A Sociophonetic Analysis of Racial Bias in Commercial ASR Systems Using the Pacific Northwest English Corpus

Michael Scott, Siyu Liang, Alicia Wassink, Gina-Anne Levow
Department of Linguistics, University of Washington
michael.kelly.scott@gmail.com, {liangsy, wassink, levow}@uw.edu

Abstract

This paper presents a systematic evaluation of racial bias in four major commercial automatic speech recognition (ASR) systems using the Pacific Northwest English (PNWE) corpus. We analyze transcription accuracy across speakers from four ethnic backgrounds (African American, Caucasian American, ChicanX, and Yakama) and examine how sociophonetic variation contributes to differential system performance. We introduce a heuristically-determined Phonetic Error Rate (PER) metric that links recognition errors to specific linguistically motivated variables derived from sociophonetic annotation. Our analysis of eleven sociophonetic features reveals that vowel quality variation, particularly resistance to the low-back merger and pre-nasal merger patterns, is systematically associated with differential error rates across ethnic groups, with the most pronounced effects for African American speakers across all evaluated systems. These findings demonstrate that acoustic modeling of dialectal phonetic variation, rather than lexical or syntactic factors, remains a primary source of bias in commercial ASR systems. The study establishes the PNWE corpus as a valuable resource for bias evaluation in speech technologies and provides actionable guidance for improving ASR performance through targeted representation of sociophonetic diversity in training data.

Keywords: automatic speech recognition, evaluation, speech corpus, bias, sociophonetics, dialectal variation

1. Introduction

Automatic Speech Recognition (ASR) systems demonstrate significantly higher error rates for speakers from marginalized communities. Koenecke et al. (2020) found that commercial ASR systems exhibit average word error rates of 35% for African American speakers compared to 19% for Caucasian American speakers across five major commercial systems. This disparity represents a critical fairness issue as speech technologies become increasingly ubiquitous in everyday applications, from virtual assistants to automated closed captioning, medical dictation, and educational technology (Cunningham et al., 2024).

While these disparities are well-documented across multiple studies (Tatman, 2017; Tatman and Kasten, 2017; Martin and Wright, 2023), the underlying linguistic mechanisms remain underexplored. Wassink et al. (2022) provided crucial evidence that phonetic rather than lexical or syntactic factors drive these differences, showing that ASR systems exhibit higher error rates for African American English even when lexical content is controlled. This finding suggests that acoustic modeling, rather than language modeling, is the primary source of bias — a hypothesis further supported by subsequent investigations (Chang et al., 2024; Johnson et al., 2024).

Recent work has begun to address these disparities through various approaches. Dorn (2019) explored dialect-specific models and targeted training strategies, while others like Zhang et al. (2023); Wang et al. (2024); Li et al. (2024) have investigated how large-scale multilingual models and self-supervised learning systems handle dialectal variation. Studies using diverse corpora such as AfriSpeech-200 have demonstrated the potential to improve ASR performance across African accents through better data representation (Olatunji et al., 2023). However, these studies often focus on model architectures or training strategies rather than systematically identifying which specific phonetic features of the underrepresented dialect contribute to transcription errors. Understanding these linguistic mechanisms is essential for developing effective mitigation strategies.

This paper contributes a systematic evaluation methodology for measuring ASR bias using the Pacific Northwest English (PNWE) corpus (Wassink, 2007), focusing on eleven sociophonetic variables that exhibit variation across the ethnic communities represented in the corpus. We evaluate four major commercial ASR systems and demonstrate that specific phonetic features, particularly resistance to the low-back merger and variation in pre-nasal /æ/ patterns, are disproportionately associated with transcription errors for African American speakers. Our approach extends existing ASR evaluation frameworks by incorporating detailed sociolinguistic analysis, providing both traditional word error rates and a novel heuristically-determined phonetic error rate that links errors to specific phonological features documented in the sociolinguistic literature (Thomas, 2007; Green, 2002). By identifying the specific phonetic characteristics to evaluate current ASR systems, this work provides actionable guidance for improving system performance and

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4787–4796
11–16 May 2026.

emphasizes the PNWE corpus as a valuable resource for bias evaluation in speech technologies.

2. Related Work

2.1. Racial Disparities in Commercial ASR Systems

The work of Koenecke et al. (2020) demonstrated substantial racial disparities in commercial ASR systems, finding almost double average word error rates for African American speakers compared to Caucasian American speakers across five major systems. Tatman (2017); Tatman and Kasten (2017) further confirmed these disparities across diverse datasets and speech contexts. These errors have real-world impacts, with users often feeling "othered" by technology that fails to recognize their speech (Cunningham et al., 2024). However, most prior work had focused on documenting disparities rather than identifying specific linguistic mechanisms driving differential performance.

2.2. Linguistic Factors in ASR Performance Disparities

Wassink et al. (2022) provided crucial evidence that phonetic rather than lexical or syntactic factors drive performance differences for African American English speakers, showing that ASR systems exhibit higher error rates for African American speech even when lexical content is controlled. This finding clarifies that the disparity arises primarily from acoustic modeling rather than from language modeling, as text-based models are not the focus of this comparison. Taken together with the findings of Koenecke et al. (2020), this implicates acoustic modeling as the primary source of bias. Sociolinguistic research has extensively documented systematic phonetic variation in AAL (Thomas, 2007; Green, 2002), including variable low-back merger, pre-nasal /æ/ raising, /ay/ monophthongization, and consonant reduction – features that create mismatches with models trained on Mainstream American English.

Recent investigations of newer architectures reveal persistent challenges. Chang et al. (2024) showed that self-supervised speech representations continue to struggle with AAL, while Li et al. (2024) highlighted confounding factors like audio quality in performance assessments. Despite advances in large-scale multilingual systems (Zhang et al., 2023) and dialect-specific models (Dorn, 2019), few studies systematically connect specific sociophonetic variables to ASR errors. Our work addresses this gap through the quantitative analysis of eleven phonetic features across multiple commercial systems.

2.3. Resources for ASR Bias Evaluation

Evaluating ASR fairness requires corpora with demographically diverse speakers and controlled recording conditions. The Corpus of Regional African American Language (Kendall and Farrington, 2018) provides extensive AAL recordings, although the audio quality varies. Mozilla Common Voice (Ardila et al., 2020) includes demographic metadata and has recently begun adding spontaneous speech, though the release used here primarily contains read prompts that may miss naturalistic variation. The Speech Accent Archive (Weinberger, 2013) focuses on L2 English, while the Artie Bias Corpus (Meyer et al., 2020) has limited sample sizes. AfriSpeech-200 (Olatunji et al., 2023) offers 200 hours across 120 African accents but targets clinical domains rather than sociolinguistic variation.

The Pacific Northwest English corpus (Wassink, 2007) offers distinct advantages: balanced sampling across five ethnic groups within one region, sociolinguistic interviews eliciting multiple speech styles, consistent high-quality recordings, detailed sociophonetic annotation, and sufficient sample sizes for statistical analysis. Our work validates PNWE for bias evaluation while demonstrating how to connect performance disparities to specific linguistic features, advancing sociolinguistic approaches to fair speech technology (Johnson et al., 2024; Martin and Wright, 2023).

3. Data

3.1. The Pacific Northwest English Corpus

The **Pacific Northwest English (PNWE)** corpus (Wassink, 2007) was designed to document regional and ethnic variation in the Pacific Northwest of the United States, encompassing linguistic and social diversity across co-regional communities in Washington State. All speakers provided informed consent and were compensated for their participation.

The PNWE corpus includes multiple tasks designed to elicit different speech styles, including:

- **Word List (WL)** task: controlled production of lexical items embedded in the carrier phrase "write ___ today," targeting specific phonetic contrasts;
- **Lexical Game (LEX)** task: semi-controlled elicitation with variable speech rate and style;
- **Conversational Speech (CS)** task: open-ended sociolinguistic interviews.

The complete corpus comprises recordings from 112 speakers across five ethnic groups: African

American, Caucasian American, Asian American, ChicanX, and Yakama. All speakers were long-term residents of the Pacific Northwest, native English speakers, and self-identified members of their respective ethnic communities. The corpus includes 73 female and 39 male speakers, with ages ranging from 20 to 93 years. Table 1 presents the detailed demographic composition of the corpus. While the corpus includes balanced representation across ethnic groups for some communities, sample sizes vary, with Caucasian American speakers comprising the majority (n=68) and smaller samples from other groups, particularly African American speakers (n=5). This distribution reflects both the regional demographics of the Pacific Northwest and the challenges of recruiting participants from smaller ethnic communities in the region.

Each recording was transcribed using CLOx Wassink et al. (2020), a wrapper based on Microsoft's ASR system (Microsoft, 2022), with all transcripts subsequently hand-checked and manually segmented by trained linguists in Praat TextGrids (Boersma, 2007). The final annotations include orthographic and broad phonetic transcriptions. A custom extension of the CMU Pronouncing Dictionary (Carnegie Mellon University, 2015), termed PNWEdict, was created for the PNWE data to handle regionally distinctive variants. Metadata for each speaker include age, gender, ethnic identity, and speech task.

Ethnicity	Spkrs	F	M	Age
African Am.	5	3	2	35–72
Caucasian Am.	68	46	22	20–93
Asian Am.	19	10	9	20–93
ChicanX	12	10	2	27–52
Yakama	8	4	4	31–59
Total	112	73	39	20–93

Table 1: Demographic composition of the PNWE corpus.

3.2. Experiment Dataset

For our experiments, only the WL task was used due to its consistent lexical content and ease of phonetic alignment. Although the full PNWE corpus includes five ethnic groups, the Asian American subsample was excluded from this analysis because their error rates were nearly identical to those of Caucasian American speakers across all systems, offering minimal contrast for investigating bias; incorporating this group is planned for future work. The evaluation subset used in this study therefore includes recordings from 16 speakers across four ethnic groups: African American (AA, $n = 3$), Caucasian American (CA, $n = 4$), ChicanX (CX, $n = 4$), and Yakama (YA, $n = 5$). The unequal group sizes reflect the availability of speakers with complete word list TextGrid annotations at the time of the study. Speaker ages ranged from 27 to 75 years old (mean $\approx$ 53). Each speaker produced approximately 750 word list utterances embedded in carrier phrases; the AA and CA speakers read one elicitation set, while CX and YA speakers read a different set targeting partially overlapping phonetic contrasts. Although the groups are not perfectly balanced, holding regional origin and recording conditions constant supports controlled comparison across ethnic groups.

Ethnicity	Spkrs	F	M	Age
African Am.	3	2	1	35–72
Caucasian Am.	4	3	1	73–75
ChicanX	4	2	2	27–51
Yakama	5	2	3	55–59
Total	16	9	7	27–75

Table 2: Demographic composition of the PNWE evaluation subset in this study.

4. Methodology

4.1. Sociophonetic Variables

To provide a sociolinguistically informed evaluation of the speech models, eleven sociophonetic variables known to exhibit variation across ethnic and regional varieties of American English (Wassink et al., 2022) were targeted, listed in Table 3. These variables include features documented in African American Language (AAL), Chicano English, and Indigenous varieties, as well as features characteristic of Pacific Northwest regional dialects. These were used to test whether specific features co-occur with transcription errors across ASR systems.

In our test data, each sociolinguistic variable was annotated for contextual occurrence (presence/absence) and measured acoustically where applicable. The annotated TextGrids were then used to cross-reference ASR transcription errors, enabling correlation of linguistic variables with system performance.

4.2. ASR Systems Evaluated

We evaluated four major commercial ASR systems that represent the current state-of-the-art in speech recognition technology and are widely deployed in consumer and enterprise applications: Amazon Transcribe (Amazon, 2022), Google Cloud Speech-to-Text (Google, 2022), Apple Speech (Apple, 2022), and IBM Watson Speech to Text (IBM, 2022).

We used default configuration settings with US English language models and no custom vocabulary or adaptation to evaluate the systems as they

Feature	Code	Description	Realization	Example
Low-back merger resistance	-AO	/ɑ/–/ɔ/ distinction	Often maintained	*cot–caught*
Pre-nasal merger	IN	/ɪ/–/ɛ/ merger before nasals	Frequent	*pin–pen*
/aɪ/ monophthongization	AY	/aɪ/ glide reduction	[aː] before voiced coda	*time* → [taːm]
R-deletion	R	Post-vocalic /r/ weakening	[aə], [ʊ] outcomes	*car* → [kaə]
/θ/, /ð/ stopping	TH-s	Dental fricative → stop	[d], [t] substitutions	*this* → [dɪs]
/θ/, /ð/ fronting	TH-f	Dental fricative → labiodental	[f], [v] substitutions	*with* → [wɪf]
Consonant cluster reduction	CC	Final cluster simplification	Frequent	*test* → [tɛs]
Word-final devoicing	Dv	Voiced obstruent → voiceless	Stylistically variable	*had* → [hæt]
Word-final debuccalization	Db	/t/, /d/ → [ʔ] or ∅	Variable realization	*side* → [saɪʔ]
Pre-lateral back merger	prel-o	/ʊ/–/u/, /ʌ/–/o/ before /l/	Variable	*fool–full*
Pre-lateral front merger	prel-i	/ɪ/–/i/ merger before /l/	Variable	*feel–fill*

Table 3: The eleven sociophonetic variables annotated in the PNWE corpus.

would perform for typical users without specialized tuning.

It is important to note that commercial ASR systems are continuously updated, and performance characteristics may change over time. The results presented here reflect system performance at the time of evaluation in 2022 and should be interpreted as a snapshot of these systems' capabilities rather than permanent characterizations. However, the patterns of phonetic errors identified may remain relevant for understanding ongoing challenges in ASR performance across dialects.

4.3. Data Preparation

Approximately 4 hours of speech from the WL task were used in the primary analysis. Speech segments were exported as 16 kHz WAV files. The analysis excluded the LEX and CS tasks because they lacked consistent time-aligned phonetic annotation at the time of the study. However, we expect that extending the same methods to CS data (around 9 additional hours) would likely yield richer results, given the higher frequency of sociophonetic features in conversational speech.

4.4. Metrics

4.4.1. Word Error Rate (WER)

Word Error Rate (WER) is the standard metric for ASR evaluation, calculated as the edit distance between the hypothesis transcription and the reference transcription, normalized by the number of words in the reference:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcription.

We computed WER using the `nist sclite` toolkit (NIST, 2021), which performs optimal string alignment between hypothesis and reference transcriptions. The alignment process accounts for minor variations in word boundaries and punctuation while identifying genuine recognition errors.

WER provides an aggregate measure of transcription accuracy but does not distinguish between different types of errors or identify the linguistic sources of errors. To address this limitation, we complement WER analysis with an additional phonetic error rate metric.

4.4.2. Phonetic Error Rate (PER)

To investigate whether phonetic variation contributes to transcription errors, we developed an heuristically-determined Phonetic Error Rate (PER) metric. This metric compares the phonetic realization of what speakers actually said (from broad phonetic transcriptions in the given TextGrids) against the canonical phonetic representation of what the ASR system hypothesized (derived from PNWEdict).

The PER calculation involves the following steps:

Step 1: Extract Reference Phonetic Transcriptions

We extracted phonetic transcriptions from the Praat TextGrids provided in the PNWE corpus, which were manually created by trained linguists during corpus annotation. These transcriptions capture the actual phonetic realizations produced by speakers, including dialectal variations.

Step 2: Generate Hypothesis Phonetic Transcriptions

For each word in the ASR system's orthographic hypothesis, we looked up its canonical pronunciation in PNWEdict, a regionally adapted version of CMUdict providing ARPABET-encoded phonemic representations. Words not found in PNWEdict were generated using standard phonological rules for English inflections (e.g., possessives, plurals, gerunds).

To illustrate the alignment process, Table 4 presents three tokens from the PNWE corpus showing how the reference and hypothesis tiers were

matched at both the orthographic and phonetic levels. Manual phonetic transcriptions were extracted from the TextGrids, while heuristic phonetic hypotheses were generated by mapping the ASR output through PNWEdict. The alignment output identifies edits at the phone level which form the basis of the error rate calculation.

Step 3: Compute Phone-Level Edit Distance We used `sclite` to align the reference phonetic transcriptions with the heuristically-determined hypothesis phonetic transcriptions at the phone level, calculating:

$$\text{PER} = \frac{S_{\text{phone}} + D_{\text{phone}} + I_{\text{phone}}}{N_{\text{phone}}} \quad (2)$$

where S_{phone}, D_{phone}, and I_{phone} are phone-level substitutions, deletions, and insertions, and N_{phone} is the total number of phones in the reference transcription.

This heuristic PER metric has important limitations: it assumes the ASR system's word choice reflects phonetic confusions when in reality the mapping from acoustics to orthography in modern systems is opaque. Nevertheless, systematic differences in PER across demographic groups can indicate where phonetic variation may be contributing to recognition failures.

4.5. Statistical Analysis

We report descriptive statistics for WER and PER by system and ethnicity, with standard errors computed across speakers. To test for significant differences across ethnic groups while accounting for repeated measures (multiple systems evaluated on the same speakers), we fit linear mixed-effects models with speaker as a random intercept.

For marker-specific analyses, we compare error co-occurrence rates across ethnic groups using proportion tests. All statistical tests are two-sided with $p < 0.05$ considered significant. Given the number of comparisons across eleven markers and four systems, we do not apply a multiple comparisons correction; results near the threshold should be treated as exploratory findings motivating further investigation rather than conclusive evidence.

4.6. Error Analysis

Beyond aggregate metrics, we conducted qualitative analysis of error patterns through manual examination of transcription outputs. We randomly sampled errors that co-occurred with targeted sociophonetic markers and categorized them by error type (substitution, deletion, insertion) and apparent phonetic motivation. We examined spectrograms and listened to audio for a subset of 200 errors to validate that the hypothesized phonetic features were indeed present in the acoustic signal. This manual validation found that approximately 85% of errors classified as phonetically-motivated showed acoustic evidence consistent with the targeted feature. The remaining 15% represent misclassifications that introduce additional uncertainty into co-occurrence counts; accordingly, single-digit co-occurrence values in Table 7 should be interpreted with caution and treated as exploratory rather than confirmatory.

5. Results

5.1. Overall Performance Disparities

Table 5 presents word error rates across the four ASR systems for each ethnic group in the PNWE corpus. Consistent with prior work (Koenecke et al., 2020), we observe significant performance disparities across ethnic groups, with African American, ChicanX, and Yakama speakers all experiencing higher error rates than Caucasian American speakers across most systems.

African American speakers experienced mean WER of 20% compared to 15% for Caucasian American speakers (33% relative increase). Mixed-effects models with speaker as random effect confirmed these differences were highly significant ($p < 0.001$). ChicanX speakers showed the highest mean error rate at 23%, while Yakama speakers experienced a mean WER of 21%, higher than Caucasian American but with less consistent patterns than African American disparities. Notably, while Caucasian American speakers consistently achieved the lowest error rates across all systems, the other three ethnic groups all experienced elevated error rates, though with different patterns of which systems performed worst for each group. Crucially, these performance gaps remained consistent across systems, indicating systematic rather than system-specific biases.

5.2. Phonetic vs. Orthographic Error Rates

Computing PER with PNWEdict yielded uniformly lower error rates than standard WER, with mean reductions of 42% for Apple, 48% for Amazon, 47% for Google, and 54% for IBM across all speakers. This difference reflects the change in evaluation granularity rather than an actual decrease in recognition mistakes: because PER operates at the phonetic level, it assigns partial credit to near-misses (e.g., "into da" vs."in today") and removes word-boundary penalties. The consistently lower PER nevertheless indicates that many apparent word errors are phonetically systematic rather than random lexical failures, supporting the hypothesis that acoustic modeling mismatches underlie performance disparities.

Word	Ref. (IPA)	ASR Output	PNWEdict (ARPABET)	Phone Alignment	Note
caught	[kɔːt]	*cot*	K AA T	K|K, AO→AA (S), T|T	Low-back merger (−AO)
pen	[pɪn]	*pin*	P IH N	P|P, IH→EH (S), N|N	Pre-nasal merger (IN)
test	[tɛst]	*tess*	T EH S	T|T, EH|EH, S|S, T→ ∅ (D)	Cluster reduction (CC)

Table 4: Example alignment between manual phonetic transcriptions (reference) and heuristic phonetic hypotheses (ASR output via PNWEdict).

System	AA	CA	CX	YA
Apple	24%	14%	24%	24%
Amazon	9%	7%	13%	12%
Google	26%	16%	21%	23%
IBM	21%	21%	35%	23%
Mean	20%	15%	23%	21%

Table 5: Word Error Rates by ethnicity and system. AA=African American, CA=Caucasian American, CX=ChicanX, YA=Yakama.

Ethnicity	Spkrs	−AO	CC	IN
African Am.	3	4.33	8.00	5.33
Caucasian Am.	4	4.00	8.75	3.50
ChicanX	4	1.75	5.25	3.00
Yakama	5	0.00	5.60	3.00

Table 6: Mean instances per speaker of realized targeted markers by ethnicity.

Using PNWEdict rather than CMUdict for canonical pronunciations further reduced PER by 2–3% on average, demonstrating the value of region-specific pronunciation dictionaries. The PNWE-augmented dictionary better captures maintained vowel distinctions and regional variants, reducing false positives in phonetic error attribution.

5.3. Sociophonetic Feature Associations

Although eleven sociophonetic variables were targeted (Table 3), only three—(−AO), (CC), and (IN)—were realized frequently enough to analyze in context with transcription errors. A fourth marker, word-final debuccalization (Db), was realized by only two speakers (one Caucasian American and one Yakama), each with a single instance and no error co-occurrence. The remaining seven markers were not realized by any speaker in the word list data, likely because the controlled word list task elicits more careful speech than conversational contexts where these features are more common.

Table 6 summarizes the realization of the three analyzable markers across ethnic groups. African American speakers realized (IN) at 5.33 instances per speaker compared to 3.5 for Caucasian American, 3 for ChicanX, and 3 for Yakama speakers. Similarly, African American speakers realized (−AO) more frequently (4.33 instances per speaker) than Caucasian American (4.0) and ChicanX (1.75) speakers, while Yakama speakers showed no (−AO) realizations. Consonant cluster reduction (CC) showed more balanced distribution, with African American speakers at 8 instances per speaker versus 8.75 for Caucasian American, 5.25 for ChicanX, and 5.6 for Yakama.

Table 7 presents raw counts of errors co-occurring with targeted sociophonetic markers across systems and ethnic groups. Two vowel-driven patterns emerge consistently. First, (−AO) (low-back merger resistance) shows higher error overlap for African American speakers than Caucasian American speakers across most systems (Apple: 6 vs. 5; Google: 5 vs. 1; IBM: 6 vs. 3). ChicanX speakers also show elevated (−AO) overlap (5–7 across systems) while Yakama speakers show zero overlap, consistent with complete merger in that group.

Second, (IN) (pre-nasal /ɪ/–/ɛ/ merger) overlaps more frequently with errors for African American speakers (4 instances for Apple and Google, 2 for Amazon and IBM) than Caucasian American speakers (1 instance each for Apple and Google, 0 for Amazon, 3 for IBM). ChicanX speakers showed no (IN) error overlap, while Yakama speakers showed 2 instances across all systems.

Consonant cluster reduction (CC) shows elevated overlap for Yakama speakers (7–9 instances across systems) while remaining more evenly distributed across other groups (1–5 instances). This suggests consonantal variation contributes to errors but less differentially by ethnicity than vowel features.

When normalized by the number of possible contexts where each marker could occur, these patterns persist. African American speakers showed (−AO) error rates of 5.8% (Apple), 2.1% (Amazon), 5.2% (Google), and 7.9% (IBM) compared to Caucasian American rates of 5.2%, 0%, 1.1%, 3.2%, and 2.2% respectively. For (IN), African American normalized error rates were 2.8%, 2.1%, 2.3%, 1.6%, and 2.4% compared to Caucasian American rates of 1.0%, 0%, 0.7%, 3.1%, and 2.3%.

These realization differences, combined with the error overlaps, indicate that vowel quality variation, particularly resistance to low-back merger and pre-nasal merger patterns, drives disproportionate transcription failures for speakers who maintain these distinctions, most notably African American speakers in our corpus.

	Apple				Amazon				Google				IBM			
Group	-AO	CC	IN	*Err*	-AO	CC	IN	*Err*	-AO	CC	IN	*Err*	-AO	CC	IN	*Err*
AA	6	3	4	1854	1	0	2	681	5	2	4	2075	6	1	2	1695
CA	5	4	1	1218	0	1	0	593	1	5	1	1340	3	2	3	1800
CX	5	4	0	2533	6	0	0	1361	6	0	0	2267	5	6	0	3889
YA	0	8	2	2753	0	9	2	1419	0	7	2	2623	0	7	2	2735

Table 7: Error overlap with targeted markers by system and ethnicity. -AO=low-back merger resistance, CC=consonant cluster reduction, IN=pre-nasal merger. *Err*=total error count

5.4. System-Specific Patterns

The ASR systems evaluated differ in sensitivity to targeted variables. Amazon exhibits the lowest overall PER and shows smallest error increases in (-AO) and (IN) contexts for African American speakers. Apple shows competitive PER with relatively balanced treatment across features. Google's error distribution is more uneven, with pronounced sensitivity to (-AO) contexts. IBM systems show the highest sensitivity to vowel quality variation, especially in pre-nasal contexts.

These differences point to distinct training distributions and accent coverage, with diverse vowel realizations from speech varieties remaining underrepresented across all systems despite architectural differences.

5.5. Example Error Patterns

Manual inspection of 200 randomly sampled errors revealed systematic patterns consistent with targeted phonetic features:

Low-back merger: Words with maintained /ɔ/ were consistently transcribed with /ɑ/ spellings: *caught* → "cot" (6 instances across African American speakers, 5-7 across ChicanX speakers), *taught* → "tot", *thought* → "that".

Pre-nasal merger: Words with /ɪ/ before nasals showed substitution or deletion: *when* → "and" (3 instances), *pin* → "pen" (2 instances).

Cluster reduction: Final clusters misrecognized as simpler forms: *next* → "neck" (4 instances across systems), *test* → "tess" (2 instances).

Approximately 85% of errors classified as phonetically-motivated showed acoustic patterns consistent with the hypothesized feature when examined spectrogrammatically, validating the PER methodology.

6. Discussion

6.1. Phonetic Variation as Primary Driver of ASR Bias

Our results provide suggestive evidence consistent with phonetic rather than just lexical or syntactic variation contributing to ASR performance disparities across ethnic groups, particularly for African American speakers, as hypothesized by Wassink et al. (2022). The systematic correlation between specific phonetic features (particularly low-back merger resistance and pre-nasal merger patterns) and transcription errors demonstrates that current ASR systems struggle with systematic phonetic variation documented across multiple ethnic varieties of American English (Thomas, 2007; Green, 2002).

The phonetic features most strongly associated with errors involve vowel quality differences, which acoustic models rely on heavily for phoneme discrimination. When training data predominantly represents merged low-back systems, acoustic models learn probability distributions that inadequately cover the acoustic space of unmerged varieties. Speakers maintaining the /ɑ/–/ɔ/ distinction produce these vowels with distinct F1 and F2 formant patterns that fall outside the learned distributions of merged-dialect models, leading to systematic misrecognition.

The consistency of these patterns across all four evaluated commercial systems suggests systematic underrepresentation of diverse dialectal phonetic features in training data rather than system-specific architectural limitations. This finding aligns with recent work showing that even self-supervised models struggle with non-mainstream dialects including AAL (Chang et al., 2024), indicating that scale alone does not resolve representational gaps.

6.2. Implications for ASR System Development

Our findings have direct implications for developing more equitable ASR systems. Rather than simply increasing data volume, developers should prioritize sociolinguistic diversity in training corpora. The identification of specific problematic features, such as low-back merger resistance and pre-nasal patterns, provides actionable targets for data collection efforts. Training sets should include sufficient examples of each phonetic variant to enable robust acoustic modeling across the full range of systematic variation.

Beyond data collection, architectural innovations could explicitly model phonetic variation. Multi-dialect acoustic models could maintain separate probability distributions for different phonetic re-

alizations rather than treating variation as noise. Recent work on phoneme-based contextualization (Hu et al., 2019) and dialect-specific models (Dorn, 2019) demonstrates promising directions, though our results suggest that commercial systems have not yet incorporated such approaches effectively.

6.3. Methodological Contributions

This study introduces a linguistically grounded evaluation framework that links ASR performance to actual phonetic realization. The proposed Phonetic Error Rate (PER) moves beyond aggregate accuracy metrics by aligning ASR hypotheses with manual phonetic transcriptions, enabling interpretation of errors in relation to sociophonetic variables rather than abstract symbol mismatches. This approach reframes ASR evaluation as an analysis of how models represent and generalize across systematic sound variation, rather than how closely they reproduce orthographic sequences.

By incorporating a region-specific pronunciation dictionary, the method operationalizes sociophonetic knowledge such as vowel mergers, cluster reduction, and regional variants, into quantitative evaluation. Instead of treating dialectal variation as noise, the framework makes it measurable, allowing identification of errors that reflect genuine acoustic divergence across speech communities.

6.4. PNWE Corpus as an Evaluation Resource

This study validates the PNWE corpus as a valuable resource for ASR bias evaluation. Compared to other available corpora, the PNWE corpus is characterized by demographic diversity, regional representation, consistent recording quality, and rich linguistic annotation. While CORAAL (Kendall and Farrington, 2018) provides extensive AAL data, it lacks comparable samples from other ethnic groups and has variable audio quality. Common Voice (Ardila et al., 2020) includes demographic metadata but uses read speech that may not capture naturalistic phonetic variation. The PNWE corpus' word list task data subset also provides useful controlled data for evaluation.

6.5. Limitations and Future Directions

Several limitations should be acknowledged. First, our sample size of 16 speakers limits statistical power for intersectional analyses examining interactions between ethnicity, gender, and age; the unequal group sizes (3 AA, 4 CA, 4 CX, 5 YA) further constrain group-level comparisons. Second, the focus on the word list task, while enabling controlled comparison, may underestimate errors in spontaneous speech where sociophonetic features are more frequent and variable. Extension to the conversational speech portions of PNWE would also provide a more comprehensive assessment. Third, the manual validation of PER relied on a single annotator inspecting spectrograms; future work should establish inter-annotator agreement and confidence intervals for the 85% validation rate. Fourth, our evaluation was restricted to four commercial systems using default US English models; comparison with open-source systems such as Whisper or dialect-adapted models would clarify whether the same sociophonetic error patterns persist in systems designed to handle greater phonetic diversity.

The regional specificity of PNWE also limits generalizability to other English varieties. For example, regional variation in AAL is well-documented (Wolfram, 1969; Thomas, 2007), and the features most salient in the Pacific Northwest may differ from those in other regions. Replication of the experiments using corpora from other regions would complement and strengthen our conclusions about systematic ASR bias.

7. Conclusion

This study provides evidence that systematic phonetic variation contributes to performance disparities in commercial ASR systems across multiple ethnic groups. Through controlled evaluation using the Pacific Northwest English corpus, we identified specific phonetic features, particularly resistance to the low-back merger and pre-nasal merger patterns, that consistently correlate with transcription errors across ethnic groups, with the most pronounced effects observed for African American speakers across four major commercial systems. The substantial reduction in error rates when using phonetic versus orthographic evaluation (42–54% across systems) confirms that acoustic modeling, rather than language modeling, is the primary locus of bias.

Our heuristically-determined Phonetic Error Rate metric provides a practical framework for connecting ASR errors to specific sociophonetic features, enabling more targeted evaluation than aggregate word error rates alone. The 85% validation rate through acoustic analysis demonstrates that this approach successfully identifies systematic phonetic patterns underlying recognition failures. This methodology can be extended to other dialects and languages where detailed phonetic annotation is available.

The consistency of bias patterns across all evaluated systems—despite differences in architecture and training approaches—points to systematic underrepresentation of diverse phonetic features in training data rather than system-specific limitations.

This finding has direct implications for ASR development: improving performance for marginalized speech communities requires not just more data, but strategically diverse data that adequately represents the full range of systematic phonetic variation documented in sociolinguistic research.

The PNWE corpus, with its balanced sampling across ethnic groups, controlled recording conditions, and detailed sociophonetic annotation, provides a valuable resource for continued bias evaluation in speech technologies. Future work should extend this analysis to conversational speech data, examine additional sociophonetic features, and replicate these methods with corpora representing other regional varieties and marginalized dialects. Only through systematic documentation of which phonetic features drive ASR failures can we develop effective strategies for building truly equitable speech recognition systems.

8. Bibliographical References

Amazon. 2022. Amazon Transcribe.

Apple. 2022. Apple Speech API.

Paul Boersma. 2007. Praat: doing phonetics by computer. *http://www. praat. org/*.

Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R. Mortensen. 2024. Self-supervised Speech Representations Still Struggle with African American Vernacular English. ArXiv:2408.14262 [cs].

Jay Cunningham, Su Lin Blodgett, Michael Madaio, Hal Daumé Iii, Christina Harrington, and Hanna Wallach. 2024. Understanding the Impacts of Language Technologies' Performance Disparities on African American Language Speakers. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12826–12833, Bangkok, Thailand. Association for Computational Linguistics.

Rachel Dorn. 2019. Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 16–20, Varna, Bulgaria. INCOMA Ltd.

Google. 2022. Google Cloud Speech-to-Text.

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Ke Hu, Antoine Bruguier, Tara N. Sainath, Rohit Prabhavalkar, and Golan Pundak. 2019. Phoneme-Based Contextualization for Cross-Lingual Speech Recognition in End-to-End Models. ArXiv:1906.09292 [cs].

IBM. 2022. IBM Watson Speech to Text.

Alexander Johnson, Natarajan Balaji Shankar, Mari Ostendorf, and Abeer Alwan. 2024. An exploratory study on dialect density estimation for children and adult's African American English. *The Journal of the Acoustical Society of America*, 155(4):2836–2848.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689. Publisher: Proceedings of the National Academy of Sciences.

Changye Li, Trevor Cohen, and Serguei Pakhomov. 2024. Reexamining Racial Disparities in Automatic Speech Recognition Performance: The Role of Confounding by Provenance. ArXiv:2407.13982 [cs].

Joshua L Martin and Kelly Elizabeth Wright. 2023. Bias in Automatic Speech Recognition: The Case of African American Language. *Applied Linguistics*, 44(4):613–630.

Microsoft. 2022. Azure AI Speech.

NIST. 2021. SCTK, the NIST scoring toolkit. Original-date: 2016-05-03T19:00:47Z.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. pages 934–938.

Erik R. Thomas. 2007. Phonological and Phonetic Characteristics of African American Vernacular English. *Language and Linguistics Compass*, 1(5):450–475.

Siyin Wang, Chao-Han Huck Yang, Ji Wu, and Chao Zhang. 2024. Can Whisper perform speech-based in-context learning? ArXiv:2309.07081 [eess] version: 2.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

Alicia Beckford Wassink, Rob Squizzero, Campion Fellin, and David Nichols. 2020. CLOx.

Walter A. Wolfram. 1969. A Sociolinguistic Description of Detroit Negro Speech. Urban Language Series, No. 5. Technical Report 5, Publications Section, Center for Applied Linguistics, 1717 Massachusetts Avenue, N. ERIC Number: ED028431.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. ArXiv:2303.01037 [cs].

9. Language Resource References

Ardila, Rosana and Branson, Megan and Davis, Kelly and Henretty, Michael and Kohler, Michael and Meyer, Josh and Morais, Reuben and Saunders, Lindsay and Tyers, Francis M. and Weber, Gregor. 2020. *Common Voice: A Massively-Multilingual Speech Corpus*. Mozilla. Mozilla Foundation. Available at https://commonvoice.mozilla.org/.

Carnegie Mellon University. 2015. *The Carnegie Mellon Pronouncing Dictionary (CMUdict)*. Carnegie Mellon University. Carnegie Mellon University, 0.7b. Available at https://github.com/cmusphinx/cmudict.

Kendall, Tyler and Farrington, Charlie. 2018. *Corpus of Regional African American Language*. University of Oregon. The Online Resources for African American Language Project. Available at https://oraal.uoregon.edu/coraal/.

Meyer, Josh and Rauchenstein, Lindy and Eisenberg, Joshua D. and Howell, Nicholas. 2020. *Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications*. Artie Inc. Artie. Available at https://github.com/artie-inc/artie-bias-corpus.

Olatunji, Tobi and Afonja, Tejumade and Yadavalli, Aditya and Emezue, Chris Chinenye and Singh, Sahib and Dossou, Bonaventure F. P. and Osuchukwu, Joanne and Osei, Salomey and Tonja, Atnafu Lambebo and Etori, Naome and Mbataku, Clinton. 2023. *AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR*. Intron Health and collaborating institutions. Intron Health. 200 hours of Pan-African English speech from 120 indigenous accents across 13 countries.

Wassink, Alicia Beckford. 2007. *Pacific Northwest English Corpus*. University of Washington Sociolinguistics Laboratory. University of Washington, Department of Linguistics. Funded by the National Science Foundation. Available at https://zeos.ling.washington.edu/PNWEnglish/index.php.

Weinberger, Steven. 2013. *Speech Accent Archive*. George Mason University. George Mason University. Available at http://accent.gmu.edu/.