doesn't jump when multimodality appears.

B Extraction of phoneme embeddings

We use the Montreal Force Aligner (McAuliffe et al., 2017) version 3.3.4 to obtain phoneme-level alignments for the entire `Sonos` corpus, which we use to extract phoneme embeddings from each S3M at every layer. We use the "english_mfa" model and corresponding "english_us_mfa" alignment dictionary, due to our corpus being focused on American English. We select the 25 phonemes which are most frequently voiced over the corpus to include as full a phoneme embedding dataset as possible (see Figure 14 for a list of phonemes). It was unfortunately impractical to select all phonemes included in "english_mfa", as many speakers didn't use some rarer phonemes enough times for them to be included in phoneme-level analysis.

Using the phoneme alignments, we then extract phoneme alignments at every layer of each S3M. Following van Niekerk et al. (2021), we first normalize all frame embeddings at each layer over the entire utterance to remove global utterance-level information. Failing to remove utterance-level information would potentially add distracting information that could add noise to KNN distance calculations (although (Liu et al., 2023) shows that phoneme and speaker information are modeled in orthogonal subspaces, thus in theory this should be redundant). After normalizing, we follow Pasad et al. (2022) to construct our phoneme embeddings by mean-pooling over frames corresponding to a given phoneme. However, rather than taking all frames that fall within the given window, we take only the middle third of frames to reduce co-articulation effects. For each speaker, we sample 30 instances of each of the 25 most common phonemes at each transformer layer. We then filter out obvious alignment errors on a by-phoneme, by-speaker basis by discarding phoneme embeddings that are over three z-scores away from the mean over all instances of each phoneme/speaker.

The first row of Table 1 shows the overall rates of PR for the best layers each finetuned S3M. That we are able to achieve such a high rate of PR validates our embedding extraction algorithm.

C Analysis of pretrained S3Ms

We replicate PR experiments on pretrained S3Ms. Figures 11 and 12 are analogous to Figures 3 and 4. We note the same patterns in pretrained models as their ASR-finetuned complements. We likewise repeat out KNN distance analyses on pretrained S3Ms in Figure 13. (We excluded `Wav2vec 2.0` models to avoid visual contamination by their strange behaving final several layers (Pasad et al., 2022)). Note that pretrained models exhibit the same patterns of increased KNN variance in the same SGs as the ASR finetuned models in Figure 6.

D By-phoneme analysis of in-domain probe training

To get a more precise picture of embedding bias, we examine which phonemes experience the biggest boost from in-domain probe training. Figure 14 shows that some phonemes receive are classified better on a phoneme classifier trained in-domain, while others experience the opposite effect. Phonemes which experience improvement are likely those which are embedded in a biased

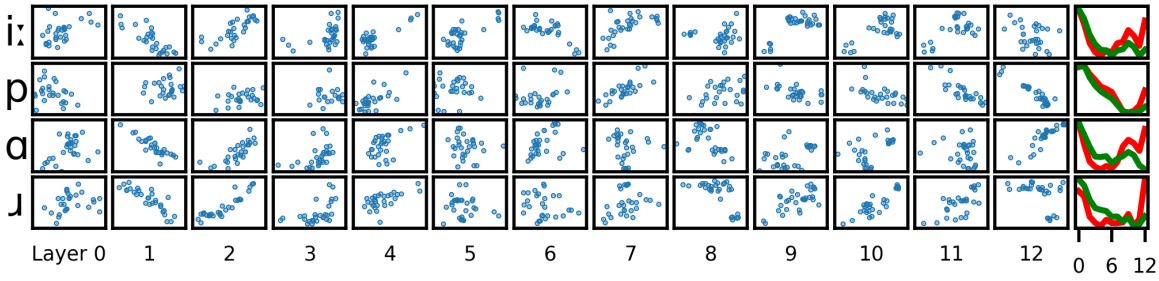

Figure 10: 2-dimensional PCA decompositions for phoneme embeddings for speaker 440 for four phonemes for `WavLM`-large finetuned on ASR, for each layer. Final column plots the average distance from the mean (red) and KNN distance (green).

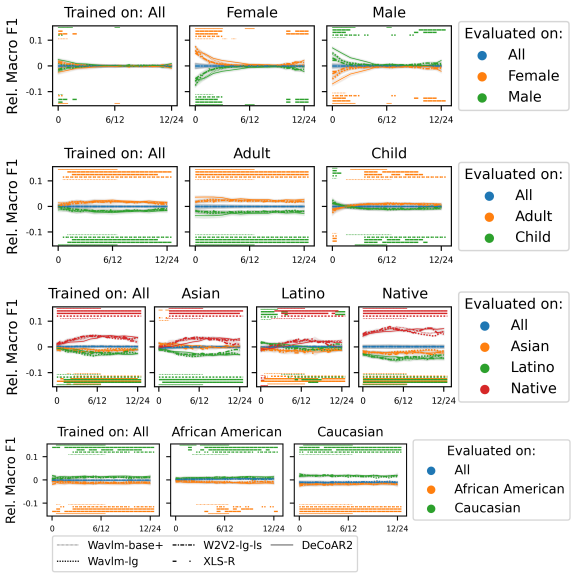

Figure 11: Macro F1 phoneme classification scores, **relative to the macro average over all SGs** (e.g. men and women) for the corresponding demographic variable (e.g. gender) for **pretrained S3Ms**. Values > 0 indicate that SG has a better-than-average macro F1 (e.g. top left: when trained on all data, females have above-average macro F1 performance in layer 0). Horizontal lines on top and bottom of each figure denote statistical significance (for $p < 0.05$) for relative F1 $< / > 0$ respectively on 1-sided 1-sample t-test.

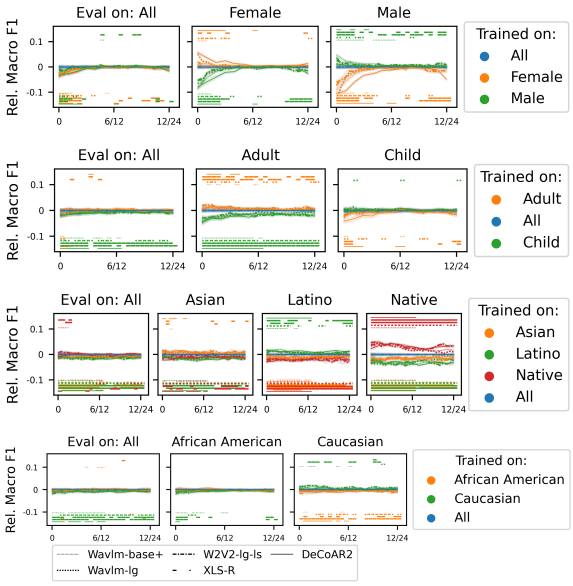

Figure 12: Macro F1 phoneme classification scores, **relative to probe training on a balanced dataset** for SGs from the corresponding demographic variable, for **pretrained S3Ms**. Values > 0 indicate that training on that SG results in better performance than for a balanced dataset (e.g. top left: when phoneme classifiers are trained on only men, men have a higher PR than for probe training on balanced data). Horizontal lines on top and bottom of each figure denote statistical significance (for $p < 0.05$) for relative F1 $< / > 0$ respectively on 1-sided 1-sample t-test.

manner by S3Ms, while those which experience degradation are likely casualties of higher variance for that SG. Note that the phonemes with significant performance improvement (or impairment) for in-domain training are largely shared across encoder models, indicating that all models exhibit similar trends of systematic and random error.

E KNN Distance for Wavlm-base-plus

Figure 15 shows the relative KNN distances for each SG for a single model to increase legibility.

E.1 Systematic error as linear SG separability

We would be interested, if possible, in directly quantifying systematic error as we did random error by KNN distance. That is to say, to calculate the extent to which phonemes are modeled around sep-

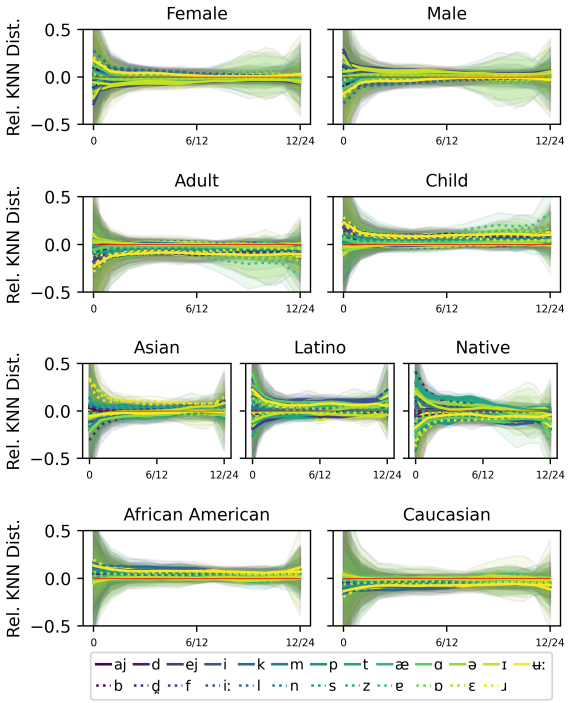

Figure 13: Relative KNN distance between embeddings of the same phoneme for the same speaker at each layer of **pretrained** S3Ms.

arate modes for different SGs (Figure 1 lower left). One potential method to this end is to train a linear classifier to predict SG (i.e. men or women) for every demographic variable (i.e. gender), based on phoneme embeddings. However, for this analysis to be sound, we would need filter out all utterance-level/speaker-level information from each phoneme embedding so that the classifier cannot use it in discriminating between SGs. One strategy to this end is to normalize out utterance-level information (recall Section B), which we have already done in data preparation.

However, in practice, we are skeptical of the completeness of this approach. Figure 16 shows that we can predict each demographic variable fairly well at every layer using every single phoneme embedding. This strongly implies that global speaker-level information is retained despite normalizing over the utterance.

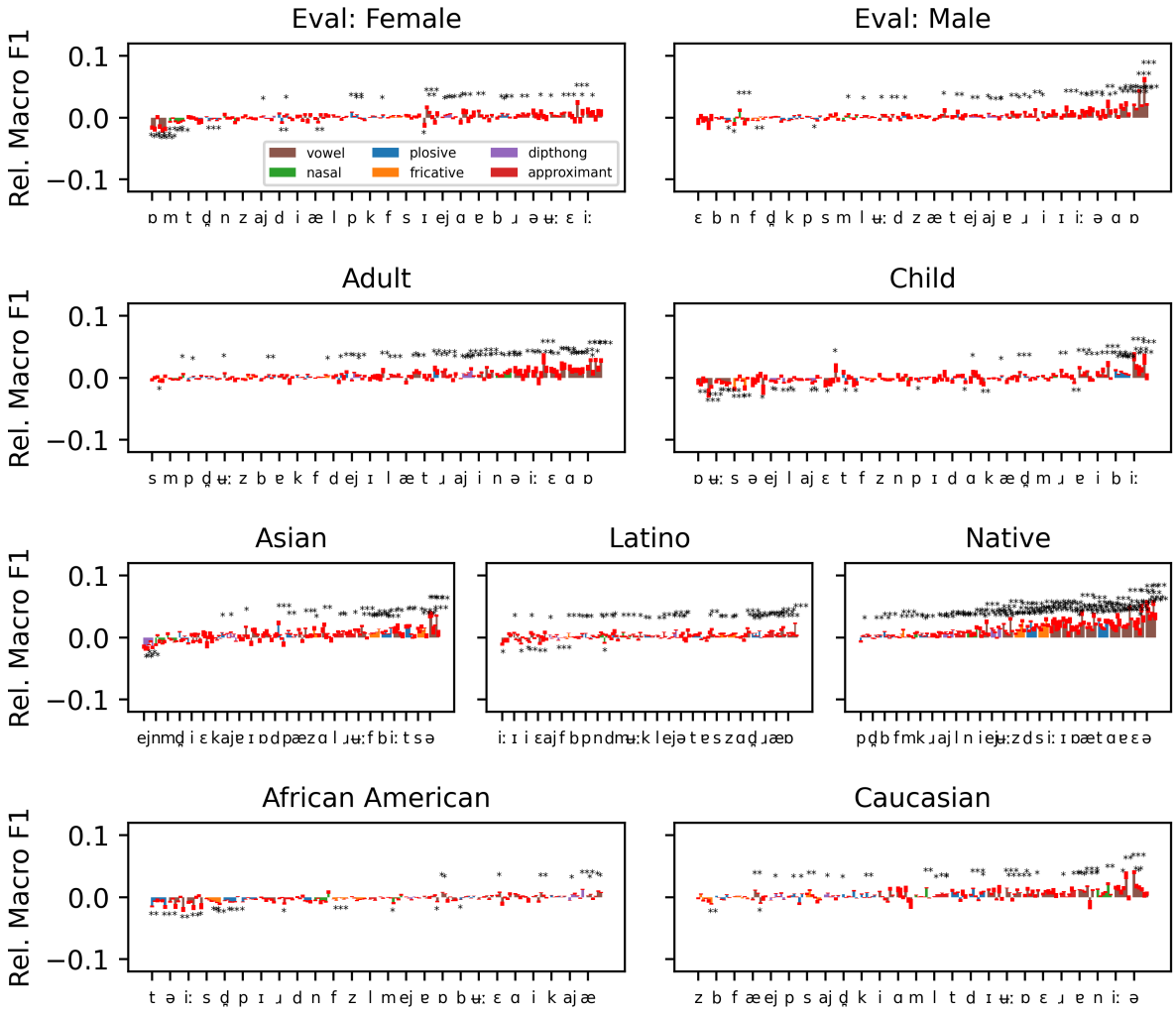

Figure 14: Macro F1 phoneme classification scores, **relative to probe training on a balanced dataset** for SGs from the corresponding demographic variable, for each phoneme separately, aggregated over the best-performing layers of all six encoder models. Values > 0 indicate that training on that SG results in better performance than for a balanced dataset. Each phoneme has six bars, one per encoder model. * represents statistical significance on a 1-sample 1-sided t-test: * indicates $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

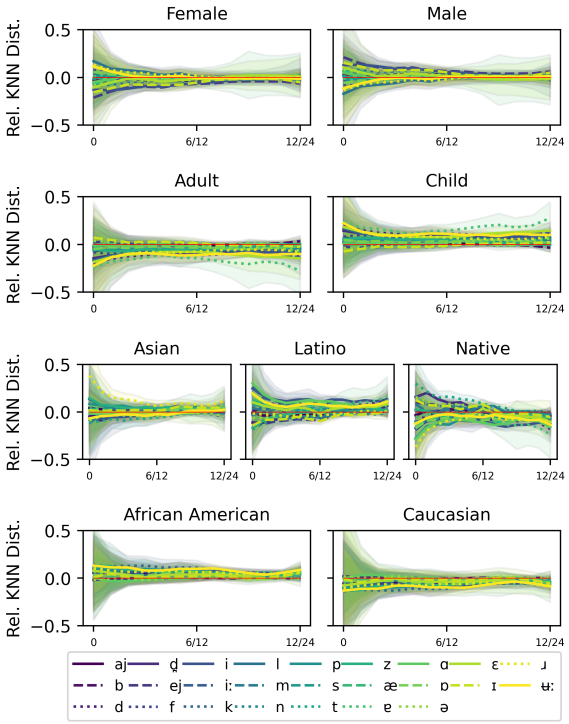

Figure 15: Relative KNN distance between embeddings of the same phoneme for the same speaker at each layer of WavLM-base-plus. This is a simplified version of Figure 6

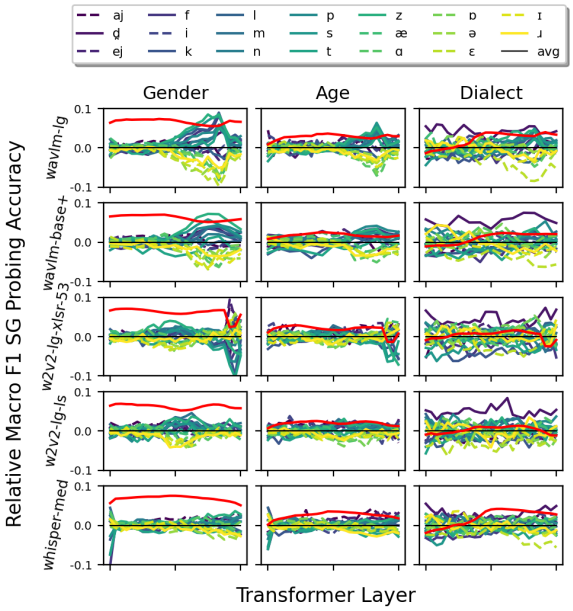

Figure 16: Relative macro F1 probing accuracy for each demographic variable (DV) based on embeddings from various phonemes. Red line is absolute average F1 over all phonemes. Scores < 0 means less SG information is present for any given phoneme/layer than the average over all phonemes at that layer.