Linguistic Distance Segregates Latent Representations in Automatic Speech Recognition Systems

Ting-Hui Cheng, Line Katrine Harder Clemmensen and Sneha Das
Department of Applied Mathematics and Computer Science
Technical University of Denmark
{tiche, lkhc, sned}@dtu.dk

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 first languages (L1) are from families distant from English. This paper investigates the relationship between first language background and English ASR performance. Through empirical analysis, we observe that the correlation between speakers' L1 distance and ASR error rates yields a systematic effect on English Speech, with its strength varying across datasets and models. This association is statistically significant in a follow-up analysis accounting for dataset-level variation in Tweedie mixed-effects models ($p < 0.001$ across evaluated models). In addition, analysis of the latent space reveals a L1-based spatial segregation across deeper acoustic layers in the majority of evaluated architectures[1].

1 Introduction

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

In this work, we define L1 speakers as those whose first language is English, and L2 speakers as those who acquired English as a second language. Prior work has documented ASR performance disparities 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, L2 speakers exhibit higher Word Error Rates (WER) than L1 across commercial English ASR systems (Google, Amazon, Microsoft) (DiChristofano et al., 2022) and open-source models such as Whisper (Graham and Roll, 2024). Similar disparities have also been reported for Swedish (Cumbal et al., 2024) and Korean (Na et al., 2024).

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

Dataset	N. Samples (hr)	Accents
Speech Accent Archive (SAA) (Weinberger, 2015)	2,015 (15.5)	17
Fair-Speech (Veliche et al., 2024)	26,471 (54.5)	27
L2-ARCTIC (L2Arc) (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
AFRISPEECH-200 (Afri200) (Olatunji et al., 2023)	6318 (18.77)	108

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 L1-L2 disparities 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 on English speech are systematically associated with structural Linguistic Distance (LD) between speakers' L1 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

[1]More details about data and code are available: https://github.com/tinghui8576/LD-Bias-EngASR

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50849–50859
October 24-29, 2026 ©2026 Association for Computational Linguistics

layers, all evaluated models maintain a clear segregation between L2 acoustic representations and those of English L1 speakers. This systematic spatial separation implies that current state-of-the-art architectures isolate accented features rather than robustly normalizing them, emphasizing 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, showing that recognition accuracy varies with speaker language and accent (Chan et al., 2022; Liu et al., 2025; Jahan et al., 2025). While prior work documents L1-L2 performance disparities, it has largely focused on quantifying these gaps. Issa et al. (2026) investigated Whisper large-v3's performance on L2 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) further showed that language embeddings cluster meaningfully, with distances correlating with geographic and lexical distances across languages, while Prasad and Jyothi (2020) found that accent-related information is largely encoded in early ASR layers. While prior studies link embeddings to linguistic and accent features, their relevance to L2-ASR performance disparities remains unexplored.

3 Methodology

Language linguistic distances: We evaluate two speaker populations: first language speakers in English (L1) and speakers who acquired English as a second language (L2). To quantify the linguistic dissimilarity between each speaker's first language and English, we apply method proposed by Vincent and Johannes (2020)[2], which provides a single composite score of linguistic relatedness between languages for subsequent analysis. The resulting linguistic distance (LD) score ranges from 0 to 100 (identical to completely unrelated languages).

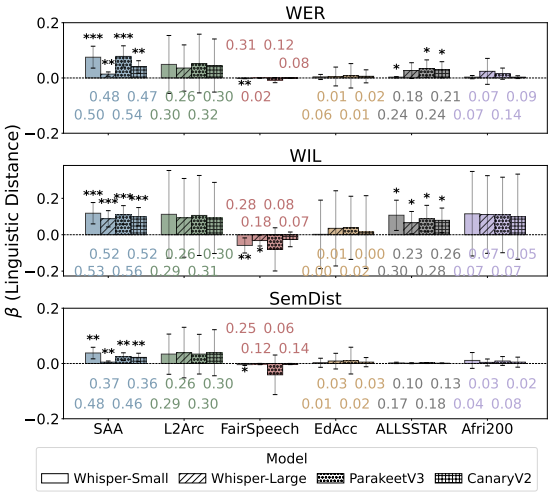


Figure 1: Coefficient (β) for Linguistic Distance (LD) relative to English ASR model performance in WER, WIL and SemDist metrics. Error bars represent 95% confidence intervals, and the value shown is the adjusted R^2 (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Statistical assumptions were verified and met.

Regression model: We use regression analysis[3] to quantify effect of L1–English LD as a measure of L1 transfer on English ASR performance.

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

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[4] to ensure a uniform baseline for comparison.

Generalized Linear Mixed Model: We further investigate the association between speakers' L1 LD from English and ASR performance across datasets by accounting for dataset differences. Towards that, we employ a generalized linear mixed model with a Tweedie likelihood[5] (Ma and Jørgensen, 2007), shown in Equation 2. The Tweedie family is suitable for continuous, non-negative response variables and can accommodate

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

[3] From SciPy package

[4] From Scikit-learn package

[5] From GBBoost package

a point mass at zero through its compound Poisson–Gamma representation, making it appropriate for modeling ASR error-related measures.

$$
\begin{aligned}
Y_{gi}^{\mathrm{WER}} \mid u_g &\sim \mathrm{Tweedie}(\mu_{gi}, \phi, p), \\
\log(\mu_{gi}) &= \beta_0 + \beta_1 x_{LD,i} + u_g, \\
u_g &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2),
\end{aligned} \tag{2}
$$

where Y_{gi}^{WER} denotes the WER for observation i in dataset g, μ_{gi} is the conditional mean, σ_u^2 is the variance component, ϕ is the dispersion parameter, and p is the Tweedie power parameter constrained to $1 < p < 2$. The conditional mean and variance are given by

$$
E\big[Y_{gi}^{\mathrm{WER}} \mid u_g\big] = \mu_{gi}, \quad \mathrm{Var}\big(Y_{gi}^{\mathrm{WER}} \mid u_g\big) = \phi \mu_{gi}^{p}.
$$

The coefficient β_1 represents population-level LD-WER association, while u_g captures dataset-specific deviations from overall intercept. WIL and SemDist trends are reported in Appendix C.

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 L2 individual utterances relative to a English L1 speakers' utterance 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 $E_{\mathrm{S},l}$, we calculate its cosine distance to the centroid of the L1 English reference embeddings, $C(E_{\mathrm{E},l})$, as:

$$
ED_l = 1 - \frac{C(E_{\mathrm{E},l}) \cdot E_{\mathrm{S},l}}{\|C(E_{\mathrm{E},l})\| \, \|E_{\mathrm{S},l}\|} \tag{3}
$$

where the L1 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:

$$
C(E_{\mathrm{E},l}) = \frac{1}{N_{\mathrm{E}}} \sum_{i=1}^{N_{\mathrm{E}}} E_{\mathrm{E},l,i} \tag{4}
$$

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. The

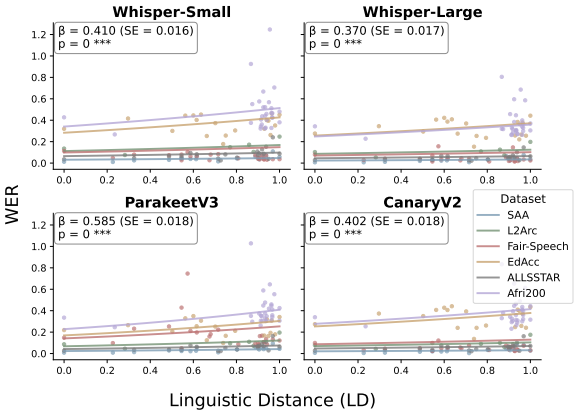


Figure 2: Association between L1's Linguistic Distance (LD) from English and ASR performance (WER). Dots represent mean WER for each L1 group, and lines show the fitted relationships between WER and L1's LD from Tweedie mixed-effects models, with dataset as a random effect. All ASR models show a statistically significant positive effect of LD ($p < 0.001$). Statistical assumptions were verified and met.

layer-wise Spearman rank correlation (ρ_l) between the embedding displacement distribution and the structural LD is calculated as:

$$
\rho_l = \frac{\mathrm{cov}(R[LD], R[ED_l])}{\sigma_{R[LD]} \sigma_{R[ED_l]}} \tag{5}
$$

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 English-capable ASR models that are architecturally different to each other, 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, chosen for their diverse L1 backgrounds among speakers who acquired English as a second language.

Beyond WER, we use 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.

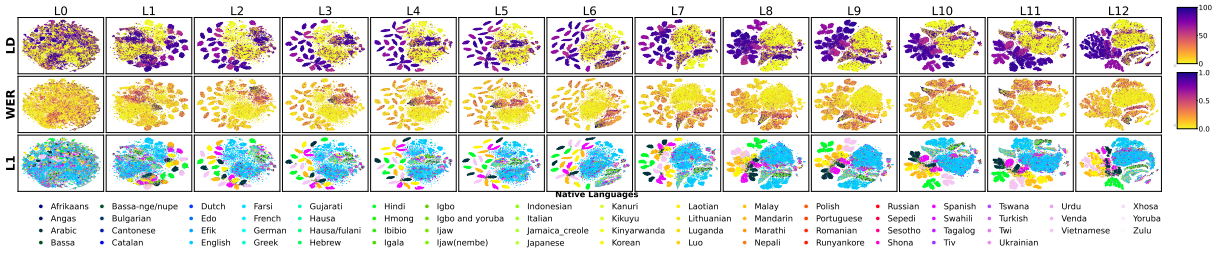


Figure 3: Layer-wise t-SNE visualizations of Whisper Small (perplexity = 30, 1,000 iterations, random state = 20): Linguistic clustering intensifies with model depth, with FairSpeech driving high-WER/low-LD observations.

4 Results

Correlation between LD and performance: As illustrated in Figure 1, LD exhibits a general positive association with all evaluated error metrics across nearly all datasets, with the sole exception of Fair-Speech. The association is particularly pronounced and statistically significant for the SAA dataset. This pattern suggests that ASR performance tends to deteriorate as the LD between speakers' L1 and English increases. Critically, this pattern remains stable across different model architectures, confirming that increased LD from English systematically drives higher downstream recognition error rates. Across all linear regressions, LD accounts for between 1% and 56% of the total performance variance. However, LD is an explanatory factor and its explanatory power varies substantially across datasets and model settings: LD explains a larger proportion of the variance in settings with higher adjusted R^2 values, while its explanatory power is weaker for datasets such as EdAcc or Afrispeech-200, (see Appendix A for further discussion). Overall, these results suggest that LD exerts a small-to-moderate yet systematic influence on model errors, given that ASR performance is simultaneously shaped by speaker-specific variables such as L2 proficiency, age, and gender.

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 31%. To investigate this anomaly, we conducted a preliminary human perception study[6] ($n = 4$) 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 L1 and L2 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 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. However, this should be interpreted as a preliminary perceptual check of whether the dataset's designated L1/L2 distinction was audible to listeners under our experimental task, rather than as an expert validation of speakers' linguistic backgrounds or the accuracy of the dataset labels.

Collectively, these findings suggest that Fair-Speech corpus may contain substantial demographic and sociolinguistic variation that influences the observed relationship between LD and ASR performance. Since all participants, including L2 speakers, are recruited in the United States, their speech may exhibit greater convergence toward general American English. This potentially reduce the retention of the distinct phonological markers associated with their primary language, which may contributing to the perceptual ambiguity documented in our human validation study. At the same time, the substantial diversity within the corpus, including variation among English L1 speakers, suggests that factors beyond L1/L2 status may contribute to the observed WER patterns and may also explain the negative LD coefficients. Therefore, we view Fair-Speech as a counterexample demonstrating that English L1/L2 status and LD are not necessarily the primary axes of ASR disparity in every corpus.

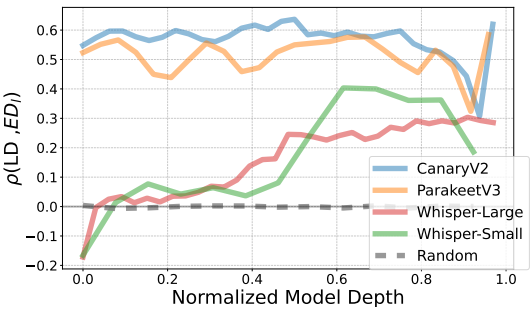
LD-ASR Association Across Datasets: To account for systematic differences in ASR performance across datasets arising from differences in recording conditions and task design,

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

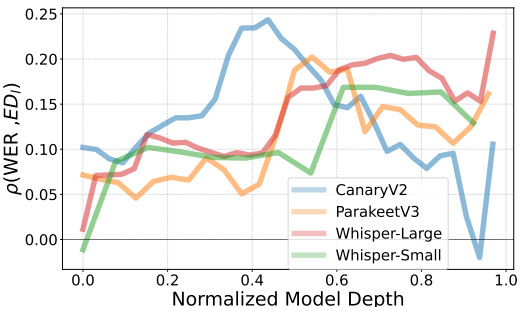
we further fitted Tweedie mixed-effects models with dataset-specific random intercepts. Results (Fig. 2) are consistent with our previous analysis, showing a positive and statistically significant association between LD and WER across models. For example, in Whisper-Small, LD has positive association with WER ($\beta = 0.410$, $SE = 0.016$, $z = 25.20$, $p < 0.001$). Given the log link, this corresponds to an approximately 52% increase in expected WER for a one-unit increase in LD, holding the dataset-specific random effect constant. The estimated dataset-level random-effect variance is $\hat{\sigma}_u^2 = 0.9997$ ($SE = 0.702$). Overall, this provides additional support for the association between LD and ASR error persists after accounting dataset-level variation[7].

Model layer depth influence on accent features: To investigate the origins of the disparity in LD-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 3), linguistic clustering becomes increasingly pronounced in the deeper layers and remains evident in the final layers, indicating that L1-related structure is preserved in the representations. Furthermore, this 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[8].

To examine our hypothesis that LD is encoded across the model layers and is associated with ASR errors, we calculate the Spearmans correlation across both LD and ED_l, and LD and Y^{metric}. Figure 4 (a) shows that Whisper, no matter the model size, has correlation between LD and ED_l becomes pronounced from the early-to-middle layers and remains evident in the later layers. This pattern is also reflected in the relationship between ED_l and Y^{WER} showing in Figure 4 (b)[9]. As for Parakeet/Canary, they both hold a high correlation across LD and ED_l. However, CanaryV2 shows lower correlation across performance metrics. It aligns with Figure 1, where Canary generally has the lowest coefficient comparing to 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 4: Evolution of Spearman correlation metrics across network layers. Embedding space divergence relative to language distance $\rho(LD, ED_l)$ and Word Error Rates $\rho(Y^{\text{WER}}, ED_l)$ 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 English Automatic Speech Recognition (ASR) accuracy across different architectures. We reveal that as a speaker's first language (L1) LD from English increase, ASR performance systematically degrades. This relationship remains after accounting for dataset-level variation in mixed-effects models. Fair-Speech provides a counterexample, highlighting that L1/L2 status and LD are not always the primary axes of the disparity. By mapping internal model dynamics, we show that L1/L2-related separation in acoustic embeddings persists across deeper neural network layers, suggesting that models retain rather than normalize accent-related variation. Crucially, CanaryV2 emerges as a notable exception on within this investigation, indicating greater robustness to accent diversity.

[7]The association also holds for WIL and SemDist; see Appendix C for results

[8]Other models visualizations are provided in Appendix D

[9]WIL and SemDist results are available in Appendix E

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 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. Dataset-provided or self-reported L1 labels can be ambiguous and may not provide a complete representation of a speaker's linguistic history. Specifically, the speech patterns of multilingual individuals often reflect blended influences that cannot be easily categorized by a single first language L1, but has been filtered out during the analysis. Treating L1 and LD as primary explanations for ASR disparities may therefore oversimplify complex linguistic histories and overlook substantial within-corpus variability, including L2 proficiency, age of acquisition and exposure to English. These factors may contribute to variation in the observed results and influence the relationship between LD and ASR performance, potentially affecting our conclusions.

Another limitation of this work is the lack of a universally standardized measure of LD. Although the approach proposed by Vincent and Johannes (2020) provides a single composite relatedness score that is convenient for our regression analysis, other commonly used resources, such as URIEL/lang2vec and WALS, capture a broader range of linguistic features but require additional choices regarding feature selection and aggregation. Consequently, the observed relationship between LD and ASR performance may depend, to some extent, on the particular distance measure adopted.

In our study, we focus on compare commonly used open-source ASR model. Although the selected models are multilingual or multilingual-capable, they are not designed the same way as massively multilingual system such as Omnilingual, or commericial system. Those models trained with broader multilingual coverage may encode L1-related variation differently and may reduce or reshape the disparities observed in our selected model.

Finally, since our empirical analysis is fully focuses on English speech, further investigation is still needed before generalize the effect of language distance to other target languages ASR system or broader setting.

Ethical Considerations

Demonstrating that ASR models separate speakers by L1 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.

Acknowledgments

This work is partly funded by the Innovation Fund Denmark, grant no. 4354-00015B and the Carlsberg Foundation, grant no. CF24-2181.

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.

Renjun Ma and Bent Jørgensen. 2007. Nested generalized linear mixed models: an orthodox best linear unbiased predictor approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):625–641.

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.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and 1 others. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Archiki Prasad and Preethi Jyothi. 2020. How accents confound: Probing for accent information in end-to-end speech recognition systems. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3739–3753.

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.

Jan Philip Wahle, Terry Ruas, Saif M Mohammad, Norman Meuschke, and Bela Gipp. 2023. Ai usage cards: Responsibly reporting ai-generated content. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 282–284. IEEE.

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.

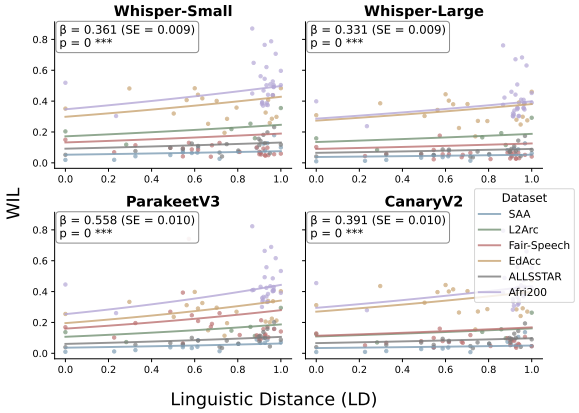
A Weak Explanatory Power of LD

LD has weaker explanatory power in datasets such as EDACC and Afrispeech-200, where additional sources of variability may influence ASR performance beyond linguistic distance. EDACC contains spontaneous dyadic conversations that includes speaker overlap and background noise, while Afrispeech-200 includes clinical domain speech, for which ASR systems have been shown to perform worse than on more general-domain speech (Olatunji et al., 2023). These dataset-specific characteristics reinforce the finding that systematic variation observed in English ASR performance cannot be attributed to LD alone. Instead, the explanatory power of LD should be interpreted alongside dataset design, sociolinguistic and demographic variation, multilingual speaker histories, and other corpus-level factors.

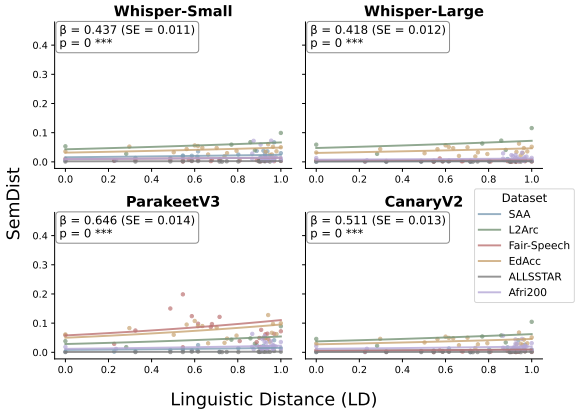
B Human Perception Study

To qualitatively verify the Fair-Speech dataset, an informal listening test was conducted with four volunteering participants based in Denmark with an English proficiency of at least C1. 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 solely on their perception of the speaker's language status, where L1 referred to speakers whose first language was English and L2 referred to speakers whose first language was not English. 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.

C Tweedie Mixed Effect Models for Alternative Performance Metrics



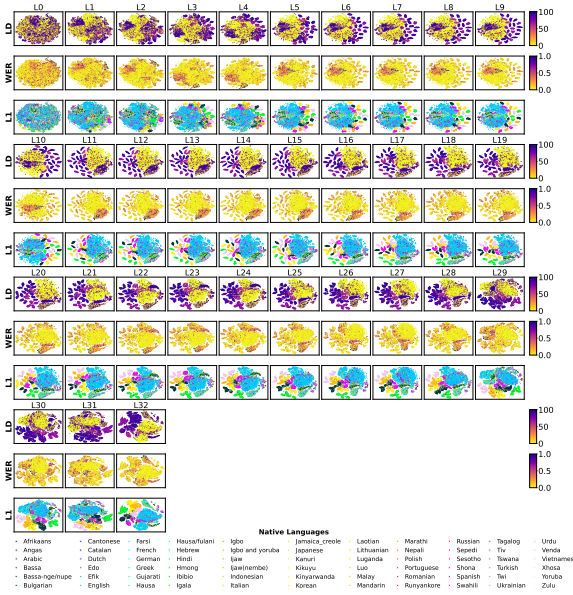
(a) Association between L1's Linguistic Distance (LD) from English and ASR performance in Word Information Lost (Y^{WIL}).



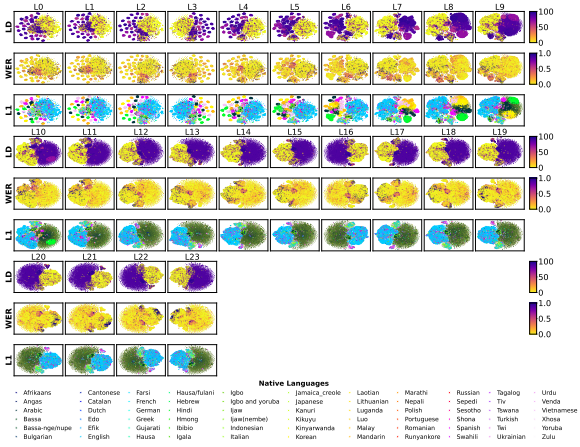
(b) Association between L1's Linguistic Distance (LD) from English and ASR performance in Semantic Distance (SemDist).

Figure 5: The association between LD and ASR performance is consistent across alternative evaluation metrics, WIL and SemDist.

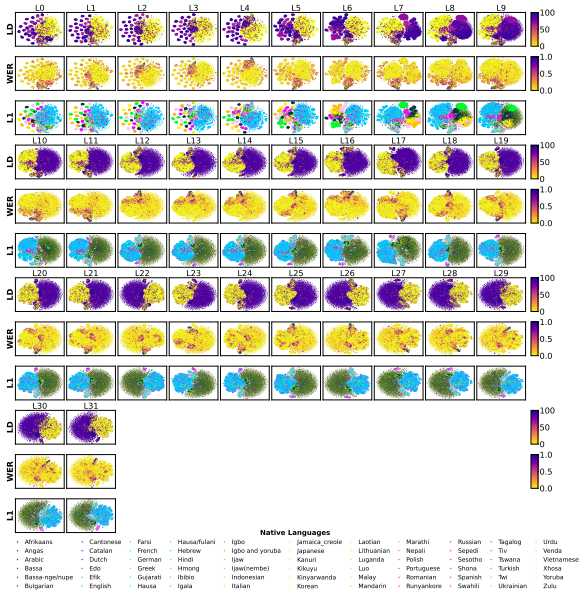
D Comparative Embedding Analysis Across ASR Architectures



(a) Whisper Large (1.5B)



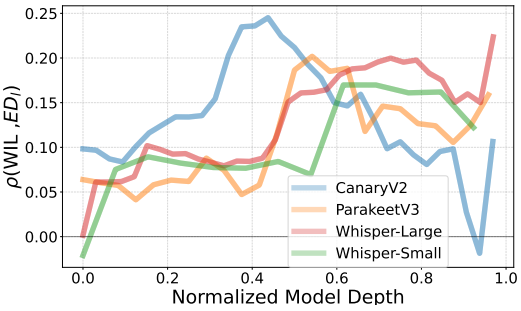
(b) Parakeet-TDT-v3 (0.6B)



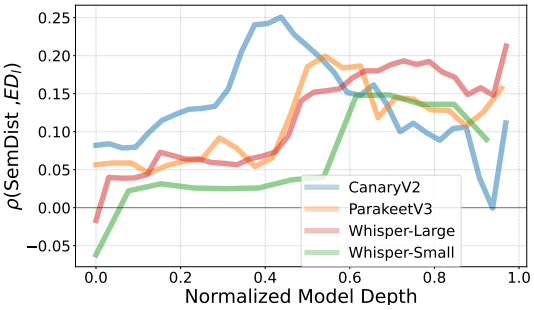
(a) Canary-v2 (1B)

Figure 7: Layer-wise evolution of latent embeddings. t-SNE (perplexity = 30, 1,000 iterations, random state = 20) projections demonstrate the progressive segregation of L2 speech features across different ASR architectures. The observed patterns are consistent across alternative t-SNE settings (perplexity = 10, random state = 49; perplexity = 30, random state = 42).

E 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 8: Evolution of Spearman correlation metrics across network layers. The plots illustrate consistent alignment between internal representation shifts and alternative performance metrics (WIL and SemDist)

F AI Usage card based on (Wahle et al., 2023)

AI Usage Card

CORRESPONDENCE(S)
Ting-Hui Cheng

CONTACT(S)
tiche@dtu.dk

AFFILIATION(S)
Technical University of Denmark

PROJECT NAME
Linguistic Distance Segregates Latent Representations in Automatic Speech Recognition Systems

KEY APPLICATION(S)
Evaluation of ASR robustness

MODEL(S)
ChatGPT
Gemini
Claude

VERSION(S)
5.1, 5.6 Luna
2.5 Pro, 3 Pro
Sonnet 5

WRITING
ChatGPT, Gemini

GENERATING NEW TEXT BASED ON INSTRUCTIONS
Not used

ASSISTING IN IMPROVING OWN CONTENT
When existing text is paraphrased or improved.

PARAPHRASING RELATED WORK
Not used

PUTTING OTHER WORKS IN PERSPECTIVE
Not used

CODING
Claude,ChatGPT, Gemini

GENERATING NEW CODE BASED ON DESCRIPTIONS OR EXISTING CODE
When new code is generated based on instructions or prompts.

REFACTORING AND OPTIMIZING EXISTING CODE
When existing or generated code is refactored or its performance optimized.

COMPARING ASPECTS OF EXISTING CODE
Not used

ETHICS

WHAT ARE THE IMPLICATIONS OF USING AI FOR THIS PROJECT?
To facilitate academic writing and improve efficiency.

WHAT STEPS ARE WE TAKING TO MITIGATE ERRORS OF AI FOR THIS PROJECT?
All AI-assisted content is critically reviewed and verified by the authors.

WHAT STEPS ARE WE TAKING TO MINIMIZE THE CHANCE OF HARM OR INAPPROPRIATE USE OF AI FOR THIS PROJECT?
Only to support writing and coding assistance and not used to generate experimental results or scientific conclusions.

THE CORRESPONDING AUTHORS VERIFY AND AGREE WITH THE MODIFICATIONS OR GENERATIONS OF THEIR USED AI-GENERATED CONTENT
Yes

AI Usage Card v1.0 https://ai-cards.org PDF | BibTeX | XML | JSON | CSV