ASR EVALUATION WITH PHONETIC EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating a speech system using mean or median Word Error Rate (WER) can obscure meaningful differences in performance among subgroups within the evaluation dataset, and among groups of speakers who will interact with the system. Prior work has shown that these disparities can be significant, and pose a challenge to adoption. However, to measure fine-grained disparities, developers typically rely on datasets that are annotated with speaker demographics. Such annotations are costly, noisy, and not aligned with speech behaviors that trigger recognition error. In this work, we propose a new data-driven technique for measuring error disparity. It is fully automatic, derived from meaningful speech behavior, and directly correlated with drivers of error. *Phonetic embeddings* are a data-driven representation of speech that can be used to cluster speakers and measure error disparity among clusters. Our technique addresses the gaps in current evaluation methods with a speaker representation that directly correlates with error, and a fully automatic method for disparity analysis. In an evaluation of an off-the-shelf commercial ASR system, we show that this method finds larger regression coefficients, performance gaps with similar or greater effect size, and better cluster quality, compared to demographic methods.

1 INTRODUCTION

By evaluating automatic speech recognition (ASR) systems only on aggregate metrics such as mean or median Word Error Rate (WER), developers may overlook significant disparities in performance across speakers. Prior work has shown that these disparities are systematic and can be large (Koenecke et al., 2020; Chang et al., 2024), posing challenges for equitable deployment. To move beyond aggregate evaluation, developers need methods that identify which speakers experience high error and why. The standard approach is to partition an evaluation dataset by demographic attributes like race, gender, and age, and look for disparities in per-group WER. However, demographic annotations are costly to collect and noisy when self-reported or externally assigned. In addition, these annotations are not explicitly provided as input to the ASR engine at inference time, meaning that they can not trigger errors directly. Intersectional demographics further complicate this approach, as they have a complex relationship to pronunciation that is not easily captured by categorical labels (Kortmann et al., 2004).

We propose a method to measure performance disparities directly in a phonetic embedding space, with no demographic labels required. We represent each speaker as a normalized vector of phone counts extracted from automatic phonetic transcriptions. By clustering these *phonetic embeddings* and measuring WER per cluster, we identify groups of speakers whose shared pronunciation patterns are associated with high or low recognition accuracy. A diagram of the processing pipeline is shown in Figure 1.

This approach has three advantages over demographic cohort analysis. First, phonetic embeddings capture speech behavior. This is a simple but important observation. While demographics and speech behaviors may overlap for many groups of speakers, the relationship is noisy. Second, phonetic embeddings reveal drivers of ASR error. Because phonetic features are derived from the speech signal, they use the same input that the ASR model consumes. Demographic categories are not an explicit part of the input to the ASR system, meaning that they can not be direct triggers of error. Third, the method is fully automatic. Phonetic embeddings require no manual annotation, making it applicable at scale to any dataset for which audio is available.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

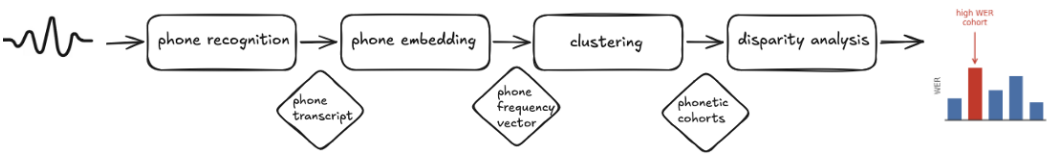

Figure 1: **Evaluation is fully automatic.** Wav files are sent for phonetic transcription and vectorization, producing a phonetic embedding. Clustering identifies speakers with similar phonetic embeddings, and disparity analysis identifies clusters with high word error rate, compared to other clusters.

2 Related Work

Bias in machine learning The field of algorithmic bias measures performance disparities that lead to negative outcomes for system users (Kearns & Roth, 2021). Prominent examples include loan approval, college admissions, and automated hiring (Barocas & Selbst, 2016). For these classification and regression models, bias is typically measured with respect to demographic labels. Martinez et al. (2021) provides a non-demographic method called Blind Pareto Fairness that can be used to measure and train fair classifiers. They point out that Pareto fair classifiers can be trained not only when group membership is unknown (e.g., racial identity of applicants), but also when we have no prior knowledge about which demographic groupings might be relevant to a particular classification problem (e.g., the list of all possible racial groups).

Slice discovery Slice discovery methods (SDMs) apply a nondemographic approach to generative models, primarily in computer vision. SDMs like Domino (Eyuboglu et al., 2022) use cross-modal embeddings and an error-aware mixture model to learn clusters of inputs that are poorly served by an image search or image generation model. Luo et al (Luo et al., 2023) also extend the non-demographic approach beyond classifiers, defining bias categories via counterfactual captions.

Disparity in speech recognition In speech recognition, specifically, work has focused on disparity in WER among speaker cohorts. Tatman (2017a) found that ASR systems perform differently for men and women, with Tatman (2017b) showing this disparity is partially attributable to non-demographic features like noise. Koenecke et al. (2020) and Chang et al. (2024) show that disparities persist in multiple commercial systems, even in current self-supervised ASR systems that provide better overall performance.

To identify bias cohorts, these works rely on demographic attributes (e.g., gender, race). However, it is not always possible to collect high quality demographic attributes alongside speech. Authors such as Kortmann et al. (2004) acknowledge that demographics have a complex relationship to speech behaviors, making inference of demographics from speech error-prone. Machine labeling for demographics also exhibits errors due to label discretization (Dong et al., 2024), and these difficulties extend to speaker recognition (Hutiri & Ding, 2022) as well as ASR (Martin & Tang, 2020; Tatman & Kasten, 2017; Dorn, 2019).

Prior work in speech recognition has produced hybrid demographic/data-driven pipelines, but not fully demographics-free evaluation. Fucci et al (Fucci et al., 2023) applied pitch features along with demographic proxy variables for weakness detection and synthetic re-training. Dheram, et al (Dheram et al., 2022) applied speaker embeddings to identify under-performing cohorts by features like pitch and power, rather than pronunciation. Phonetic speaker clusters were applied to dialect classification by (Nagy et al., 2006), supporting the hypothesis that phonetic clusters can reveal meaningful speech behaviors.

3 Methods

3.1 Data

We use two corpora of read speech, allowing us to control for variables of grammar and word choice. This means that differences in phone frequency among speakers are due to pronunciation, rather than differences in the words used by those speakers.

The TIMIT Complete Test Set Garofolo et al. (1993) is a corpus of spoken American English with 168 total speakers uttering 8 sentences each. Speakers self-identify with 8 regional dialects, 3 race values, and 2 gender values. We use the Complete Test Set, augmented with two additional "dialect" sentences per speaker, designed to reveal dialectical differences. The speakers in this augmented test set are completely disjoint with the speakers in the TIMIT Training Set, which appears in the training data sets of many ASR systems. The TIMIT dataset includes close phonetic transcriptions that were reviewed for quality by human linguists. We use these transcriptions in Section 4.5 to measure the accuracy of automatic phone recognizers.

We also test on a vendor-curated corpus of spoken American English that is larger and more balanced than TIMIT. The EN-US-structured dataset contains 2964 speakers uttering 35 sentences each. Speakers self-identify with 9 regional dialects, 3 gender identities, 4 race identities, and open-text responses that include identity intersections. This dataset has not been used as training data for any speech system, including the phone recognizers and commercial ASR model that we use for bias assessment.

3.2 ASR MODEL

We test our evaluation pipeline on the output of an off-the-shelf speech recognition service that relies on a multi-billion parameter speech foundation model. According to service documentation[1], the model is trained on millions of hours of unlabeled audio data from over 100 languages. This represents a state-of-the-art recognizer that performs well in aggregate[2], with reported WER of $< 1\%$ on native US English speech.

3.3 PHONE RECOGNITION

This work relies on the distinction between phonemes and phones. *Phonemes* are abstract sound units that are meaningful for a language (e.g. /p/ in "peak" vs /b/ in "beak"). *Phones* are the concrete sounds that are produced in a sample of speech (e.g. [p^h] vs [p] in "peak"). Different speakers may use different phones when pronouncing the same phoneme, due to regional accents or other dialect differences.

To generate an embedding for a new speech sample, we use a close phonetic transcription, which labels the speech with a string of phones. A sample is shown in Figure 2. Every transcription in the figure is unique, demonstrating how phonetic transcription can distinguish among different pronunciations of the same sentence. The transcripts in this figure come from the manually-annotated data in TIMIT. To produce a phonetic transcription automatically, we use machine-learned phone recognizers.

The Vitouphy phone recognizer Phy (2022) is a wav2vec model fine-tuned on the TIMIT training dataset. It has a vocabulary of 61 English-language phones. The Allosaurus phone recognizer Li et al. (2020) is a multilingual model with a customizable phone vocabulary of up to 263 phones[3].

In our experiments, we use two transcription settings: automatic transcripts from the Vitouphy English-language phone recognizer, denoted as *phone-en*; and transcripts from the multilingual Allosaurus model, denoted as *phone-mlt*. By testing both conditions, we can understand the robustness of the evaluation pipeline across different recognizer conditions. The *phone-en* condition applies when a language-specific phone recognizer is available in the language under consideration. The *phone-mlt* condition is more realistic for deployment, where ASR systems themselves are increasingly language-independent, and a customized phone recognizer may not be available.

3.4 PHONETIC EMBEDDINGS AND COHORTS

Phonetic cohorts are clusters in phonetic embedding space. To create these cohorts, we acquire phonetic transcripts from the recognizers described in Section 3.3. We build simple phonetic embedding vectors using the normalized counts of phones in each transcript. Next, we apply 95%

[1] https://aws.amazon.com/blogs/machine-learning/amazon-transcribe-announces-a-new-speech-foundation-model-powered-asr-system-that-expands-support-to-over-100-languages/

[2] https://docs.aws.amazon.com/ai/responsible-ai/transcribe-speech-recognition/overview.html

[3] In our experiments, we use the full IPA vocabulary of 263

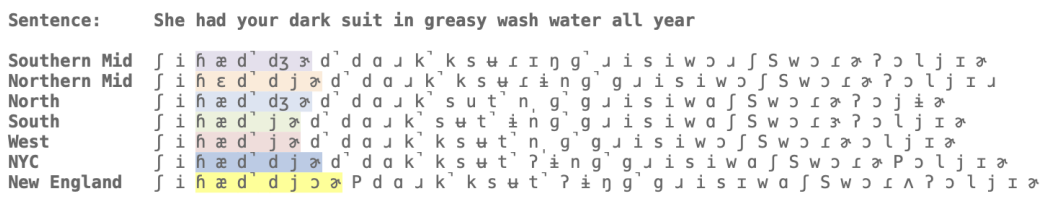

Figure 2: **Seven speakers pronounce the same sentence differently.** Close phonetic transcription from the TIMIT corpus, shown in IPA symbols. Phones for the phrase "had your" are highlighted, for comparison.

PCA to the vectors and build a Bayesian GMM with $k = 50$ components, pruning out centroids with weights below a constant threshold (0.015). After pruning, this yields a number of cohorts n between 23 and 28 on TIMIT, and 14 to 15 on EN-US-structured. Cohort size k ranges from 2 to 13 on TIMIT, and from 38 to 1546 on EN-US (see Table 3). Finally, we apply our pruned model to re-label the data. For each speaker, the model produces a distribution of membership likelihoods over all clusters; we label speakers with their highest-likelihood cluster ID. The resulting phonetic cohorts group speakers based on their distribution of phones.

Phones have the advantage of being the smallest contrastive unit of pronunciation. Speakers who differ in dialect or accent will differ in their phone distributions, such as presence of rhotic [r] or vowel mergers. For example, Wells (1982) describes phone-level variation as the primary marker of regional accent, confirmed by dialect atlases like WALS Online (Dryer & Haspelmath, 2013). As a result, phonetic embeddings capture variation that is linguistically meaningful and directly relevant to ASR difficulty.

3.5 Demographic cohorts

Our pipeline is demographics-free. However, we construct demographic cohorts as an experimental baseline. Demographic cohorts group speech samples based on a single attribute label (e.g. Race) or on intersections (e.g. Race-Gender). For TIMIT data, this yields a number of cohorts n between 3 and 5, depending on the attribute, and minimum cohort size k between 3 and 5. Unlike the learned phonetic cohorts, neither n nor the minimum cohort size k can be tuned or optimized. The nature of demographic labels is that they are hard-coded into the dataset at a level of granularity that is fixed by social norms.

In addition, demographic cohorts have the disadvantage of being noisy with respect to speech behavior. Speakers with identity features in common can still show systematic variations in pronunciation. Jones (2015) confirms that race-based communities may not be monolithic in their dialect features, while Rickford et al. (2015) provide additional data showing that dialect markers can vary significantly and systematically within a demographic class.

To illustrate, Table 1 shows two speakers from the EN-US-structured dataset who share the same race (White) and native language (English) but have very different phonetic profiles and very different WER. Their phonetic embeddings diverge most on the vowels [ʌ] (STRUT) and [ɪ] (KIT), the alveolar flap [ɾ], and the r-colored vowel [ɚ]. A demographic cohort groups these speakers together; phonetic cohorts separate them, exposing a nearly 4× difference in error rate.

Table 1: **Demographics do not map 1:1 onto speech.** Two speakers from the same race group (White) with divergent phonetic profiles and WER.

	Speaker A	**Speaker B**
Race	White	White
Native lang	English	English
Dialect	West	Inland South
WER	0.045	0.171

3.6 METRICS

Cluster Quality: We use cluster quality to measure whether our phonetic embeddings support good separation of data into coherent groups. We apply three clustering metrics: Davies Bouldin Index (Davies & Bouldin, 1979), Calinski Harabasz Index (Caliński & Harabasz, 1974), and the Silhouette Score (Shahapure & Nicholas, 2020). The Davies Bouldin Index (DBI) is the average similarity measure between each cluster and its most similar one, where similarity is the ratio of within-cluster distances to between-cluster distances, and a lower overall score indicates better clustering. The Calinski Harabasz Index (CHI) is the ratio of between-cluster variance to within-cluster variance, with a higher score indicating a better clustering result. The Silhouette Score (SIL) measures how well-separated clusters are by calculating the average silhouette coefficient for all data points, where a high value (ranging from -1 to 1) indicates that points are well-matched to their own cluster and poorly-matched to neighboring ones.

Disparity Detection: Evaluation methods are useful if they can go beyond mean WER to identify fine-grained disparities in system performance. We measure WER disparity for a single cohort as *cohort-vs-rest*, the difference in mean Word Error Rate (WER) over all data points within the cohort, compared to mean WER over all data points outside the cohort. Following Martinez et al. (2020), we apply a Rawlsian criterion (Rawls, 2001) for assessing embedding methods. The disparity value for the method is equal to the maximum disparity of any cohort found by the method.

To determine if these differences are statistically significant, we conduct Welch's ANOVA tests (Welch, 1951) to identify if significant pairwise differences exist between cohorts, and, when they do exist, measure the maximum effect size of the differences using Hedges g.

Each method for discovering cohorts (e.g., *phone-mlt* or *phone-en*) yields a clustering result C with individual cohorts $\{c_0...c_I\}$. Each c_i itself contains a set of speakers $S_i = \{s_0...s_J\}$ with an associated error measurement for each speaker, $err(s_j)$. Error for the cohort $err(c_i)$ is calculated as mean error across speakers. We apply ANOVA tests for the significance of pairwise differences in mean error between cohorts. For pairs that show significant differences, we calculate the effect size of the difference using Hedges g:

$$g(c_i, c_j) = J\frac{err(c_i) - err(c_j)}{RMSE(C)}$$

Where $RMSE(C)$ is the root mean squared error of the full cohort group and J is a correction factor for small samples ($J = 1 - \frac{3}{4(|c_i|+|c_j|)-9}$).

The maximum effect size for a cohort method like $C_{phone-mlt}$ is the max of $g(c_i, c_j)$ for all cohort pairs where significant differences were found.

Correlation with WER: We measure the salience of individual phonetic features to WER using Spearman rank correlation (Spearman, 1904), calculated between the speaker-level normalized count for each phone and speaker-level WER. Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995) at α=0.05 is applied to control for false positives over a large number of tests, one for every feature in the phone vocabulary.

The Spearman coefficient ρ ranges from -1 to $+1$; values above $|\rho|$=0.6 are conventionally considered strong, 0.4–0.6 moderate, and below 0.4 weak. Statistical significance (low p_{FDR}) and correlation strength ($|\rho|$) are distinct. With a large sample, even a weak correlation can be statistically significant.

Because demographic features are categorical rather than continuous, we measure salience to WER using eta-squared (η^2) from one-way ANOVA (Fisher, 1925). For each demographic variable (e.g. race, gender, education), η^2 gives the proportion of WER variance explained by group membership.

To compare the total impact of phonetic vs. demographic features, we fit two Ridge regression models (Hoerl, 2020). One model predicts speaker-level WER from all phonetic features jointly, a second model does the same across all demographic features[4]. We report 10-fold cross-validated R^2, to determine whether phonetic features explain more of the variance in WER than demographics.

[4]Demographic variables are one-hot encoded before fitting

4 RESULTS

4.1 CLUSTER QUALITY

Phonetic embeddings produce cohorts with meaningful internal structure. Table 2 reports our three standard clustering metrics for each dataset and phonetic condition. On TIMIT (168 speakers), both conditions yield positive CHI and SIL values, confirming that the clusters capture real structure rather than random partitions. On EN-US-structured (2,475 speakers), the larger sample produces stronger separation: CHI values reach 108–275 and SIL is positive for *phone-mlt* (0.087), though slightly negative for *phone-en* (-0.035), indicating some overlap between adjacent clusters in that condition.

Overall, the SIL values are modest. This is expected for high-dimensional, soft-boundary data. The DBI and CHI scores confirm that the clustering is not degenerate.

Table 2: **Phonetic cohorts have internal structure.** Cluster quality metrics for phonetic cohorts. Lower DBI is better; higher CHI and SIL are better.

Dataset	Condition	DBI	CHI	SIL
TIMIT	phone-en	2.80	3.19	0.0003
TIMIT	phone-mlt	1.78	6.59	0.1083
EN-US	phone-en	3.95	108.30	−0.0346
EN-US	phone-mlt	8.93	274.93	0.0866

4.2 DISPARITY DETECTION

Table 3: **Phonetic cohorts (gray) can produce larger effect sizes than demographic cohorts (white).** ANOVA results for pairwise disparity on two datasets, using phonetic and demographic cohort methods. Effect size is shown only for methods where ANOVA results are significant. See Section 3.6 for metric details.

TIMIT Dataset				
Method	**ANOVA p**	**Effect Size**	**N**	**Min k**
phone-mlt	4.48e-11	8.207	23	3
phone-en	7.77e-09	6.931	25	2
race	4.24e-01		3	5
race-gender	7.79e-01		5	3
EN-US-structured Dataset				
Method	**ANOVA p**	**Effect Size**	**N**	**Min k**
phone-mlt	6.03e-59	1.514	14	47
phone-en	2.28e-93	2.730	15	38
race	2.14e-14	0.704	6	9
race-gender	2.03e-09	1.533	16	2

In Figure 3, the disparity found by phonetic methods (blue) is generally larger than disparity found by demographic methods (brown). For the TIMIT dataset, both phonetic methods out-perform demographics.

To determine whether these differences are meaningful, we turn to ANOVA tests for statistical significance of the pairwise differences (Table 3). On TIMIT data, the differences observed using multilingual phonetic cohorts *phone-mlt* are significant and drive the largest effect size (8.207) of the two phonetic methods, beating English-language phonetic cohorts *phone-en* (6.931) Demographic methods do not reveal significant disparities on this dataset. This is unsurprising given that TIMIT's race and race-gender groups are severely imbalanced: the race groups contain 155, 8, and 5 speakers respectively, and race-gender subgroups have as few as 3 speakers.

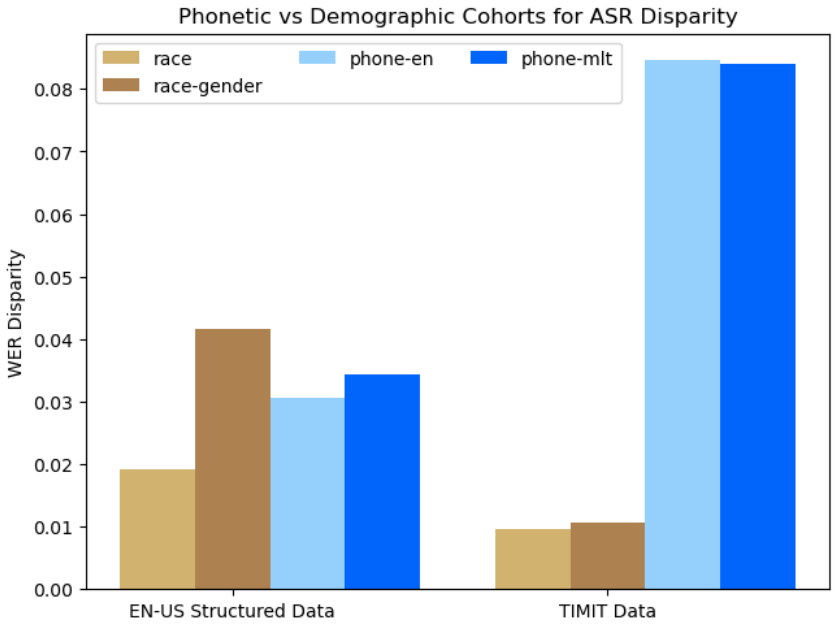

Figure 3: **Phonetic cohorts (blue) can reveal greater disparity than demographic cohorts (brown), particularly single-factor demographics (light brown)**

On the EN-US-structured dataset, phonetics perform as well or better than demographics. *phone-en* cohorts achieve the highest effect size (2.730) while *phone-mlt* (1.514) and *race-gender* (1.533) have similar scores. Cohorts identified by *race* do reveal significant disparities, but with small effect size.

4.3 Correlation with Error

Phonetic features are directly correlated with speaker-level WER in the EN-US-structured dataset. Based on Spearman coefficients for the *phone-mlt* condition, 114 features out of 229 (50%) are significantly correlated with WER. The five most predictive phones are shown in Table 4.

Table 4: **Individual phones have significant effect on WER.** Top 5 phones by Spearman $|\rho|$ with speaker-level WER on EN-US-structured (*phone-mlt* condition). Negative ρ means more of this phone predicts lower WER.

Phone	Description	ρ	p_{FDR}
iː	close front unrounded vowel (long)	$+0.452$	<0.001
ɒ	open back rounded vowel	$+0.403$	<0.001
t	voiceless alveolar stop	-0.400	<0.001
d	voiced alveolar stop	-0.361	<0.001
ɹ̩	syllabic alveolar approximant	-0.352	<0.001

Speakers who produce more long vowels [iː] and open back vowels [ɒ] tend to have higher WER, while speakers with more alveolar stops [t] and [d] and syllabic-r [ɹ̩] tend to have lower WER. The *phone-en* condition shows similar patterns at slightly lower magnitude, with 30 of 62 phones (48%) significant and top $|\rho|$=0.37.

Table 5 shows the total explanatory power of phonetic versus demographic features. Phonetic embeddings explain roughly five times more WER variance than all demographic variables combined (R^2=0.49 vs. 0.10), supporting the claim that phonetic features are directly correlated with drivers of error.

Table 5: **Phonetics drive more total error than demographics.** Cross-validated R^2 (10-fold, Ridge regression) predicting speaker-level WER on EN-US-structured. Phonetic features explain $\sim 5\times$ more variance than demographics.

Feature set	R^2_{CV}	Features
phone-en (vitouphy)	0.491 ± 0.066	62
phone-mlt (allosaurus)	0.415 ± 0.108	229
all demographics	0.102 ± 0.048	60
race-gender	0.090 ± 0.041	15
race + gender	0.080 ± 0.034	7
native language	0.022 ± 0.044	40

4.4 Interpretability

The phonetic profile of the highest-WER cohort on TIMIT (*phone-en*, mean WER = 0.077 vs. dataset mean 0.044) illustrates what the pipeline reveals in practice. This cohort shows reduced rhoticity: r-colored vowels appear at $0.72\times$ the dataset rate, and the alveolar flap at $0.79\times$, while the FOOT vowel is over-represented at $2.6\times$. This phonetic signature is consistent with Eastern New England speech, where r-dropping and distinct vowel realizations are well documented (Wells, 1982), and where most of the cohort's speakers originate. The cohort thus points an engineer directly at a phonetic register the system handles poorly, with no demographic labels required.

4.5 Robustness to Phone Recognition Quality

By measuring the error rate of both phone recognizers, we can assess the effect of phone recognition accuracy on the end to end pipeline. We perform this experiment on the TIMIT dataset, taking the linguist-supplied phonetic transcripts as ground truth and comparing them to the transcripts produced by our phone recognizers.

The Vitouphy phone recognizer (*phone-en*) matches linguist annotations perfectly (PER=0%). This recognizer was trained on the TIMIT training partition. Even though the speakers and sentence texts are disjoint in training and test data, the recognition accuracy benefits from a small vocabulary size and the controlled setting of TIMIT recordings (acoustics and IPA annotation conventions).

The Allosaurus phone recognizer (*phone-mlt*) achieves a phone error rate of 42% on the TIMIT test data. This accuracy is comparable to results on other English-language corpora such as Tedlium (Rousseau et al., 2012) and Switchboard (Godfrey et al., 1992) when retraining data is limited (Li et al., 2020). Even with low exact-match accuracy, the recognizer is able to spot enough distinguishing phones to uncover pronunciations that trigger error disparity.

Figure 3 shows that the *phone-mlt* cohorts, constructed with phonetic transcripts from Allosaurus, reveal similar absolute WER disparity to the cohorts constructed with Vitouphy transcripts, while both outperform demographics in most conditions. This suggests that the pipeline reveals disparity, even under realistic deployment scenarios where phone recognition will be imperfect.

5 Discussion and Limitations

This work demonstrates that phonetic embeddings can drive a meaningful evaluation of ASR beyond mean WER, identifying individual speech behaviors that may trigger error. The intended impact is to build momentum for disparity evaluation in ASR systems, given this technique to evaluate in a fully automatic way. Practitioners can seamlessly add disparity metrics to their existing evaluations, using their existing datasets.

Results in Section 4.3 show that phonetics are a powerful and overlooked representation for ASR evaluation. Phonetic features explain $\sim 49\%$ of WER variance, versus $\sim 10\%$ of variance explained by all demographics combined. This finding validates the choice of a data-driven embedding that aligns with the input to the ASR system itself.

This result holds up even under imperfect phone recognition conditions. Although the multilingual phone recognition rates described in Section 4.5 are low, the *phone-mlt* pipeline still recovered meaningful drivers of error in the realistic EN-US dataset. Our interpretation is that, with a larger phone vocabulary, the multilingual phone recognizer is able to spot rare error-triggering phones that were missed by the English-only recognizer. We will investigate this tension between phone recognition accuracy and phonetic vocabulary in future work.

Although we intentionally used a multilingual phone recognizer, this work is limited to English data. The feature representation is also limited, based on unigram phone frequencies. Future work using N-gram phone information or additional features like speaking rate may add to the interpretability and power of the clustering results. Experiments on multiple ASR models will also shed light on whether the results generalize well. However, we note that the method is not limited to ASR; any speech task with a well-defined error metric is in scope.

In spite of these limitations, the cumulative results show that even a simple phonetic embedding represents speech in a meaningful way for error analysis. The top Spearman phones are recognizable as characteristics of dialect speech (e.g., voiced/voiceless alveolar stops). Critically, our method identifies behaviors without assigning socially-constructed identities to the speakers themselves. This social context is also important. Linguistics literature shows us that social identities may influence, but do not dictate, speech patterns. This is borne out by the example in Table 1.

6 Conclusion

We propose a fully automatic method for evaluating speech systems at a granular level, providing actionable insight beyond mean WER and identifying speech behaviors that trigger error. Instead of subjective annotations or self-disclosure of demographic identity, this method relies on phonetic embeddings that we extract directly from the speech signal.

Phonetic embeddings reveal disparity as well as, or better than, demographic cohorts. They explain variance in WER at nearly five times the rate of demographic features. They can be learned even when the underlying phonetic transcript is noisy. The approach is scalable and aligned with features that are used by the speech system itself. High-error clusters are defined by interpretable phonetic profiles, providing a direct path to targeted data collection and model improvement. This technique brings meaningful disparity analysis into reach for any developer of an ASR system, using data already in their evaluation environment, and expanding the industry capability to understand fine-grained behaviors of speech systems.

AI use statement

In this work, we used generative AI tools to suggest a refactored outline for the paper, after a draft was written completely by hand. We created content for new sections by hand, and used generative AI to refine the language for a better fit with ICLR standards. We wrote a substantial related work section by hand, and submitted it for error-checking and gaps using AI tools. Experiments were designed by humans, with AI review and addition of some steps. We used AI to provide feedback on research methodology in the following way: code for phonetic embedding extraction, clustering, and disparity analysis was written manually by the authors, generating .csv files with raw WER results and pipeline-generated clustering and disparity results. We used AI to provide feedback on these results, to uncover additional findings and suggest next steps. We used AI to format jupyter output as LaTeX tables.

We have not used generative AI tools to generate datasets, develop conceptual frameworks, formulate or prove mathematical claims, or write proofs. We have reviewed all AI-assisted work. For example, we manually applied the refactored outline, to make sure information was not lost, mis-interpreted, or added by mistake. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

Although this work operates on human language, these studies did not involve human subjects directly. The speech datasets used here were either licensed explicitly for speech research purposes (TIMIT),

or collected locally with consent from participants for use in research and development of speech systems (en-us).

Reproducibility statement

The phone recognizers used in this paper are publicly available on github with links provided in the paper. Algorithms implemented by the authors are documented with citations for the algorithmic method (e.g. GMM clustering) with relevant parameters described in the Methods section, and additional detail provided in Appendix A. Reproducibility is limited for the EN-US dataset, which is proprietary. To mitigate this limitation we also test on the TIMIT dataset, which is available through licensing through the LDC Catalog (catalog.ldc.upenn.edu).

References

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 671, 2016. URL http://dx.doi.org/10.2139/ssrn.2477899.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3: 1–27, 1974. doi: 10.1080/03610927408827101.

Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R. Mortensen. Self-supervised speech representations still struggle with african american vernacular english. In *Interspeech 20224*, 2024.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. In *Interspeech 2022*, 2022.

Evan Dong, Aaron Schein, Yixin Wang, and Nikhil Garg. Addressing discretization-induced bias in demographic prediction. In *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2024. Association for Computing Machinery. URL https://arxiv.org/abs/2405.16762.

Rachel Dorn. Dialect-specific models for automatic speech recognition of African American Vernacular English. In Venelin Kovatchev, Irina Temnikova, Branislava Šandrih, and Ivelina Nikolova (eds.), *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pp. 16–20, Varna, Bulgaria, September 2019. INCOMA Ltd. doi: 10.26615/issn.2603-2821.2019_003. URL https://aclanthology.org/R19-2003/.

Matthew S. Dryer and Martin Haspelmath (eds.). *WALS Online (v2020.4)*. Zenodo, 2013. doi: 10.5281/zenodo.13950591. URL https://doi.org/10.5281/zenodo.13950591.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. No pitch left behind: Addressing gender unbalance in automatic speech recognition through pitch manipulation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023*, 2023. doi: 10.1109/ASRU57964.2023.10389767.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93, 1993.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pp. 517–520. IEEE, 1992.

Roger Hoerl. Ridge regression: A historical context. *Technometrics*, 62:420–425, 10 2020. doi: 10.1080/00401706.2020.1742207.

Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 230–247, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533089. URL https://doi.org/10.1145/3531146.3533089.

Taylor Jones. Toward a description of african american vernacular english dialect regions using "black twitter". *American Speech*, 90:403–440, 11 2015. doi: 10.1215/00031283-3442117.

Michael Kearns and Aaron Roth. The ethical algorithm: The science of socially aware algorithm design. *Perspectives on Science and Christian Faith*, 73:55–56, 03 2021. doi: 10.56315/PSCF3-21Kearns.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 2020. ISSN 10916490. doi: 10.1073/pnas.1915768117.

Bernd Kortmann, Kate Burridge, Rajend Mesthrie, Edgar W. Schnieder, and Clive Upton. *A Handbook of Varieties of English: A Multimedia Reference Tool.* De Gruyter, 12 2004. ISBN 9783110197181. doi: 10.1515/9783110197181. URL https://www.cur.ac.rw/mis/main/library/documents/book_file/digital-64749f6b7caec1.80729830.pdf. Volume 1: Phonology. Volume 2: Morphology and Syntax.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. Universal phone recognition with a multilingual allophone system. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2020-May, 2020. doi: 10.1109/ICASSP40776.2020.9054362.

Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De La Torre. Zero-shot model diagnosis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2023-June, 2023. doi: 10.1109/CVPR52729.2023.01119.

Joshua L. Martin and Kevin Tang. Understanding racial disparities in automatic speech recognition: The case of habitual ?be? In *Proc. Interspeech 2020*, volume 2020-October, pp. 626–630, 2020. doi: 10.21437/Interspeech.2020-2893.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, pp. 6755–6764. PMLR, 2020.

Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. Blind pareto fairness and subgroup robustness. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7492–7501, 7 2021. URL http://proceedings.mlr.press/v139/martinez21a/martinez21a.pdf.

Naomi Nagy, Xiaoli Zhang, George Nagy, and Edgar W Schneider. Clustering dialects automatically: A mutual information approach. *University of Pennsylvania Working Papers in Linguistics*, 12, 2006. ISSN 1524-9549, 1524-9549.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Vitou Phy. Automatic Phoneme Recognition on TIMIT Dataset with Wav2Vec 2.0, 2022. URL https://huggingface.co/vitouphy/wav2vec2-xls-r-300m-timit-phoneme. If you use this model, please cite it using these metadata.

John Rawls. *Justice as Fairness: A Restatement*. Harvard University Press, Cambridge, MA, 2001.

John R. Rickford, Greg J. Duncan, Lisa A. Gennetian, Ray Yun Gou, Rebecca Greene, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, Lisa Sanbonmatsu, Andres E. Sanchez-Ordoñez, Matthew Sciandra, Ewart Thomas, and Jens Ludwig. Neighborhood effects on use of african-american vernacular english. *Proceedings of the National Academy of Sciences*, 112(38):11817–11822, 2015. doi: 10.1073/pnas.1500176112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1500176112.

Anthony Rousseau, Paul Déléglise, and Yannick Esteve. TED-LIUM: An automatic speech recognition dedicated corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pp. 125–129, 2012.

Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747–748, 2020. doi: 10.1109/DSAA49011.2020.00096.

Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach (eds.), *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 53–59, Valencia, Spain, April 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-1606. URL https://aclanthology.org/W17-1606.

Rachael Tatman. How well do google and microsoft and recognize speech across dialect gender and race, 2017b. URL http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2024-08-17.

Rachael Tatman and Conner Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech 2017*, pp. 934–938, 2017. doi: 10.21437/Interspeech.2017-1746.

Raphael Vallat. Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31):1026, 2018.

B. L. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38 (3-4):330–336, 1951.

John C. Wells. *Accents of English. 3 Volumes.* Cambridge University Press, 1982. URL https://www.yorku.ca/earmstro/courses/phonetics/lexical_sets.pdf.

A Clustering Packages

Phonetic cohorts are constructed using scikit-learn Pedregosa et al. (2011). Speaker embeddings (normalized phone-count vectors) are standardized with `StandardScaler` and reduced to 95% explained variance using `PCA`. We then fit a `BayesianGaussianMixture` model with k=50 components and a weight concentration prior of 0.01. Components with weight below 0.015 are pruned, and the surviving weights are renormalized. The pruned means, covariances, and precision matrices are transferred to a standard `GaussianMixture` instance for final cluster assignment.

Cluster quality is evaluated with three scikit-learn metrics: `davies_bouldin_score`, `calinski_harabasz_score`, and `silhouette_score`. Statistical significance of pairwise WER differences across cohorts is assessed with Welch's ANOVA and Games-Howell post-hoc tests via the pingouin package Vallat (2018).