Evaluating Bias in Phoneme-Based Automatic Speech Recognition Systems: An Analysis of IPA Transcription Models

Maneesha Rani Saha *
University of Utah
maneesha.saha@utah.edu

Catherine Bao *
University of Utah
u1459030@utah.edu

Neal Patwari
University of Utah
neal.patwari@utah.edu

Abstract

As automatic speech recognition (ASR) systems shift toward multilingual support and low-resource language modeling, phoneme-based layers serve as a critical language-agnostic foundation. However, most evaluations of ASR's demographic biases related to race, age, gender, and accent focus on standard grapheme-based ASR systems with comparatively little emphasis on phoneme-based systems. In this study, we evaluate the performance of WhisperIPA and ZIPA, two state-of-the-art open-source systems that generate International Phonetic Alphabet (IPA) transcriptions. Our evaluation includes existing multilingual speech corpora and demographically annotated English-language corpora, comparing model-generated IPA transcriptions against grapheme-to-phoneme (G2P) systems using both standard phoneme error rate (PER) and a proposed Soft PER metric that tolerates linguistically similar phoneme substitutions. Our analysis examines how performance varies across language, gender, accent, ethnicity, and age, revealing persistent disparities even after accounting for acceptable phonemic variation. These findings, while limited, provide insight into potential sources of bias and inform the development of more inclusive and linguistically robust phoneme-based ASR systems. Our code and data are publicly available.[1]

1 Introduction

Automatic Speech Recognition (ASR) systems are a foundational component of modern agent-based technology. Current ASR systems typically rely on grapheme representations: graphemes are written units of language (e.g., letters or characters), while phonemes are units of sound that represent how words are pronounced. Existing research shows that grapheme-based ASR systems exhibit systematic disparities across demographic groups, including race, gender, age, accent, and language background (Tatman and Kasten, 2017; Tatman, 2017; DiChristofano et al., 2023; Jahan et al., 2025; El-Ghazaly et al., 2025; Scott et al., 2025; Serditova et al., 2025; Cunningham et al., 2025). ASR evaluation currently mostly focuses on word error rates (WER) (Feng et al., 2021).

While a majority of industry models remain grapheme-based, there has been a growing trend toward integrating IPA layers for multilingual capability (Li et al., 2020; Xu et al., 2021). The International Phonetic Alphabet (IPA) captures fine-grained phonetic details and cross-linguistic sound variations. Multilingual training on IPA tokens reduces phonetic token error rate by 60–70% in high-resource languages and 14–42% in low-resource languages (Żelasko et al., 2022), and including IPA layers enables more effective generalization by providing a more direct mapping between the acoustic signal and its phonological categories. This makes recent work in phonetic (Bharadwaj et al., 2026; Goriely and Buttery, 2025) and IPA models (Fang et al., 2020; Taguchi et al., 2023; Zhu et al., 2024; Lee et al., 2025; Zhu et al., 2025), integrated as standalone or intermediate layers, valuable for multilingual and low-resource ASR systems.

Despite this growing importance, bias research has focused almost exclusively on grapheme-based, WER-measured systems; comparatively little has been done on phoneme representations. This gap matters in both directions: if disparities are already present at the phoneme level, they may propagate through downstream components of integrated ASR systems; conversely, if phonetic representations prove more robust, they may lead to more equitable speech recognition. Phonemic-level demographic bias thus remains largely underexplored, motivating the present study.

Since the same phoneme can be realized differ-

*These authors contributed equally.

[1]https://github.com/Maneesha28/ipa-asr-transcription-bias-evaluation

ently across accents, dialects, and languages, strict Phonetic Error Rate (PER) may penalize valid pronunciation variation, introducing noise when evaluating demographic disparities. To reduce this effect, we introduce a soft variant of PER that tolerates minor phonemic variation. We evaluate how this PER variant changes phoneme ASR bias analysis. Our contributions include:

1. Introducing and evaluating a Soft PER that tolerates minor phonemic variation rather than strictly penalizing all substitutions. We compare Soft PER with standard PER to determine whether it yields a more robust metric for evaluating demographic bias.

2. Evaluating the biases of IPA-based phoneme ASR models across accent, gender, age, ethnicity and language. Our evaluation spans 11 languages (English, Hindi, Bangla, Panjabi, Tamil, Telugu, Arabic, French, Spanish, Turkish, and Shona). We also analyze different age groups, gender, four ethnicity groups, and English accents across eight accent/dialect groups.

2 Related Work

End-to-end ASR systems have increased accuracy by scaling data and model size, especially with transformer-family architectures (Radford et al., 2023). Large models such as wav2vec 2.0 (Baevski et al., 2020), Whisper (Radford et al., 2023), and XLS-r (Babu et al., 2022) demonstrate that grapheme-based models can generalize to dozens of languages. Recent work focuses on even larger multilingual models that replace language-specific systems and target lower-resource languages (Meng et al., 2025; Omnilingual ASR team et al., 2025; Ramirez et al., 2024).

Despite these advances, ASR is concentrated in high-resource languages with standardized orthographic transcriptions (Gale et al., 2023; Gorisch and Schmidt, 2024). Grapheme-based models rely on widely adopted writing systems in paired speech-text data. For languages and conversation without a formal writing system, non-standardized orthography, limited textual resources, or code-switching, grapheme models are less reliable or unavailable. Most grapheme mappings do not consider phonetic variation (Fang et al., 2020; Omnilingual ASR team et al., 2025; Ramirez et al., 2024).

By mapping acoustically similar sounds to shared phonetic symbols, these systems can transfer knowledge from high-resource to truly under-resourced languages (Daul et al., 2026). Currently, architectures introduce an explicit phonetic layer that improves robustness and data efficiency in low-resource settings (Lee et al., 2025). Incorporating phonemic supervision can reduce the ultimate WER (Fu et al., 2025; Scott et al., 2025).

Phonemic models are particularly advantageous for orthographic limitations or linguistic variability. Phonemic/phonetic token models outperform orthographic baselines in both data-constrained and cross-lingual domains (Daul et al., 2026). Shared phonetics allow easier cross-lingual transfer and generalize underrepresented languages (Goriely and Buttery, 2025; Zhu et al., 2025; Devi and Sarma, 2026). Phonemic modeling directly captures acoustic distinctions, thus improving robustness to pronunciation variability (Lyu and Lyu, 2008; Solorio and Liu, 2008; Onda et al., 2026).

Rather than replacing grapheme-based systems, phonetic layers are often integrated within them. Multi-task learning frameworks jointly predict phonemes and graphemes. Hybrid approaches have demonstrated improved robustness in noisy, low-resource, and multilingual settings, suggesting that phonetic supervision can complement larger ASR architectures (Fang et al., 2020).

Bias in ASR systems has been extensively studied in grapheme-based systems, with prior research showing significant disparities in WER across demographic groups — such as race, gender, dialect, and accent (Tatman and Kasten, 2017; DiChristofano et al., 2023; Jahan et al., 2025). An ASR system exhibits bias when its performance differs by user demographics (Zhang et al., 2023). Within grapheme-based models, commercial systems consistently have higher error rates for women compared to men (Ngueajio and Washington, 2022; Aryal et al., 2023). Models often have higher error rates for children (Bhardwaj et al., 2022). Non-native and non-urban accents have been associated with higher error rates (Serditova et al., 2025; Torgbi et al., 2025; Wassink et al., 2022), with second-language speech accents showing higher error rates than speakers of non-standard or regional accents (Feng et al., 2021; Zhang et al., 2022). For example, commercial models produce twice as many errors for African American speakers as for white speakers of English (Koenecke et al., 2020; Martin and Tang, 2020).

Standardized evaluation benchmarks and bias audits have been introduced to quantify these performance gaps and assess fairness in widely used ASR models. In contrast, investigations into phonetic or IPA-based ASR models remain limited. As phoneme-level systems become more common, either as full speech transcription models or as intermediate components within larger architectures, it is increasingly important to understand how they may contribute to or mitigate existing biases.

3 Methods

3.1 Datasets

We evaluate phoneme recognition models across two complementary dimensions: cross-lingual robustness and demographic variation.

Cross-lingual Datasets. To assess multilingual phoneme recognition performance, we use IPAPACK (FLEURS-IPA) (Zhu et al., 2024), MediaSpeech (Kolobov et al., 2021) and WAXAL (Diack et al., 2026). From these multilingual corpora, we select 11 languages for evaluation, including high-resource languages (English, Spanish, and French) and lower-resource languages (Hindi, Bangla, Panjabi, Tamil, Telugu, Arabic, Turkish, and Shona). Languages are selected based on support from both the G2P pipeline and the evaluated IPA models. Table 1 in Appendix A.1 provides more details about the language distributions of these datasets.

Demographic Datasets. To analyze demographic bias and variation in phoneme recognition, we use English speech datasets with speaker metadata: the Corpus of Regional African American Language (CORAAL) (Kendall and Farrington, 2018), the Edinburgh International Accents of English Corpus (EdAcc) (Sanabria et al., 2023), the Sonos Voice Control Bias Assessment Dataset (SVC) (Sekkat et al., 2024), and WAXAL (Diack et al., 2026). These datasets contain annotations for demographic attributes, including age, gender, ethnicity, accent, geographic region, and native-speaker status, with the demography presented in Table 2 in Appendix A.1.

3.2 Models

In this work, we evaluate two IPA-based ASR models which allows us to maximize coverage in cross-lingual evaluation, since they have the largest number of overlapping supported languages.

WhisperIPA. We evaluate WhisperIPA[2], a transformer-based encoder–decoder model built on the Whisper architecture (Radford et al., 2023). Whisper is originally pre-trained on 680,000 hours of weakly supervised multilingual speech data. The base WhisperIPA variant, with approximately 74M parameters, is further fine-tuned for direct speech-to-IPA transcription on 15,000 labeled synthetic IPA audio samples. The fine-tuning data is derived from Common Voice 21, which covers >70 languages with corresponding phonetic annotations.

ZIPA. We evaluate ZIPA (Zhu et al., 2025), a family of efficient multilingual phone recognition models built on the Zipformer architecture (Yao et al., 2024), that includes both transducer-based and CTC-based variants. In this work, we use the large consistency-regularized CTC model with noisy-student training. The ZIPA-CR-NS[3] large variant has approximately 300M parameters and is trained on IPAPack++, a large-scale multilingual speech corpus with 17,132 hours of normalized phone transcriptions across 88 languages. The noisy-student training stage further incorporates approximately 11,000 hours of pseudo-labeled multilingual speech from >4,000 languages, improving cross-linguistic phone recognition performance.

3.3 Experimental Setup

All models are evaluated in a zero-shot setting without additional fine-tuning on the evaluation datasets in order to measure out-of-the-box phoneme recognition performance. After both models generate IPA transcriptions for every audio sample, outputs are post-processed to remove special tokens, normalize whitespace, and ensure consistency with the reference phoneme format. Output post-processing details are described in Appendix A.2.

Reference phoneme sequences are generated from orthographic transcripts using a multilingual grapheme-to-phoneme (G2P) pipeline. We use G2P+ (Goriely and Buttery, 2025) to generate IPA-based phonemic transcriptions across languages. Although generated phonemic references introduce limitations due to potentially incorrect, oversimplified phoneme mappings, incomplete phonetic detail, and inconsistencies across languages, they provide a scalable framework for multilingual evaluation. These generated phoneme sequences serve

[2]https://huggingface.co/neurlang/ipa-whisper-base

[3]https://huggingface.co/anyspeech/zipa-large-crctc-ns-800k

as the ground-truth references for all experiments.

Soft PER We evaluate phoneme recognition accuracy using phoneme error rate (PER). However, because phoneme realizations can vary across accents, dialects, and languages without reflecting true recognition failures, we also introduce and evaluate a Soft PER metric that reduces penalties for linguistically similar substitutions. This evaluation uses a two-tier phoneme mapping to (1) establish broad interchangeable phonemic variation and (2) create language-specific pronunciation overlaps that are not global equivalents. A single equivalence tier is insufficient because many phonetic relationships are not identical between languages.

The first tier groups phonemes into transitive equivalence classes using two allophonic resources. AlloVera (Mortensen et al., 2020) provides language-specific mappings between surface phones and canonical phonemes across 14 languages (e.g., aspirated /pʰ/→/p/ where a strongly breathed "p" sound is treated as a standard /p/, devoiced /b̥/→/b/ where a partially unvoiced "b" is mapped to its canonical form, etc.), with English given priority, because English defines the largest proportion of the demographic evaluation data, when a surface phone appears in multiple languages. We then use articulatory feature representations from PHOIBLE (Moran and McCloy, 2019) to identify pairs of canonical phonemes that differ in only one of 34 articulatory features. These pairs are merged into shared equivalence classes. Through this grouping, phonemes may belong to the same class even if they differ in more than one feature, as long as they are connected through phonemes that differ by only one feature at each step. Together, these mappings create 62 equivalence classes spanning 254 phones. Table 4 in Appendix A.3 provides the full mapping for the first tier.

The second tier captures non-transitive cross-language similarities, derived from AlloVera (Mortensen et al., 2020). When the same surface phone maps to different canonical phonemes across languages (e.g., the flap sound /ɾ/ is mapped to /t/ in English words such as "water," but to /ɾ/ in Spanish words such as "pero"), those canonical phonemes are recorded as a direct pair without being merged into a shared equivalence class. This preserves the individual relations /t/≈/ɾ/ and /ɾ/≈/ɾ/ without implying /t/≈/ɾ/ and keeping language-specific relationships. A total of 90 such language-specific pairs are extracted from cross-language allophonic overlaps. These pairs are scored as equivalent within their respective language mappings, but are not transitively combined. Table 5 in Appendix A.3 provides the full mapping for the second tier.

For example, Tier 1 places devoiced /b̥/ and canonical /b/ in the same equivalence class, so substituting one for the other incurs no error. In contrast, /d/ and /t/ belong to different Tier 1 classes and would initially be counted as a substitution error. Tier 2 contains this pair as a language-specific similarity in English, Kazakh, Russian, and Vietnamese, so the substitution isn't counted when it happens in one of these languages. Unlike Tier 1 equivalences, this relation is not extended transitively to other phonemes.

During evaluation, Tier 1 equivalence classes are applied universally across all languages, while Tier 2 similarities are only applied when relevant to the target language. Substitutions within either similarity tier receive a zero penalty when computing Soft PER, allowing the metric to better distinguish genuine recognition errors from acceptable phonetic variation.

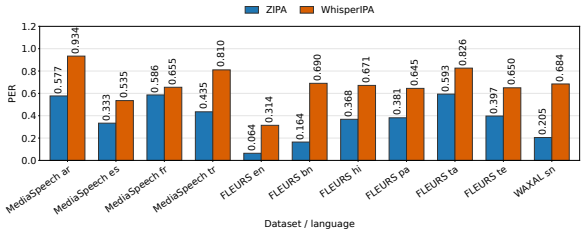
4 Results

4.1 Language Performance Evaluation

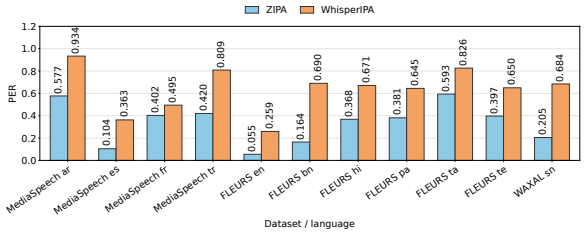
We first evaluate model performance across languages to understand how IPA-based ASR system accuracy varies in a multilingual setting. This analysis compares standard PER and Soft PER for both ZIPA and WhisperIPA across the available language datasets. Standard PER measures exact phoneme-level mismatch, while Soft PER applies the equivalence mappings described in Section 3.3 to reduce penalties for linguistically similar phoneme substitutions.

Figure 1 compares ZIPA and WhisperIPA across the evaluated languages using both standard PER and Soft PER. ZIPA achieves lower error rates than WhisperIPA for every language under both metrics, indicating stronger overall phoneme-recognition performance in this multilingual evaluation. Under standard PER, English has the lowest error rate for both models. For ZIPA, the highest error rates occur for Tamil, French, and Arabic, whereas WhisperIPA produces its highest error rates for Arabic, Tamil, and Turkish. These results demonstrate substantial variation in performance across languages.

Under Soft PER, ZIPA continues to outperform WhisperIPA for every language, although the ef-

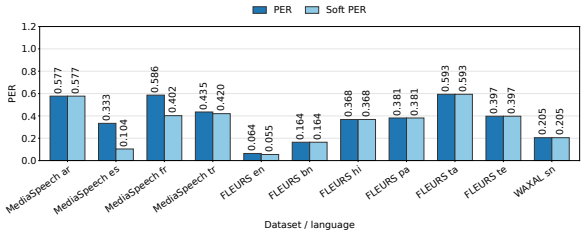


(a) Standard PER across languages for ZIPA and WhisperIPA.

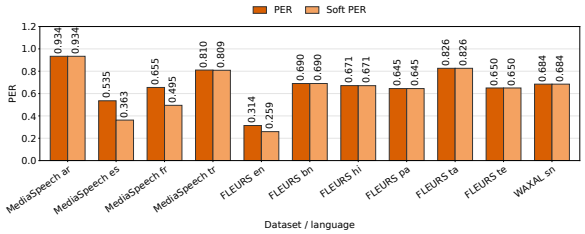


(b) Soft PER across languages for ZIPA and WhisperIPA.

Figure 1: Comparison of standard PER and Soft PER across evaluated languages: (ar=Arabic, es=Spanish, fr=French, tr=Turkish, en=English, bn=Bangla, hi=Hindi, pa=Panjabi, ta=Tamil, te=Telugu, and sn=Shona).



(a) Standard PER and Soft PER comparison for ZIPA.



(b) Standard PER and Soft PER comparison for WhisperIPA.

Figure 2: Effect of Soft PER on language-level evaluation for each model.

fect of the soft-equivalence mapping varies considerably across languages. The largest reductions occur for Spanish and French. For ZIPA, Spanish decreases from 0.333 to 0.104 and French decreases from 0.586 to 0.402. WhisperIPA shows similar reductions, with Spanish decreasing from 0.535 to 0.363 and French decreasing from 0.655 to 0.495. English also improves under Soft PER, particularly for WhisperIPA, while Turkish shows only a small reduction for ZIPA and a negligible change for WhisperIPA. In contrast, the values for Arabic, Bangla, Hindi, Panjabi, Tamil, Telugu, and Shona remain unchanged. Thus, Soft PER affects only a subset of the evaluated languages rather than reducing errors uniformly across the multilingual evaluation.

Figure 2 presents the effect of Soft PER separately for each model. For ZIPA, the reduction is largest for Spanish and French, smaller for Turkish and English, and absent for the remaining languages. WhisperIPA follows a similar pattern, with substantial reductions for Spanish and French, a smaller reduction for English, and almost no change for Turkish. The other languages remain unchanged. Despite these language-specific reductions, WhisperIPA continues to produce higher error rates than ZIPA across all languages.

Key Takeaways: Overall, ZIPA consistently outperforms WhisperIPA across all evaluated languages. ZIPA achieves an average standard PER of 0.373, compared with 0.674 for WhisperIPA, corresponding to an absolute difference of 0.301. Under Soft PER, the averages decrease to 0.333 and 0.639, respectively, while the absolute difference remains nearly unchanged at 0.305. Thus, Soft PER lowers the aggregate error rates but does not narrow the performance gap between the models. This suggests that ZIPA's advantage is not explained primarily by the phonemically similar substitutions captured by the current Soft PER mapping.

Considering English, Spanish, and French as high-resource languages and Hindi, Bangla, Panjabi, Tamil, Telugu, Arabic, Turkish, and Shona as low-resource languages, average error rates are higher for the low-resource group. Averaged across both models, standard PER is 0.565 for low-resource languages and 0.415 for high-resource languages, representing a 36.2% relative increase. Under Soft PER, the corresponding averages are 0.563 and 0.280, producing an absolute difference of 0.284 and a relative difference of approximately 101%.

The widening relative gap under Soft PER is driven primarily by the substantial reductions observed for Spanish and French, both of which belong to the high-resource group. English also improves, although by a smaller amount. In comparison, most low-resource languages remain unchanged, with only a small reduction for Turk-

ish. One possible explanation is that the AlloVera-derived equivalence mappings, which prioritize English when mappings overlap across languages, do not adequately represent phonemic variation in many of the low-resource languages. Differences in model architecture, training objectives, and multilingual training data may also contribute to the observed pattern. Overall, Soft PER identifies phonemically similar substitutions in a limited subset of languages but does not change the model ranking or reduce the average performance disparity between the high- and low-resource groups.

4.2 Demographic Performance Evaluation

We next analyze how phoneme recognition performance varies across demographic groups using CORAAL, EdAcc, SVC, and WAXAL, focusing on gender, age, ethnicity, and accent/dialect. We interpret demographic patterns within each dataset rather than across datasets, since dataset characteristics may contribute more to performance variation than the demographic attributes themselves. We report bootstrap 95% confidence intervals for all group means shown as error bars in the figures and test every pairwise group contrast with a Welch's t-test. Full pairwise results are reported by demographic attribute in Tables 6–11 in Appendix A.4.

Gender: Figure 3 presents PER and Soft PER across gender groups in SVC, EdAcc, WAXAL, and CORAAL datasets. In SVC, EdAcc, and CORAAL datasets, male and female speakers exhibit similar error rates, with no consistent direction of disparity across models or metrics. WAXAL shows the only significant gender contrast: PER and Soft PER for ZIPA are significantly higher for female speakers ($+0.067$, $p < 0.01$; Table 6). In contrast, WhisperIPA shows a larger but not significant descriptive gap in the opposite direction (-0.149), likely reflecting the small male sample ($n = 52$). Soft PER lowers absolute error rates while preserving within-dataset patterns (identical to PER for WAXAL, since none of its substitutions fall under the soft-equivalence mapping). Overall, aside from this single ZIPA effect in WAXAL, the results show limited evidence of systematic gender-based disparity.

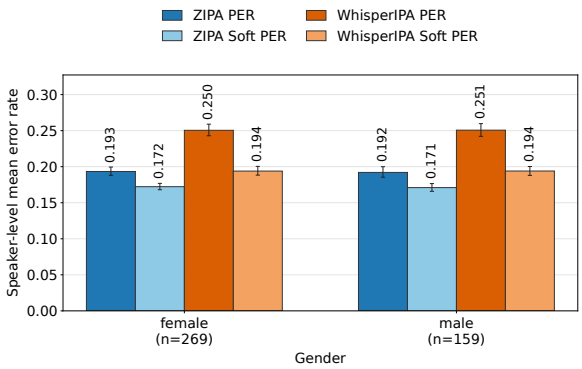
Age: Figure 4 presents PER and Soft PER across age groups in SVC, EdAcc, and CORAAL datasets. In SVC, both models show small absolute differences across age groups (Figure 4a), though several are statistically significant: the 42–54 group has significantly lower PER than the 17–28 group (all four metrics, $p < 0.05$, $p < 0.01$ for ZIPA) and the 9–16 group (ZIPA PER/Soft PER and WhisperIPA Soft PER, $p < 0.05$; Table 8), with no other SVC pair reaching significance. In EdAcc, ZIPA is stable across ages, while WhisperIPA peaks for the 45–59 group (Figure 4b); this should be read cautiously, as the group has only five speakers and none of EdAcc's three pairwise age contrasts reach significance (Table 9). CORAAL[4] shows the clearest pattern: the 0–19 group has significantly higher PER than the 20–29, 30–50, and 51+ groups on all four metrics ($p < 0.01$, mostly $p < 0.001$) and higher ZIPA PER/Soft PER than the 0–29 group ($p < 0.01$), though the WhisperIPA contrast with 0–29 is not significant ($p = 0.11$, $p = 0.13$; Table 10). No other pair differs significantly. Soft PER lowers absolute error rates while largely preserving these patterns. Overall, age effects are dataset-dependent: a robust effect for the youngest speakers in CORAAL, a modest but significant effect in SVC's 42–54 group, and no significant effect in EdAcc despite a large descriptive gap.

Ethnicity: Figure 5 presents PER and Soft PER across ethnicity groups in the EdAcc dataset. "White", "Asian", "Black", and "South Asian" are possible self-selected ethnicity descriptors in EdAcc.[5] As shown in Figure 5a, WhisperIPA produces substantially higher error rates than ZIPA across all groups, consistent with the model-level differences observed throughout the demographic evaluation.

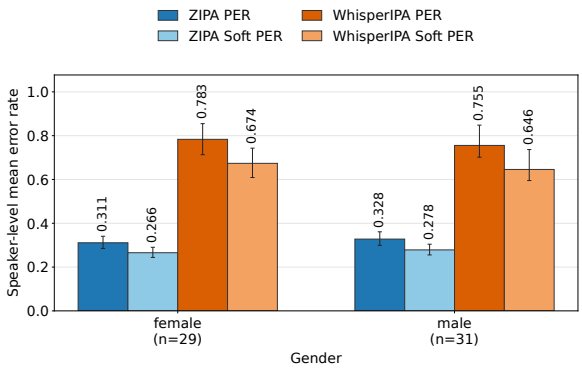
Pairwise Welch's t-tests (Table 7) show "Black" speakers have significantly higher PER than "South Asian" speakers for ZIPA PER, WhisperIPA PER ($+0.157$, $p < 0.01$), and WhisperIPA Soft PER ($+0.144$, $p < 0.05$), and higher PER than "White" speakers for ZIPA PER and Soft PER ($+0.074$, $+0.054$; $p < 0.05$). No other contrast is significant: "Asian" speakers show an elevated disparity from the overall mean (Figure 5b) but are not significantly different from any individual group, including "Black" speakers. All estimates rest on small samples ($n = 6$–23) and should be interpreted accordingly.

[4] CORAAL component corpora use nonuniform age bins; although the speaker groups are disjoint, the source label 0–29 overlaps in range with 0–19 and 20–29 and may partly capture subcorpus differences. See the CORAAL User Guide.

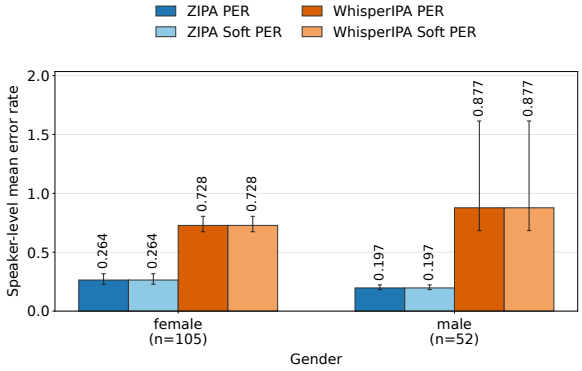
[5] https://groups.inf.ed.ac.uk/EdAcc/EdAcc_statement.pdf



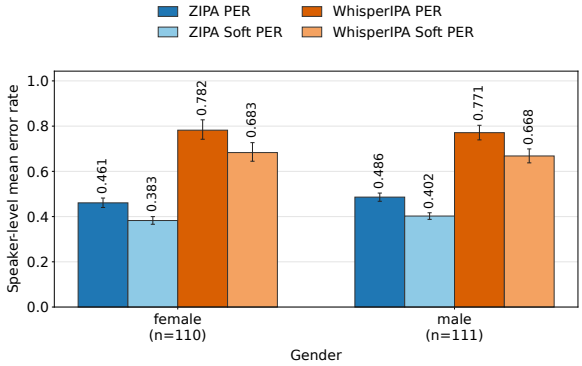
(a) PER and Soft PER by gender for SVC Dataset



(b) PER and Soft PER by gender for EdAcc Dataset



(c) PER and Soft PER by gender for WAXAL Dataset



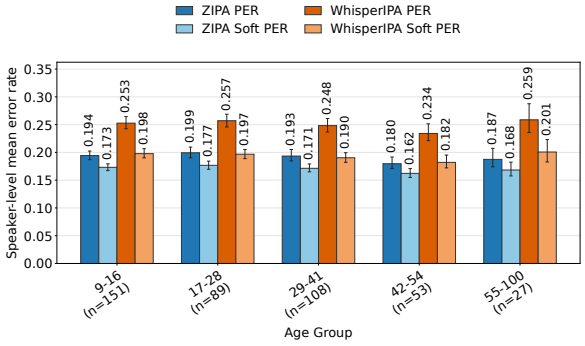
(d) PER and Soft PER by gender for CORAAL Dataset

Figure 3: PER and Soft PER comparison by gender across demographic datasets. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.

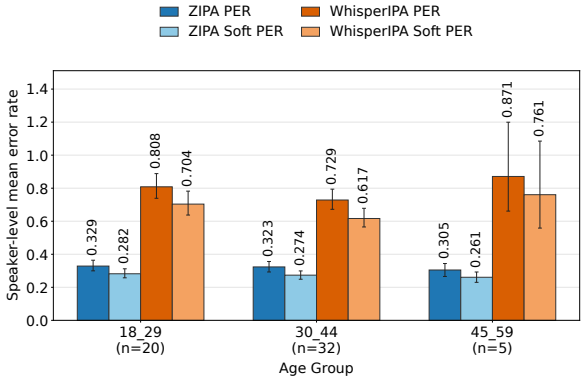
Accent/Dialect: Figure 6 presents PER and Soft PER across accent and dialect groups in the SVC dataset. In SVC, six regional groups—"USA Southern", "USA Western", "USA Mid-Atlantic", "USA Midland", "USA Inland-North", and "USA New England"—represent native speakers of American English of any racial or ethnic group. The "Asian" and "Latino" groups represent non-native speakers from Asia and Latin America who reside in any region of the USA.

As shown in Figure 6a, most USA regional groups show similar or slightly below-average error for both models, while the "Latino" and "Asian" groups have the highest error, particularly for WhisperIPA — reflecting greater difficulty with accents that differ from SVC's native regional varieties. The disparity heatmap (Figure 6b) and pairwise Welch's t-tests confirm this is a robust effect (Table 11): "Latino" speakers have the largest positive disparity from the overall mean ($+0.057$/$+0.037$ ZIPA PER/Soft PER, $+0.063$/$+0.034$ WhisperIPA), and both "Latino" and "Asian" speakers have significantly higher PER than nearly every individual USA regional group ($p < 0.001$ in most cases). "Latino" speakers are also significantly higher than "Asian" speakers under ZIPA ($+0.033$, $+0.021$; $p < 0.05$), though not under WhisperIPA. The six USA regional groups are largely indistinguishable from one another, with only a handful of pairwise contrasts reaching significance. Soft PER reduces the magnitude of these disparities but preserves the relative pattern, indicating the differences are not fully explained by the minor phonemic variation captured by the soft-equivalence mapping.

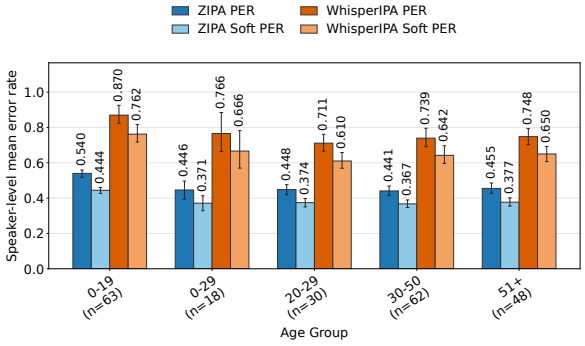
Key Takeaways: Overall, gender shows limited evidence of disparity, aside from a small but significant ZIPA effect in WAXAL. Accent and ethnicity show the most robust effects: in SVC, "Latino" and "Asian" speakers have significantly higher PER than nearly every USA regional group across both models; in EdAcc, "Black" speakers have significantly higher PER than "South Asian" speakers (and, for ZIPA, "White" speakers as well), while "Asian" speakers are not significantly different from any other group despite an elevated disparity from



(a) PER and Soft PER by age group for the SVC dataset.



(b) PER and Soft PER by age group for the EdAcc dataset.



(c) PER and Soft PER by age group for the CORAAL dataset.

Figure 4: PER and Soft PER comparison by age group across datasets. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.

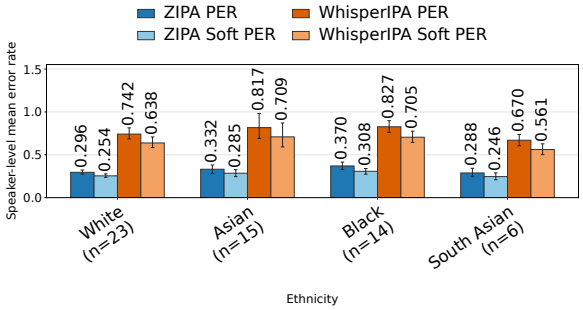
the overall mean. Age effects are more dataset-dependent: a robust elevation for the youngest speakers in CORAAL, a modest but significant effect in SVC's 42–54 group, and no significant effect in EdAcc despite a large descriptive gap. Several estimates rest on small samples and should be interpreted cautiously. Overall, demographic performance patterns reflect an interaction of model behavior, dataset characteristics, and group representation rather than any single attribute.

PER vs. Soft PER: Soft PER generally reduces absolute error rates while preserving the relative performance patterns observed across demographic groups. If the disparities measured by standard PER were driven primarily by phonologically similar substitutions, including variations associated with accents or dialects, applying the soft-equivalence mapping would be expected to narrow the gaps between groups substantially. Instead, the persistence of similar group-level patterns under both metrics suggests that the observed differences also involve recognition errors not covered by the current mapping. The limited effect of Soft PER may additionally reflect incomplete coverage of demographic and cross-lingual phonemic variation within the equivalence classes. In this evaluation, standard PER and Soft PER therefore lead to broadly similar qualitative conclusions about relative group performance. Standard PER remains useful for comparison with prior work, whereas Soft PER provides a complementary measure for examining how acceptable phonemic variation contributes to the reported error rates.

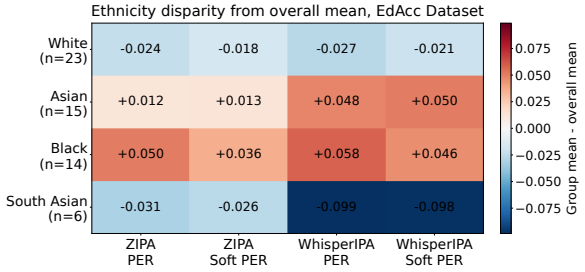
5 Conclusion

In this work, we evaluate bias and performance variation in IPA-based phoneme recognition models across languages and speaker demographics. We compare two open-source IPA transcription systems, ZIPA and WhisperIPA, using both standard PER and a proposed Soft PER metric that reduces penalties for linguistically similar phoneme substitutions. Across the multilingual evaluation, ZIPA consistently achieves lower error rates than WhisperIPA, while both models show substantial variation across languages. This performance gap may be partly attributable to differences in model architecture, training objective, and training data scale. Although Soft PER reduces absolute error rates for several languages, it does not remove cross-lingual performance disparities or change the overall model ranking. This indicates that some standard PER errors reflect acceptable phonemic variation, but that meaningful language-level differences remain even under relaxed matching.

The demographic evaluation also shows that the performance variation is not uniform across speaker identity characteristics. Gender-based differences are small and largely inconsistent across datasets and models, with one exception: ZIPA shows a small but statistically significant gender

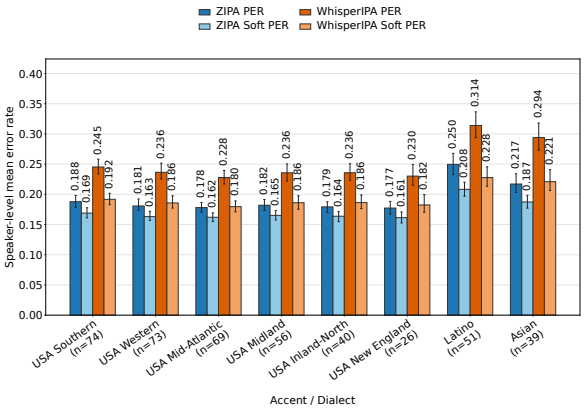


(a) PER and Soft PER by ethnicity group for the EdAcc dataset.

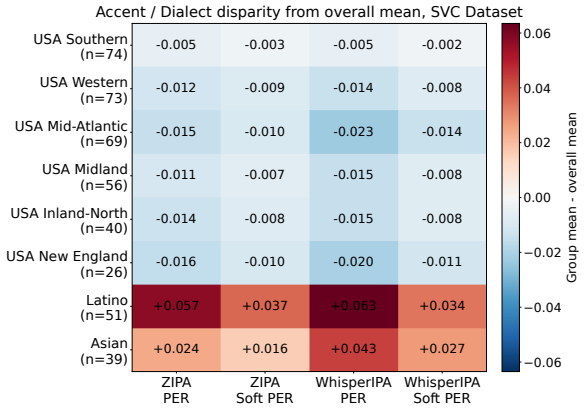


(b) Ethnicity disparity from the overall mean for the EdAcc dataset.

Figure 5: PER, Soft PER, and group-level disparity by ethnicity for the EdAcc dataset. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.



(a) PER and Soft PER by accent/dialect group for the SVC dataset.



(b) Accent/dialect disparity from the overall mean for the SVC dataset.

Figure 6: PER, Soft PER, and group-level disparity by accent/dialect for the SVC dataset. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.

effect in WAXAL. Accent and ethnicity show the most robust effects: in SVC, "Latino" and "Asian" speakers have significantly higher PER than nearly every USA regional group across both models, while in EdAcc, "Black" speakers have significantly higher PER than "South Asian" speakers, and, for ZIPA specifically, than White speakers as well; "Asian" speakers, despite an elevated disparity from the overall mean, are not significantly different from any individual ethnicity group under pairwise testing. Age effects are more dataset-dependent, with a robust elevation for the youngest speakers in CORAAL, a modest but significant effect in SVC's 42–54 group, and no statistically significant effect in EdAcc despite a large descriptive gap for its smallest age group. Across demographic attributes, Soft PER lowers absolute error rates but generally preserves relative group-level patterns, suggesting that observed disparities are not fully explained by minor phonemic variation alone.

Overall, our findings show that IPA-based ASR systems exhibit demographic and language-level disparities in measured PER. However, these disparities are measured against G2P references, which may not capture legitimate pronunciation variation and can conflate differences in realized pronunciation with ASR error. Although phoneme-based representations provide a promising foundation for low-resource and multilingual speech technologies, their performance can still vary substantially between languages, accents, and demographic groups. These results highlight the need for more careful evaluation of phoneme-level ASR systems, richer demographic and sociophonetic benchmarks, and evaluation metrics that distinguish genuine recognition errors from acceptable pronunciation variation.

Limitations

The key limitation of phoneme-level evaluation is the reliance on auto-generated IPA ground truth, because of the high cost of specialized manual phonemic annotation. Unlike orthographic transcriptions, phonemic annotations depend on G2P conversion and forced alignment, which can introduce systematic noise and model bias in the ground truth. Because our pipeline converts written transcripts into IPA representations, the resulting "ground truth" often reflects standardized pronunciation, which can encode biases of what constitutes "correct" speech into our analysis (Vythelingum et al., 2017; MacKenzie and Turton, 2020). This issue will naturally increase the error rates for accented speech, dialectal variation, and non-canonical pronunciations, where no single canonical phonemic form may exist (Fatema et al., 2024). Even with a softened PER, it may reflect annotation artifacts rather than true model error. Addressing this limitation may require future evaluation frameworks that utilize expert-validated phonemic annotations.

Our Soft PER hopes to partially mitigate sensitivity to annotation noise. While this acts as a necessary smoothing, it introduces a tradeoff between linguistic fidelity and metric standardization. Our definition of Soft PER may not be directly comparable to standard PER or WER, limiting cross-study comparability. Additionally, the equivalence mapping relies on English-prioritized relations when a surface phone appears across multiple languages. Although this provides a consistent normalization strategy, it may bias the metric toward English phonological interpretations and over-normalize distinctions that remain contrastive in other languages. Future work should explore language-adaptive mappings that better preserve cross-lingual phonological variation.

The analysis is further constrained by the demographic metadata available in the evaluated datasets. These categories do not capture finer sociophonetic variation such as code-switching, intra-speaker variability, or regional micro-dialects. In addition, certain demographic groups are underrepresented, which may limit the statistical robustness and generalizability of the results. Future work could address this by incorporating or creating a dataset with richer sociolinguistic annotations optimized for multilingual analysis to better understand the relationship among different dimensions.

Finally, while IPA provides a more fine-grained representation of speech than grapheme-based transcription, it is an abstract acoustic representation. There is ongoing debate in linguistics regarding the extent to which phonemic representations can fully capture gradient phonetic variation (Mallaband, 2024; Mitterer et al., 2018). Dialectal variation and allophonic realizations may not be fully represented within standardized IPA inventories, which may further influence both the accuracy of the ground truth and evaluation results (Fatema et al., 2024). Hybrid evaluation frameworks in the future that jointly leverage phonemic representations and acoustic similarity measures could potentially provide a more accurate representation of ASR performance across diverse demographics.

References

Surakshya Aryal, Mikel K Ngueajio, Saurav Keshari Aryal, and Gloria Washington. 2023. Hey, siri! why are you biased against women?(student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16154–16155.

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. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Shikhar Bharadwaj, Chin-Jou Li, Yoonjae Kim, Kwanghee Choi, Eunjung Yeo, Ryan Soh-Eun Shim, Hanyu Zhou, Brendon Boldt, Karen Rosero Jacome, Kalvin Chang, Darsh Agrawal, Keer Xu, Chao-Han Huck Yang, Jian Zhu, Shinji Watanabe, and David R. Mortensen. 2026. Prism: Benchmarking phone realization in speech models. *Preprint*, arXiv:2601.14046.

Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. 2022. Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9):4419.

Jay L Cunningham, Adinawa Adjagbodjou, Jeffrey Basoah, Jainaba Jawara, Kowe Kadoma, and Aaleyah Lewis. 2025. Toward responsible ASR for african american english speakers: A scoping review of bias and equity in speech technology. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 665–678.

Massimo Marie Daul, Alessio Tosolini, and Claire Bowern. 2026. Linguistically informed tokenization improves ASR for underresourced languages. In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 31–37, Rabat, Morocco. Association for Computational Linguistics.

Kumari Jayashree Devi and Parismita Sarma. 2026. Automated international phonetic alphabet (ipa) representation for assamese speech. In *ICT for Intelligent Systems*, pages 603–614, Singapore. Springer Nature Singapore.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokpor, Fiifi Baffoe Payin Winful, Kafui Kwashie Solaga, Jamal-Deen Abdulai, Akon Obu Ekpezu, and 24 others. 2026. Waxal: A large-scale multilingual african language speech corpus. *Preprint*, arXiv:2602.02734.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Performance disparities between accents in automatic speech recognition (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16200–16201.

Hend ElGhazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025. Exploring gender disparities in automatic speech recognition technology. *arXiv preprint arXiv:2502.18434*.

Anjie Fang, Simone Filice, Nut Limsopatham, and Oleg Rokhlenko. 2020. Using phoneme representations to build predictive models robust to asr errors. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 699–708.

Kanij Fatema, Fazle Dawood Haider, Nirzona Ferdousi Turpa, Tanveer Azmal, Sourav Ahmed, Navid Hasan, Mohammad Akhlaqur Rahman, Biplab Kumar Sarkar, Afrar Jahin, Md. Rezuwan Hassan, Md Foriduzzaman Zihad, Rubayet Sabbir Faruque, Asif Sushmit, Mashrur Imtiaz, Farig Sadeque, and Syed Shahrier Rahman. 2024. Ipa transcription of bengali texts. *Preprint*, arXiv:2403.20084.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Li Fu, Yu Xin, Sunlu Zeng, Lu Fan, Youzheng Wu, and Xiaodong He. 2025. Pac: Pronunciation-aware contextualized large language model-based automatic speech recognition. *arXiv preprint arXiv:2509.12647*.

Robert C. Gale, Alexandra C. Salem, Gerasimos Fergadiotis, and Steven Bedrick. 2023. Mixed orthographic/phonemic language modeling: Beyond orthographically restricted transformers (BORT). In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 212–225, Toronto, Canada. Association for Computational Linguistics.

Zébulon Goriely and Paula Buttery. 2025. IPA CHILDES & G2P+: Feature-rich resources for cross-lingual phonology and phonemic language modeling. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 502–521, Vienna, Austria. Association for Computational Linguistics.

Jan Gorisch and Thomas Schmidt. 2024. Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6564–6574, Torino, Italia. ELRA and ICCL.

Maliha Jahan, Priyam Mazumdar, Thomas Thebaud, Mark Hasegawa-Johnson, Jesús Villalba, Najim Dehak, and Laureano Moro-Velazquez. 2025. Unveiling performance bias in asr systems: A study on gender, age, accent, and more. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tyler Kendall and Charlie Farrington. 2018. The corpus of regional african american language. *Version*, 6:1.

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.

Rostislav Kolobov, Olga Okhapkina, Olga Omelchishina, Andrey Platunov, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. Mediaspeech: Multilanguage asr benchmark and dataset. *Preprint*, arXiv:2103.16193.

Jaeyoung Lee, Masato Mimura, and Tatsuya Kawahara. 2025. Leveraging ipa and articulatory features as effective inductive biases for multilingual asr training. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and 1 others. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Interspeech*, pages 711–714.

Laurel MacKenzie and Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of british english. *Linguistics Vanguard*, 6(s1):20180061.

Laura Jane Mallaband. 2024. The agreement of phonetic transcriptions between paediatric speech and language therapists transcribing a disordered speech sample. *International Journal of Language & Communication Disorders*, 59(5):1981–1995.

Joshua L Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual" be". In *Interspeech*, pages 626–630.

Yangyang Meng, Jinpeng Li, Guodong Lin, Yu Pu, Guanbo Wang, Hu Du, Zhiming Shao, Yukai Huang, Ke Li, and Wei-Qiang Zhang. 2025. Dolphin: A large-scale automatic speech recognition model for eastern languages. *arXiv preprint arXiv:2503.20212*.

Holger Mitterer, Eva Reinisch, and James M McQueen. 2018. Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, 98:77–92.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

David R Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan W Black, Florian Metze, and Graham Neubig. 2020. Allovera: A multilingual allophone database. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5329–5336.

Mikel K Ngueajio and Gloria Washington. 2022. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *International conference on human-computer interaction*, pages 421–440. Springer.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

Kentaro Onda, Satoru Fukayama, Daisuke Saito, and Nobuaki Minematsu. 2026. Advanced modeling of interlanguage speech intelligibility benefit with l1-l2 multi-task learning using differentiable k-means for accent-robust discrete token-based asr. *Preprint*, arXiv:2601.19767.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Francis McCann Ramirez, Luka Chkhetiani, Andrew Ehrenberg, Robert McHardy, Rami Botros, Yash Khare, Andrea Vanzo, Taufiquzzaman Peyash, Gabriel Oexle, Michael Liang, Ilya Sklyar, Enver Fakhan, Ahmed Etefy, Daniel McCrystal, Sam Flamini, Domenic Donato, and Takuya Yoshioka. 2024. Anatomy of industrial scale multilingual asr. *Preprint*, arXiv:2404.09841.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Michael Scott, Siyu Liang, Alicia Wassink, and Gina-Anne Levow. 2025. A sociophonetic analysis of racial bias in commercial asr systems using the pacific northwest english corpus. *arXiv preprint arXiv:2510.22495*.

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.

Dana Serditova, Kevin Tang, and Jochen Steffens. 2025. Automatic Speech Recognition Biases in Newcastle English: an Error Analysis. In *Interspeech 2025*, pages 3204–3208.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal Automatic Phonetic Transcription into the International Phonetic Alphabet. In *Interspeech 2023*, pages 2548–2552.

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. In *Interspeech 2017*, pages 934–938.

Melissa Torgbi, Andrew Clayman, Jordan J Speight, and Harish Tayyar Madabushi. 2025. Adapting whisper for regional dialects: Enhancing public services for vulnerable populations in the united kingdom. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 29–38.

Kévin Vythelingum, Yannick Estève, and Olivier Rosee. 2017. Error detection of grapheme-to-phoneme conversion in text-to-speech synthesis using speech signal and lexical context. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 692–697. IEEE.

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

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *International Conference on Learning Representations*, volume 2024, pages 44440–44455.

Yuanyuan Zhang, Aaricia Herygers, Tanvina Patel, Zhengjun Yue, and Odette Scharenborg. 2023. Exploring data augmentation in bias mitigation against non-native-accented speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yuanyuan Zhang, Yixuan Zhang, Bence Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Interspeech 2022*, pages 3168–3172.

Jian Zhu, Farhan Samir, Eleanor Chodroff, and David R Mortensen. 2025. Zipa: A family of efficient models for multilingual phone recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19568–19585.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of ipa: Towards open-vocabulary keyword spotting and forced alignment in any language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 750–772, Mexico City, Mexico. Association for Computational Linguistics.

Piotr Żelasko, Siyuan Feng, Laureano Moro Velázquez, Ali Abavisani, Saurabhchand Bhati, Odette Scharenborg, Mark Hasegawa-Johnson, and Najim Dehak. 2022. Discovering phonetic inventories with crosslingual automatic speech recognition. *Computer Speech & Language*, 74:101358.

A Appendix

A.1 Dataset Details

Table 1 summarizes the multilingual speech datasets used for cross-lingual evaluation, including the languages represented in each dataset and the dataset split used for analysis. Together, these datasets support evaluation across both high-resource and lower-resource languages.

Table 2 provides a detailed summary of the English-language datasets and the speaker-level demographic metadata available in each dataset. These metadata fields, including age, gender, ethnicity, accent, and dialectal region, are used to evaluate whether IPA-based ASR models exhibit systematic performance variation across demographic groups.

A.2 IPA Transcription Post-Processing

To ensure that the evaluation focused on phonemic correspondence rather than superficial transcription differences, we apply the same normalization procedure to both the G2P-generated ground-truth IPA transcriptions and the model-generated outputs before computing PER and Soft PER as described in Table 3. The normalization rules are designed to reduce variation caused by stress notation, IPA symbol variants, tokenization conventions, and formatting artifacts, while preserving the underlying phone sequence as much as possible.

After applying these transformations, the normalized phone strings were re-tokenized into whitespace-separated phone sequences. This produced a consistent representation across reference and generated outputs, reducing the impact of orthographic IPA variants and tokenization artifacts on downstream evaluation.

A.3 Soft PER Details

Table 4 lists the transitive phonetic equivalence classes used in the first tier of Soft PER. These classes group phones that are treated as interchangeable during evaluation, allowing the metric to avoid penalizing substitutions that correspond to linguistically similar or allophonic variants. The Tier 1 mapping is applied universally across languages and is intended to capture broad phonemic variation that should not necessarily be counted as a recognition error.

The Tier 1 equivalence classes are constructed from two sources of phonetic similarity. First, language-specific allophonic mappings are derived from AlloVera, which links surface phones to canonical phonemes across multiple languages. Second, articulatory feature representations from PHOIBLE are used to identify canonical phonemes that differ by only one articulatory feature. These phones are merged into shared transitive classes, producing a mapping that spans 62 equivalence classes and 254 phones. During Soft PER computation, substitutions within the same Tier 1 class receive zero penalty, allowing the metric to better distinguish substantial recognition errors from acceptable phonetic variation.

Table 5 lists the language-specific non-transitive similarity pairs used in the second tier of Soft PER. Unlike the Tier 1 classes, these pairs are not merged into broader equivalence classes. Instead, they capture cases where the same surface phone corresponds to different canonical phonemes across languages. This allows Soft PER to recognize language-specific pronunciation overlap without incorrectly making all connected phonemes globally interchangeable.

The Tier 2 mapping is also derived from AlloVera. When a surface phone maps to different canonical phonemes in different languages, the corresponding canonical phonemes are recorded as a direct similarity pair. During evaluation, these pairs are applied only when relevant to the target language. Substitutions covered by Tier 2 receive zero penalty, but the relation remains non-transitive; for example, if two separate language-specific mappings relate one phone to two different phones, Soft PER does not automatically treat those two phones as equivalent to each other. This design preserves language-specific phonological distinctions while still reducing penalties for acceptable cross-lingual or allophonic variation.

A.4 Pairwise Welch's t-test Details

Tables 6–11 report pairwise comparisons of speaker-level PER between demographic groups within each dataset, for the demographic attributes examined in Section 4.2: gender, age, ethnicity, and accent/dialect. For every pair of groups within a demographic attribute, we compute the difference in group-mean PER (row minus column) and test the difference using a two-sided Welch's t-test. This is reported separately for ZIPA and WhisperIPA, under both standard PER and Soft PER.

In each table, the reported value is the point-estimate difference in speaker-level mean PER be-

Table 1: Dataset Metadata and Speech Characteristics

Dataset	**Languages / Varieties**	**Subset**[†]	**Collection Period**	**Speech Style**	**Recording Environment**
Fleurs_IPA	English, Hindi, Bangla, Panjabi, Tamil, Telugu	Test	2020–2022	Read speech (utterance-level)	Unedited original recordings
MediaSpeech	Arabic, French, Spanish, Turkish	N/A	2020	Read and spontaneous speech	YouTube broadcasts/channels
Waxal	Shona	Train	Jan. 2021–Mar. 2024	Natural, spontaneous speech	Natural environments
EdAcc	English	Test	2023	Naturalistic, spontaneous interactions	Zoom recordings
SVC	English	Train	2024	Read speech	Clean, close-field environments
CORAAL	African American English	Train	2023	Sociolinguistic interviews	Unedited original recordings

[†] Dataset partition used for bias evaluation.

Table 2: Dataset Languages and Demographic Attributes

Dataset	**Language(s)**	**Demographic Attributes**
WAXAL	Shona	gender
EdAcc	English	gender, year_of_birth, ethnicity, education, l1, l2, accent, english_since, countries_lived, lang_home, lang_friends
SVC Bias Assessment	English	age_group, gender, dialectal_region
CORAAL	African American English	city, se_group, age_group, gender

tween the row groups and column groups. A positive value indicates that the row group has a higher (worse) mean PER than the column group.

Group-level means and their 95% confidence intervals, computed using bias-corrected and accelerated (BCa) bootstrap with 10,000 resamples (percentile intervals when a group contains fewer than 10 speakers or when the BCa procedure does not converge), are the same estimates used for the error bars in the demographic performance figures in Section 4.2.

Table 3: Normalization rules applied to both ground-truth IPA transcriptions and model-generated outputs prior to evaluation.

Category	Normalization behavior
Stress marks	Primary and secondary stress marks, such as ˈ and ˌ, were removed.
Length marks	Length markers, including ː, ˑ, and the ASCII colon :, were removed.
Rhotics	Rhotic consonants such as ɹ, ɾ, ʁ, and ʀ were mapped to ɾ. Rhotic vowels such as ɚ and ɝ were expanded to ə ɾ and ɛɾ, respectively. Rhoticity marks, such as əʳ, were also normalized to a vowel followed by ɾ.
Diphthongs	Joined diphthong symbols were tokenized as separate phone sequences. For example, eɪ was normalized to e ɪ.
Affricates	Joined and tie-barred affricates were tokenized as separate phones. For example, tʃ, t͡ʃ, and ʧ were normalized to t ʃ.
Token separators	Whitespace, SentencePiece markers such as ▁, hyphens, slashes, and similar separators were collapsed into phone-token boundaries.
Combining marks	Separated combining marks were attached to the preceding phone. For example, ɑ ̃ was normalized to ɑ̃.
Symbol choices	Common spelling variants were standardized. For example, ASCII g was normalized to IPA ɡ, and dark-*l* variants such as lˠ, lˠ, and separated l ˠ were normalized to ɫ.

Table 4: Phonetic Transitive Equivalence Classes (Tier 1).

Index	Representations
0	b’, p’
1	b, b̥, b̤
2	dʒ, d͡ʒ
3	d, d̤
4	e, e̤
5	kʷ’, k’ʷ
6	k, k̚
7	o, o̥, o̤
8	pf, p͡f
9	ts, t͡s
10	tɕ, t͡ɕ
11	tʃ, t͡ʃ
12	tʷ’, t’ʷ
13	t, t̚
14	v, v̥
15	xʲ, x̥ʲ
16	ŋ, ŋ̊
17	ɑ, ɑ̱
18	ɕ, ɕː
19	ɡ, ɡ̊
20	ɾʲ, ɾ̤ʲ
21	ʉ̥, ʏ

Table 5: Language Specific Non-Transitive Pairs (Tier 2).

Phoneme 1	Phoneme 2	Language	Phoneme 1	Phoneme 2	Language
a	aː	jpn, tur	a	æ	deu, eng
a	ɑ	cmn, eng	a	ɔ	eng, jav
a	ə	eng, rus	a	ɛ	cmn, eng
a	ʊ	eng, rus	aː	ɑ	deu, eng
b	bʲ	jpn, rus	b	p	amh, eng, kaz, rus, spa
b	v	spa, tur	bʲ	pʲ	rus
d	dʒ	ita, tgl	d	t	eng, kaz, rus, vie
d	ð	eng, spa	d	ɳ	spa, tgl
dʲ	tʲ	rus	dʒ	ts	ita, tgl
dʒ	ʒ	amh, eng, ita, tgl	e	eː	deu, eng, jpn, tur
e	i	eng, tgl	e	æ	eng, tur
e	ɛ	deu, eng	e	ɪ	eng, tgl
e	ʊ	cmn, eng	h	ɾ	fra, kaz
h	x	deu, tur	h	ɕ	deu, jpn
i	iː	deu, eng, jpn, tur	i	ɛ	eng, tgl
i	ɪ	deu, eng	j	y	eng, jpn
j	ɕ	deu, rus, spa	j	ɲ	cmn, fra
j	ɪ	eng	j	ʎ	deu, spa
k	q	eng, jav	k	x	deu, tgl
k	ɡ	amh, deu, eng, kaz, rus	k	ɡʲ	rus
m	mʲ	jpn, rus	m	n	eng, ita
m	ŋ	eng, ita, tgl	m	ɲ	eng, fra, ita
n	ŋ	eng, spa, tgl	n	ɲ	eng, fra, ita
o	oː	deu, eng	o	u	eng, tgl
o	ɔ	eng	o	ʊ	eng, tgl
p	pʲ	jpn, rus	p	q	jpn, vie
q	ɾ	fra, kaz	q	s	eng, jpn
r	x	deu, fra, kaz, spa	r	ɪ	eng, spa
r	ɳ	jpn, spa, tgl	r	ʃ	eng, kaz
r	ʒ	eng, kaz	s	z	amh, deu, eng, kaz
s	ʃ	eng, tgl	s	θ	eng, spa
t	tʃ	spa, tgl	ts	tʃ	spa, tgl
ts	z	ita, jpn	u	uː	deu, eng, jpn, tur
u	w	eng, ita	u	ɨ	jpn, rus
u	ʊ	deu, eng	w	ʊ	eng
x	ɡ	rus, spa	y	yː	deu, fra
ø	øː	deu, fra	ø	œ	deu, fra
ŋ	ɲ	eng, ita	ɔ	ə	amh, eng
ə	ɛ	amh, eng	ɡ	ɳɻ	tgl, tur
ɨ	ɪ	amh, eng	ɨ	ŋ	rus, vie
ɨ	ʊ	amh, eng	ɪ	ɳ	eng, tgl
ʃ	ʒ	eng, kaz	ʎ	ʒ	eng, spa

Table 6: Gender bias analysis across all four demographic datasets. Values represent the difference in speaker-level mean PER between the row and column groups (row − column). Statistically significant difference in mean between two groups via Welch's t-test is shown as: $*$: $p < 0.05$; $**$: $p < 0.01$; $***$: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

Dataset	Groups	ZIPA PER	ZIPA Soft PER	Whisper PER	Whisper Soft PER
SVC	Female vs Male	+0.001	+0.001	-0.001	+0.000
EdAcc	Female vs Male	-0.017	-0.012	+0.028	+0.028
WAXAL	Female vs Male	**+0.067****	**+0.067****	-0.149	-0.149
CORAAL	Female vs Male	-0.025	-0.019	+0.011	+0.015

Table 7: Ethnicity bias analysis on the EdAcc dataset. Values are the difference in speaker-level mean PER between the row and column groups (row − column). Statistically significant difference in mean between two groups via Welch's t-test is shown as: $*$: $p < 0.05$; $**$: $p < 0.01$; $***$: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		Black	South Asian	White
ZIPA PER	Asian	-0.038	+0.044	+0.036
ZIPA Soft PER		-0.023	+0.039	+0.031
Whisper PER		-0.010	+0.147	+0.075
Whisper Soft PER		+0.004	+0.148	+0.071
ZIPA PER	Black		**+0.082***	**+0.074***
ZIPA Soft PER			+0.062	**+0.054***
Whisper PER			**+0.157****	+0.085
Whisper Soft PER			**+0.144***	+0.067
ZIPA PER	South Asian			-0.008
ZIPA Soft PER				-0.008
Whisper PER				-0.072
Whisper Soft PER				-0.077

Table 8: Age-group bias analysis on the SVC dataset. Values are the difference in speaker-level mean PER between the row and column groups (row − column). Statistically significant difference in mean between two groups via Welch's t-test is shown as: $*$: $p < 0.05$; $**$: $p < 0.01$; $***$: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		17-28	29-41	42-54	55-100
ZIPA PER	9-16	-0.005	+0.001	**+0.014***	+0.007
ZIPA Soft PER		-0.004	+0.002	**+0.011****	+0.005
Whisper PER		-0.004	+0.005	+0.019	-0.006
Whisper Soft PER		+0.001	+0.008	**+0.016***	-0.003
ZIPA PER	17-28		+0.006	**+0.019****	+0.012
ZIPA Soft PER			+0.006	**+0.015****	+0.009
Whisper PER			+0.009	**+0.023***	-0.002
Whisper Soft PER			+0.007	**+0.015***	-0.004
ZIPA PER	29-41			+0.013	+0.006
ZIPA Soft PER				+0.009	+0.003
Whisper PER				+0.014	-0.011
Whisper Soft PER				+0.008	-0.011
ZIPA PER	42-54				-0.007
ZIPA Soft PER					-0.006
Whisper PER					-0.025
Whisper Soft PER					-0.019

Table 9: Age-group bias analysis on the EdAcc dataset. Values are the difference in speaker-level mean PER between the row and column groups (row − column). Statistically significant difference in mean between two groups via Welch's t-test is shown as: $*$: $p < 0.05$; $**$: $p < 0.01$; $***$: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		30–44	45–59
ZIPA PER	18–29	+0.006	+0.024
ZIPA Soft PER		+0.008	+0.021
Whisper PER		+0.079	-0.063
Whisper Soft PER		+0.087	-0.057
ZIPA PER	30–44		+0.018
ZIPA Soft PER			+0.013
Whisper PER			-0.142
Whisper Soft PER			-0.144

Table 10: Age-group bias analysis on the CORAAL dataset. Values are the difference in speaker-level mean PER between the row and column groups (row − column). Statistically significant difference in mean between two groups via Welch's t-test is shown as: $*$: $p < 0.05$; $**$: $p < 0.01$; $***$: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		0-29	20-29	30-50	51+
ZIPA PER	0-19	**+0.094****	**+0.092*****	**+0.099*****	**+0.085*****
ZIPA Soft PER		**+0.073****	**+0.070*****	**+0.077*****	**+0.067*****
Whisper PER		+0.104	**+0.159*****	**+0.131*****	**+0.122*****
Whisper Soft PER		+0.096	**+0.152*****	**+0.120****	**+0.112****
ZIPA PER	0-29		-0.002	+0.005	-0.009
ZIPA Soft PER			-0.003	+0.004	-0.006
Whisper PER			+0.055	+0.027	+0.018
Whisper Soft PER			+0.056	+0.024	+0.016
ZIPA PER	20-29			+0.007	-0.007
ZIPA Soft PER				+0.007	-0.003
Whisper PER				-0.028	-0.037
Whisper Soft PER				-0.032	-0.040
ZIPA PER	30-50				-0.014
ZIPA Soft PER					-0.010
Whisper PER					-0.009
Whisper Soft PER					-0.008

Table 11: Accent/dialect bias analysis on the SVC dataset. Region names abbreviated (Inland-N.=USA Inland-North, Mid-Atl.=USA Mid-Atlantic, New Eng.=USA New England). Values are the difference in speaker-level mean PER between the row and column groups (row − column). Statistically significant difference in mean between two groups via Welch's t-test is shown as: $*$: $p < 0.05$; $**$: $p < 0.01$; $***$: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		Latino	Inland-N.	Mid-Atl.	Midland	New Eng.	Southern	Western
ZIPA PER	Asian	**-0.033****	**+0.038*****	**+0.039*****	**+0.035*****	**+0.040*****	**+0.029****	**+0.036*****
ZIPA Soft PER		**-0.021***	**+0.023*****	**+0.025*****	**+0.022*****	**+0.026*****	**+0.018****	**+0.024*****
Whisper PER		-0.020	**+0.058*****	**+0.066*****	**+0.058*****	**+0.064*****	**+0.049*****	**+0.058*****
Whisper Soft PER		-0.007	**+0.035****	**+0.041*****	**+0.035****	**+0.039****	**+0.029****	**+0.035****
ZIPA PER	Latino		**+0.071*****	**+0.072*****	**+0.068*****	**+0.073*****	**+0.062*****	**+0.069*****
ZIPA Soft PER			**+0.044*****	**+0.046*****	**+0.043*****	**+0.047*****	**+0.039*****	**+0.045*****
Whisper PER			**+0.078*****	**+0.086*****	**+0.078*****	**+0.084*****	**+0.069*****	**+0.078*****
Whisper Soft PER			**+0.042*****	**+0.048*****	**+0.042*****	**+0.046*****	**+0.036*****	**+0.042*****
ZIPA PER	Inland-N.			+0.001	-0.003	+0.002	-0.009	-0.002
ZIPA Soft PER				+0.002	-0.001	+0.003	**-0.005***	+0.001
Whisper PER				+0.008	+0.000	+0.006	-0.009	+0.000
Whisper Soft PER				+0.006	+0.000	+0.004	-0.006	+0.000
ZIPA PER	Mid-Atl.				-0.004	+0.001	-0.010	-0.003
ZIPA Soft PER					-0.003	+0.001	**-0.007***	-0.001
Whisper PER					-0.008	-0.002	**-0.017***	-0.008
Whisper Soft PER					-0.006	-0.002	-0.012	-0.006
ZIPA PER	Midland					+0.005	-0.006	+0.001
ZIPA Soft PER						+0.004	-0.004	+0.002
Whisper PER						+0.006	-0.009	+0.000
Whisper Soft PER						+0.004	-0.006	+0.000
ZIPA PER	New Eng.						-0.011	-0.004
ZIPA Soft PER							**-0.008***	-0.002
Whisper PER							-0.015	-0.006
Whisper Soft PER							-0.010	-0.004
ZIPA PER	Southern							+0.007
ZIPA Soft PER								+0.006
Whisper PER								+0.009
Whisper Soft PER								+0.006