Linguistic Distance Segregates Latent Representations in Automatic Speech Recognition Systems

Anonymous ACL submission

Abstract

While automatic speech recognition (ASR) models have achieved remarkable improvements in recent years, performance disparities persist across different speaker populations. One such disparity is for speakers whose native languages are from families distant from English. This paper investigates the relationship between native language background and ASR performance. Through empirical analysis, we observe that the correlation between typological L1 distance and ASR error rates yields a systematic and dominant effect, with our regression models explaining up to 48% of the variance ($R^2 \leq 0.5$). In addition, analysis of the latent space reveals a progressive intensification of L1-based spatial segregation across deeper acoustic layers in the majority of evaluated architectures.

1 Introduction

Automatic Speech Recognition (ASR) systems are widely deployed in real-world applications. Ensuring robust performance across diverse speaker populations is therefore essential. However, performance disparities across speaker subgroups continue to be reported, raising questions about when and who the systems underperform for?

Prior work has documented performance disparities in ASR across various speaker attributes. Differences have been observed with respect to gender (Attanasio et al., 2024) and racial background (Zolnoori et al., 2024). Beyond these demographic factors, a consistent body of evidence shows that second-language (L2) speakers exhibit higher Word Error Rates (WER) than first-language (L1) speakers across commercial ASR systems (Google, Amazon, Microsoft) (DiChristofano et al., 2022) as well as open-source models such as Whisper (Graham and Roll, 2024). Such performance disparity can also be found in Swedish (Cumbal et al., 2024) and Korean (Na et al., 2024).

These performance gaps are commonly attributed to speaker characteristics, including geographic origin (Jahan et al., 2025) and first-language influence (Chan et al., 2022). Prior research suggests that disparities between L1 and L2 speakers may stem from systematic structural differences between languages. Linguistic research has shown that L1–L2 distance plays an important role in second language processing and acquisition (Kuperman, 2025). Similar effects have been observed in NLP systems, where linguistic proximity impacts cross-lingual transfer performance, such as in automatic abusive language detection (Eronen et al., 2022). These findings suggest that linguistic distance may systematically impact model generalization across languages. However, its role in explaining disparities across ASR remains limited.

In this work, we examine whether L1–L2 ASR performance gaps are systematically associated with the structural Linguistic Distance (LD) between speakers' native language and English. Furthermore, we conduct a comparative analysis across distinct ASR architectures to evaluate their respective sensitivities to L2 speech. We observe that as input representations propagate into deeper layers, all evaluated models progressively segregate L2 acoustic spaces from the native baseline. This systematic spatial separation strongly implies that current state-of-the-art architectures isolate accented features rather than robustly normalizing them, emphasize the need for enhanced accent-invariant representation learning in future acoustic frameworks.

2 Related Work

Several studies have examined the impact of speakers' L1 on ASR performance and showing that recognition accuracy varies with speaker language and accent (Chan et al., 2022; Liu et al., 2025; Ja-

Table 1: Overview of the datasets, including total samples (hours) and the number of accents.

Dataset	N. Samples (hr)	Accents
Speech Accent Archive (Weinberger, 2015)	2,015 (15.5)	17
Fair-Speech (Veliche et al., 2024)	26,471 (54.5)	27
L2-ARCTIC (Zhao et al., 2018)	26,867 (27.1)	6
EdAcc (Sanabria et al., 2023)	17,965 (32.7)	20
ALLSSTAR (Bradlow et al., 2010)	699 (20.9)	22

han et al., 2025). While these studies confirm that performance disparities exist across L1-L2 groups, existing work has primarily focused on documenting the extent of these gaps. Issa et al. (2026) investigated Whisper large-v3's performance on non-native Egyptian Arabic speech and revealed existing correlation across WER and high accentedness (rated by human) but comprehensibility only had a marginal positive association. Bartelds et al. (2021) showed that there exists a correlation between acoustic distances in neural speech representation and human native-likeness judgments. Tsoukala et al. (2026) revealed a clear performance gradient with dialectal distance from Standard Modern Greek. Similar research has been shown in Finnish (Törö et al., 2025). Törö et al. (2025) also applied LDA and demonstrated that language embeddings are clustered. This embedding-based distances are correlated to geography- and lexicon-based distances in world languages. While these studies establish a structural link between embeddings and linguistics, they do not bridge the gap to the performance variances seen in L2-ASR.

3 Methodology

Language linguistic distances: We evaluate two speaker populations: native English speakers (L1) and non-native English speakers (L2) - who acquired English as a second language. We adopt the method proposed by Vincent and Johannes (2020)[1] to quantify how linguistically distant each language is from English. The distance is ranging from 0 to 100 ((identical to completely unrelated languages).

Regression model: We performed a regression analysis[2] with linguistic distance as the main parameter to quantify the impact of speakers' L1 transfer on ASR performance.

$$Y_i^{metric} = \beta_{const} + \beta_{LD} x_{LD,i} + \epsilon \qquad (1)$$

[1] http://www.elinguistics.net/Compare_Languages.aspx

[2] From SciPy package

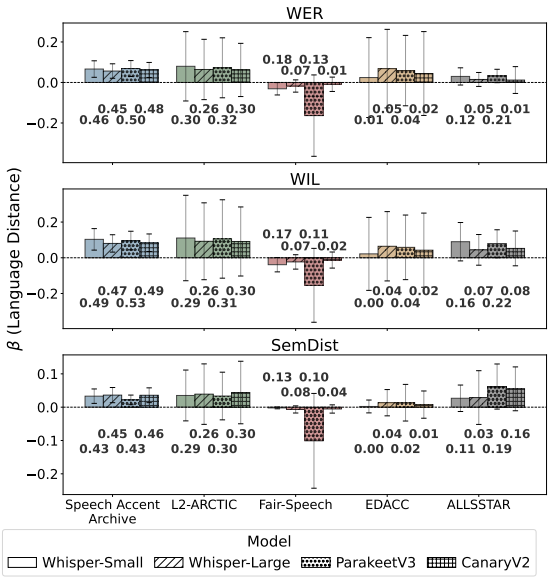


Figure 1: Coefficient (β) for language relative to ASR model performance in WER, WIL and SemDist metrics. Error bars represent 95% confidence intervals, and the value shown is the adjusted R^2. All statistical test assumptions were verified and met.

where Y_i^{metric} is the mean of performance metric evaluated on language i; β_{const} represents constant intercept term, capturing the model's baseline error rate for L1 speakers; $x_{LD,i}$ denotes the LD from language i to English; β_{LD} is the slope coefficient quantifying model's performance sensitivity to increasing linguistic distance; and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the residual error term. Both Y_i^{metric} and $x_{LD,i}$ are normalized via Min-Max scaling[3] to ensure a uniform baseline for comparison.

Embedding distance and correlation: To evaluate how L2 phonetic variants are encoded dynamically within the internal representations across the network's depth, we compute the Spearman rank correlation coefficient (ρ) between LD and the layer-specific (l) Embedding Distance (ED_l) of individual accented utterances relative to a native English baseline. Following similar approaches utilized in recent literature to measure embedding distance (Durrheim et al., 2023; Bolukbasi et al., 2016; Ohamadike et al., 2025) , we apply cosine distance as our core metric. For any given target utterance embedding vector $E_{\text{S},l}$, we calculate its distance to the centroid of the native English reference embeddings, $C(E_{\text{E},l})$, as:

$$ED_l = \frac{C(E_{\text{E},l}) \cdot E_{\text{S},l}}{\|C(E_{\text{E},l})\| \, \|E_{\text{S},l}\|} \qquad (2)$$

[3] From Scikit-learn package

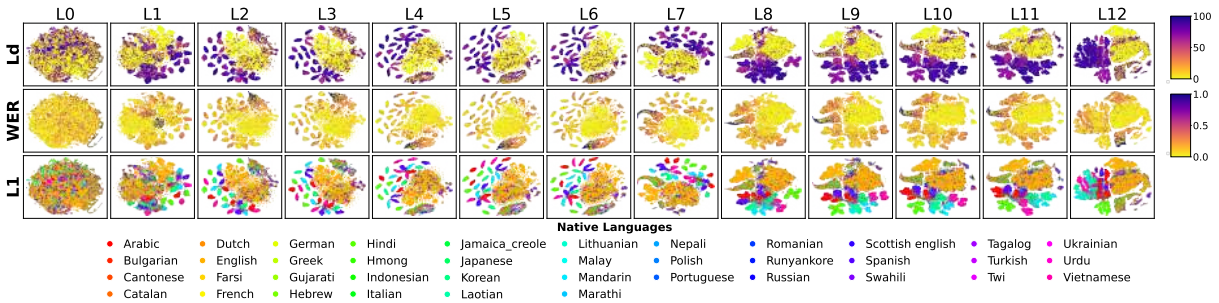


Figure 2: Layer-wise Embedding Analysis of Whisper small: Linguistic clustering intensifies with model depth, with FairSpeech driving high-WER/low-LD observations.

where the native English centroid $C(E_{\mathrm{E},l})$ represents the geometric mean of all L1's utterance embeddings in all evaluating datasets within the corresponding network layer l, calculated as:

$$C(E_{\mathrm{E},l}) = \frac{1}{N_{\mathrm{E}}} \sum_{i=1}^{N_{\mathrm{E}}} E_{\mathrm{E},l,i} \tag{3}$$

Here, N_{E} denotes the total number of L1 speaker utterances, and $E_{\mathrm{E},l,i}$ represents the embedding vector of the i-th L1 speaker sample in layer l. Subsequently, the layer-wise Spearman rank correlation (ρ_l) between the embedding displacement distribution and the structural LD is calculated as:

$$\rho_l = \frac{\mathrm{cov}(R[LD], R[ED_l])}{\sigma_{R[LD]}\sigma_{R[ED_l]}} \tag{4}$$

where $R[LD]$ and $R[ED_l]$ represent the assigned ranks of the LD and ED_l vectors, respectively; $\mathrm{cov}(\cdot)$ denotes the covariance of these ranks; and σ represents their respective standard deviations.

Experiment Setup: We choose to evaluate ASR models that are architecturally diverse to assess bias against L2 speakers, including the transformer-based Whisper (Radford et al., 2022) (Small, 244M, and Large, 1.5B), as well as the Conformer-based Parakeet-TDT-0.6B-v3 and Canary-1B-v2 (Sekoyan et al., 2025). Table 1 shows the selected datasets during our experiment which are chosen because of their rich L2 speakers' native languages, where English is their L2.

Beyond the WER, we incorporate Word Information Loss (WIL) (Morris et al., 2004) to better align our results with human perceptual judgment, and Semantic Distance (SemDist) (Kim et al., 2021) to capture errors that may be lexically significant but semantically minor. SemDist, specifically, uses RoBERTa-base to generate semantic embeddings of the reference and hypothesis transcripts, quantifying their similarity via cosine distance.

4 Results

Correlation between LD and performance: As illustrated in Figure 1, LD exhibits a consistent and statistically significant positive association with all evaluated error metrics ($p < 0.05$) across nearly all datasets, with the sole exception of FairSpeech. This trend demonstrates that ASR systems encounter progressively greater optimization difficulties when processing speech from individuals whose native languages are typologically distant from the target domain. Critically, this pattern remains stable across models leveraging fundamentally different architectures, confirming that increased linguistic distance from English systematically drives higher downstream recognition error rates. Across all linear regressions, LD accounts for between 1% and 50% of the total performance variance. Given that ASR performance is simultaneously shaped by localized speaker-specific variables such as L2 proficiency, age, and gender, an effect size of this magnitude indicates that LD exerts a small-to-moderate yet highly systematic influence on model behavior.

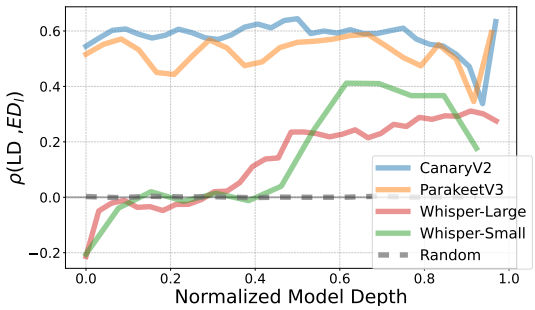
Conversely, Fair-Speech dataset shows negative regression coefficients across all evaluated models and performance metrics, with the total variance explained by LD ranging between 2% and 18%. To investigate this anomaly, we conducted a preliminary human perception study[4] ($n = 4$) to evaluate whether listeners could reliably distinguish between the corpus's designated L1 and L2 speaker cohorts. Our perceptual results reveal that the acoustic boundaries between native and non-native speakers in this dataset are remarkably thin. While the average classification accuracy across evaluators was 58.8%, binomial tests against a chance baseline (50%) demonstrated that only Evaluator

[4]Detail of setup is available in Appendix A. All relevant statistical assumptions were verified and met.

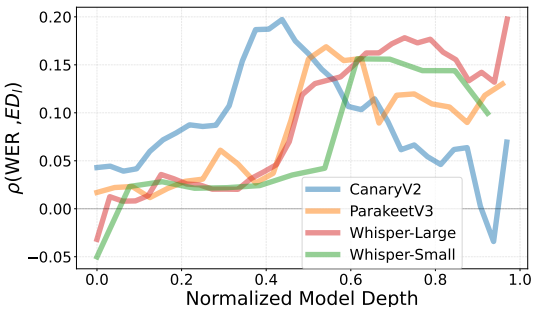
1 ($p = 0.003$) and Evaluator 3 ($p = 0.016$) performed significantly better than random guessing. Furthermore, the low inter-rater reliability (Fleiss' $\kappa = 0.29$) confirms a distinct lack of perceptual consensus among the human annotators. Collectively, these findings suggest that Fair-Speech corpus may suffer from demographic confounding and label noise that limits its utility as an authentic non-native benchmark. A substantial portion of L2 participants represented in this dataset reside within the United States, likely resulting in assimilated speech patterns that conform closely to general American english rather than retaining the distinct phonological markers of their primary language. This structural convergence not only explains the perceptual ambiguity documented in our human validation study but also accounts for the unexpectedly high, native-level ASR performance observed for the L2 cohort within this specific corpus.

Model layer depth influence on accent features: To investigate the origins of the disparity in L1-ASR performance within the latent space, we visualize the embeddings using t-SNE projections to identify potential separation based on speakers' L1. Taking Whisper-small as an example (Figure 2), linguistic clustering becomes significantly more pronounced in the deeper layers of the model, suggesting that language-specific features are increasingly refined as the signal propagates through the neural network. Furthermore, this typological clustering within the embedding space emerges as early as the first few layers, aligning with the higher error rates observed in early-stage representations. This phenomenon is not unique to Whisper architecture. A similar layer-wise spatial separation can also be observed in both Parakeet and Canary[5].

To confirm our hypothesis that LD is encoded across the model layers, and that it increases error rates, we calculate the Spearmans correlation across both LD and ED_l, and LD and Y^{metric}. Figure 3 (a) shows that Whisper, no matter the model size, has increased correlation between LD and ED_l with deeper layers, and that this is also a trend between ED_l and Y^{WER} showing in Figure 3 (b)[6]. As for Parakeet/Canary, they both hold a high correlation across LD and ED_l. However, CanaryV2 shows the lower correlation across performance metrics. It aligns with previous plot in Figure 1, where Canary in general has the lowest coefficient when compared with other models.



(a) Spearman correlation across Linguistic Distance (LD) and Embedding Distance (ED_l).



(b) Spearman correlation across Word Error Rates (Y^{WER}) and Embedding Distance (ED_l).

Figure 3: Evolution of Spearman correlation metrics across network layers. Both embedding space divergence relative to language distance $\rho(LD, ED_l)$ and Word Error Rates $\rho(Y^{WER}, ED_l)$ intensify with layer depth. Due to non-normality, Spearman's rank correlation was used; all statistical assumptions were satisfied.

5 Conclusion

This work explores how Linguistic Distance (LD) impacts Automatic Speech Recognition (ASR) accuracy across different architectures. We demonstrate that as a speaker's native language drifts further from English, ASR performance systematically degrades. A key exception is the Fair-Speech dataset, which we show lacks representative non-native (L2) traits due to its localized, US-resident L2 cohort. By mapping internal model dynamics, we reveal that deeper neural network layers progressively separated native (L1) and L2 acoustic embeddings, proving that most models actively isolate accented traits rather than normalizing them. Crucially, CanaryV2 emerges as a more robust exception within this investigation, demonstrating an advanced capacity for diverse accent.

[5]Due to page limits, the corresponding visualization plots for other models are provided in Appendix B

[6]WIL and SemDist performance is available in Appendix C

6 Limitations

While our study incorporates diverse datasets, the L2 samples from available dataset are still primarily drawn from specific geographic populations, particularly European and East Asian speakers. This leaves other regions, such as Africa, underrepresented in our current analysis, potentially limiting the generalizability of our findings to all marginalized speaker groups. Furthermore, our reliance on LD as a metric may not fully capture the nuance of diverse linguistic backgrounds. Specifically, the speech patterns of multilingual individuals often reflect blended influences that cannot be easily categorized by a single first language L1.

7 Ethics Considerations

Demonstrating that ASR models separate speakers by native language in their latent space reveals a dual-use risk. These internal representations could be extracted without consent to profile or surveil users based on their background. Furthermore, we show that optimizing only for output parity (WER, WIL, SemDist) hides this internal bias, allowing structurally biased models to be deployed while falsely appearing fair.

References

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340.

Martijn Bartelds, Wietse de Vries, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2021. Measuring foreign accent strength using an acoustic distance measure. In *12th International Seminar on Speech Production*, pages 17–20. Haskins Press.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Ann R Bradlow, L Ackerman, L Burchfield, L Hesterberg, J Luque, and K Mok. 2010. Allsstar: Archive of l1 and l2 scripted and spontaneous transcripts and recordings. In *Proceedings of the International Congress on Phonetic Sciences*, volume 356, page 359.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole R Holliday. 2022. Training and typological bias in asr performance for world englishes. In *INTERSPEECH*, pages 1273–1277.

Ronald Cumbal, Birger Moell, José Lopes, and Olof Engwall. 2024. You don't understand me!: Comparing asr results for l1 and l2 speakers of swedish. *arXiv preprint arXiv:2405.13379*.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2022. Global performance disparities between english-language accents in automatic speech recognition. *arXiv preprint arXiv:2208.01157*.

Kevin Durrheim, Maria Schuld, Martin Mafunda, and Sindisiwe Mazibuko. 2023. Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, 62(1):617–629.

Juuso Eronen, Michal Ptaszynski, Fumito Masui, Masaki Arata, Gniewosz Leliwa, and Michal Wroczynski. 2022. Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing & Management*, 59(4):102981.

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).

Elsayed Issa, Mahmoud Ali, and Kevin Hirschi. 2026. Measuring linguistic bias in asr: Whisper large-v3 on non-native speech versus human perception. *Procedia Computer Science*, 275:692–699.

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.

Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael L Seltzer. 2021. Evaluating user perception of speech recognition system quality with semantic distance metric. *arXiv preprint arXiv:2110.05376*.

Victor Kuperman. 2025. How does language distance affect reading fluency and comprehension in english as second language? *Studies in Second Language Acquisition*, 47(3):757–773.

Wei Liu, Yukun Xiong, Bin Hu, and Daehan Kwak. 2025. Evaluating automatic speech recognition models: How well do they handle accents? In *International Conference on the AI Revolution*, pages 447–458. Springer.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Jonghwan Na, Yeseul Park, and Bowon Lee. 2024. A comparative study on the biases of age, gender, dialects, and l2 speakers of automatic speech recognition for korean language. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6.

Nnaemeka Ohamadike, Kevin Durrheim, and Mpho Primus. 2025. Whose voice matters? word embeddings reveal identity bias in news quotes. *EPJ Data Science*, 14(1):30.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

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*.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*.

Tuukka Törö, Antti Suni, and Juraj Šimko. 2025. Neighbors and relatives: How do speech embeddings reflect linguistic connections across the world? *PLoS One*, 20(8):e0330755.

Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos, and Stella Markantonatou. 2026. Extending asr evaluation resources for modern greek dialects. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 210–222.

Tuukka Törö, Antti Suni, Leena Dihingia, Juraj Šimko, and Priyankoo Sarmah. 2025. Exploring dialects with speech embeddings: Insights from two speech databases in assamese and finnish. In *2025 28th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

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

Beaufils Vincent and Tomin Johannes. 2020. Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration.

Steven Weinberger. 2015. Speech accent archive. george mason university.

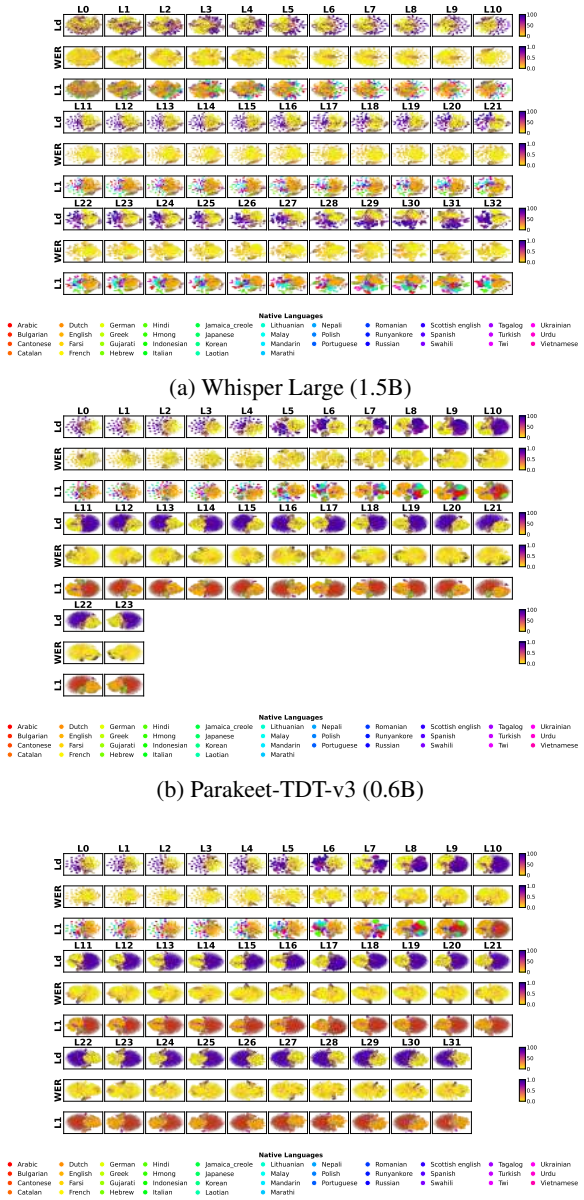
Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, page 2783–2787.

Maryam Zolnoori, Sasha Vergez, Zidu Xu, Elyas Esmaeili, Ali Zolnour, Krystal Anne Briggs, Jihye Kim Scroggins, Seyed Farid Hosseini Ebrahimabad, James M Noble, Maxim Topaz, and 1 others. 2024. Decoding disparities: evaluating automatic speech recognition system performance in transcribing black and white patient verbal communication with nurses in home healthcare. *JAMIA open*, 7(4):ooae130.

A Human Perception Study

To qualitatively verify the Fair-Speech dataset, an informal listening test was conducted with four volunteering participants. Each participant was presented with 100 randomized audio samples, consisting of 50 L1 and 50 L2 speakers. The participants were instructed to label each sample based on their perception of the speaker's language status (L1 vs. L2). Prior to the task, participants were briefed on the study's objective: validating ASR model performance across diverse speaker backgrounds. They were also informed that their responses would be fully anonymized and used strictly for internal result verification, with no individual-level data to be publicly released.

B Comparative Embedding Analysis Across ASR Architectures



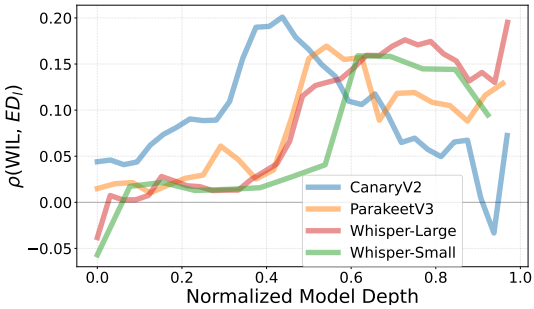
(a) Whisper Large (1.5B)

(b) Parakeet-TDT-v3 (0.6B)

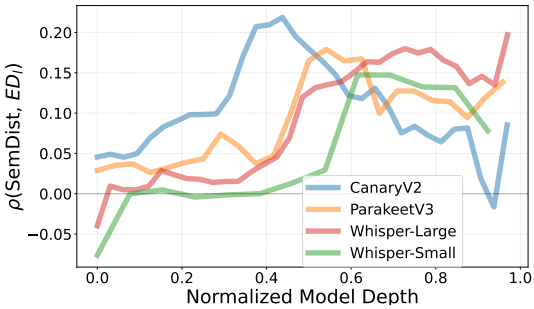
(c) Canary-v2 (1B)

Figure 4: Layer-wise evolution of latent embeddings. t-SNE projections demonstrate the progressive segregation of L2 speech features across different ASR architectures.

C Correlation Analysis for Alternative Performance Metrics



(a) Correlation between Word Information Lost (Y^{WIL}) and Embedding Distance (ED_l).



(b) Correlation between Semantic Distance (Y^{SemDist}) and Embedding Distance (ED_l).

Figure 5: Evolution of Spearman correlation metrics across network layers. The plots illustrate consistent alignment between internal representation shifts and alternative performance metrics (WIL and SemDist)