

Accent Bias at Scale: Word Error Rate Disparities Between Indian and American English in OpenAI's Whisper

Arnav Jagetiya

DOI:10.37648/ijrst.v16i03.006

¹ Received: 20 June 2026; Accepted: 13 August 2026; Published: 06 September 2026

Abstract

Automatic speech recognition (ASR) systems are evaluated by an aggregate word error rate (WER) that conceals which speakers a model fails on. In this paper, we measure that hidden disparity for OpenAI's Whisper, comparing its transcription accuracy on Indian-accented and American-accented English. Sixty speakers from each group read an identical passage from the Speech Accent Archive, so accent is the only variable between conditions. We transcribed each recording with four Whisper sizes (tiny, base, small and medium), computed per-clip WER using jiwer, and tested group differences with a one-sided Mann-Whitney U -test. We also ran the same corpus through Meta's MMS-1B-all and NVIDIA's Parakeet to test whether the gap is specific to Whisper.

Indian-accented speech produced a significantly higher WER at every model size, from 8.86 percent against 4.52 percent on tiny to 5.65 percent against 1.09 percent on medium. Scaling reduced absolute error for both groups but widened the relative gap, from a factor of 1.96 on tiny to 5.20 on medium, and the disparity persisted across both alternative architectures. Our error analysis shows that the largest single contributor is orthographic rather than phonetic: on Indian audio, Whisper repeatedly rendered spoken number words as numerals, an error absent from American clips. Once we normalize numerals, a smaller but still measurable gap remains, driven by phonetic confusions such as "slabs" transcribed as "slaps". These results indicate that accent bias in Whisper is a property of scale, not a deficiency that scale removes, and point toward targeted fine-tuning as the more promising remedy.

Introduction

Automatic Speech Recognition (ASR) is present everywhere around us nowadays. From voice-based virtual assistants like Siri to live captions and call recordings, ASR has become deeply embedded in daily life. ASR models built by dominant companies like IBM, Google, and Microsoft are trained on advanced infrastructure with massive amounts of data, time, and money invested into them. Consequently, these models report low aggregate Word Error Rates (WER) of 5 to 8 percent, which has become the undisputed industry standard for evaluating ASR performance (Xiong et al. 2017, 2410). However, this single aggregate number conceals exactly who the model fails on. Research has shown that these same commercial systems exhibit an average WER of around 35% for Black speakers, compared to just 19% for white speakers (Koenecke et al. 2020, 7684). This demonstrates that ASR error is not randomly distributed; it

¹ How to cite the article: Jagetiya A.; September 2026; Accent Bias at Scale: Word Error Rate Disparities Between Indian and American English in OpenAI's Whisper; *International Journal of Research in Science and Technology*, Vol 16, Issue 3, 66-78, DOI: <http://doi.org/10.37648/ijrst.v16i03.006>

disproportionately burdens certain groups over others. In this paper, we build on this idea of hidden disparity, narrowing the focus specifically to the WER gap between Indian English speakers and American English speakers, using one of the most widely deployed ASR models today: Whisper.

There are several underlying reasons for this disparity in WER across different speaker groups. The foundation of any ASR system is its acoustic model, which learns phoneme patterns directly from its training data (Huang and Deng 2010). As a speaker's accent shifts away from the dominant accent represented in that training data, the corresponding phoneme patterns shift as well, toward patterns the acoustic model was never exposed to during training (Faria 2006, 285). As a result, these unfamiliar phonemes get misread. For instance, the word “very” might be transcribed as “wary” when spoken with certain accents. Crucially, this happens even when the exact same sentence, in the exact same language, is spoken by two different speakers; the resulting WER gap arises purely from an underlying accent bias in the acoustic model itself, not from any difference in the linguistic content being spoken (Sailaja 2009).

Indian English speakers represent one of the largest English-speaking populations in the world and yet remain classified as a “low-resource” accent group in most ASR training pipelines. Large-scale commercial models from IBM, Google, and Microsoft have all been shown to struggle with Indian-accented English, producing visibly elevated WER of 12 to 22 percent compared to American English. This raises an important question: can simply scaling up training data inherently resolve accent bias and produce a more genuinely accurate aggregate WER, or does the disparity persist regardless of scale? This research paper investigates exactly that question using Whisper, one of the most widely available large-scale ASR models, trained on 680,000 hours of audio across 96 languages (Radford et al. 2023).

A word error rate that is five points higher has real consequences for the speakers it affects. Speech recognition already screens job applicants through automated interviews, transcribes clinical notes, and routes emergency and customer service calls, and in each case a failed transcript transfers the cost of the error onto the person speaking. No major vendor currently publishes error rates disaggregated by accent, so this cost stays invisible to the people paying it. This study demonstrates that such an audit requires no proprietary access and no specialized hardware: a controlled comparison can be run on a consumer laptop using open models and a public corpus. Because the tools for such an audit are freely available, the absence of accent-level error reporting reflects a choice by developers rather than a technical limitation.

Related Works

ASR performance is reported as a single word error rate on shared benchmarks, a convention that produced the milestone claim of human parity on Switchboard (Xiong et al. 2017) and, more recently, Whisper's assertion that training on 680,000 hours of web audio yields accent robustness that improves with model size (Radford et al. 2023). Koenecke et al. (2020) showed what that aggregate conceals: five commercial systems produced substantially higher error rates for Black speakers than for white speakers, and because the gap persisted on identical phrases spoken by both groups, the cause was traced to the acoustic model rather than to vocabulary or grammar. Faria (2006) demonstrated that non-native WER falls when accent-specific models are used, and Ike, Polsley, and Hammond (2022) found consistently lower accuracy for under-represented English accents across popular commercial systems.

Graham and Roll (2024) evaluated its largest model on the Speech Accent Archive and found native accents recognized more accurately than non-native ones, with American English performing best. Indian English in particular remains under-served: Javed et al. (2023) note that the Speech Accent Archive contains only 59 Indian speakers, and their Svarah benchmark shows that every ASR system tested, Whisper included, degrades substantially on Indian accents relative to native-speaker data.

No prior study combines these strands. Koenecke et al. controlled content but tested neither Whisper nor Indian

English; Graham and Roll tested Whisper on the same corpus used here but at one model size and without isolating Indian speakers; Javed et al. isolated Indian English but with content that varies across speakers. This study holds the spoken passage fixed across sixty Indian and sixty American speakers, sweeps four Whisper sizes, and repeats the comparison on two alternative architectures, so that the disparity, its response to scale, and its dependence on architecture can be measured with accent as the only variable.

Methodology

Dataset. All audio comes from the Speech Accent Archive (Weinberger 2015), in which every speaker reads the same 69-word elicitation paragraph beginning “Please call Stella”. Sixty recordings were selected for each group. The American group comprises native English speakers born in the United States; the Indian group comprises speakers whose first language is an Indian language, Hindi, Telugu, Marathi and Tamil. Because the passage is identical across all 120 recordings, lexical content, syntax and sentence length are held constant, and accent remains the only systematic difference between groups. Clips were downloaded as MP3 and converted to 16 kHz mono WAV with ffmpeg. A validation script confirms file counts and that each clip’s reference transcript is the canonical passage.

Models. Four sizes of OpenAI’s Whisper were evaluated: tiny (39M parameters), base (74M), small (244M) and medium (769M). Large and large-v3 were excluded because they exceed the memory available on the local machine. To test whether the disparity is specific to Whisper, the same 120 recordings were transcribed by two architecturally distinct systems: Meta’s MMS-1B-all, a wav2vec 2.0 encoder with a CTC head, and NVIDIA’s Parakeet TDT 0.6B v3, a FastConformer model. These two runs were executed in Google Colab on a GPU because their weights do not fit comfortably on the local machine.

Evaluation. Transcripts and references were normalized identically before scoring: lowercased, with punctuation replaced by whitespace and apostrophes retained. WER was computed per clip using jiwer (Vaessen 2023) and averaged within each group, so that every speaker contributes equally regardless of transcript length. Because neither MMS nor Parakeet emits numerals, a secondary pass mapped digits in Whisper output back to their spoken word forms, allowing a like-for-like cross-architecture comparison and separating orthographic from phonetic error. Group differences were tested with a one-sided Mann-Whitney U test under the directional hypothesis that Indian-accented WER exceeds American-accented WER, with Cohen’s d reported as the effect size. Substitution pairs were extracted from jiwer’s word alignments and counted by group to identify the words most sensitive to accent.

Software environment. All Whisper runs, scoring and figure generation were performed locally in Python, using pandas for data handling, SciPy for statistics and Matplotlib for figures. Results for each model are written to CSV with ground truth and prediction side by side, and every figure and table in this paper is regenerated from those CSVs by a single script so that the analysis is fully reproducible from the repository.

Interactive dashboard. Alongside the paper, the results are published as an interactive web application. It presents the WER comparison across model sizes, error-type breakdowns, phrase-level sensitivity, a scroll-driven walkthrough of the Whisper pipeline from waveform to mel spectrogram to decoder, and a live WER calculator that runs Whisper in the browser. The dashboard is intended to make the disparity legible to non-technical readers without requiring them to run the code.

Experimental Set Up

Datasets. The corpus consists of 120 recordings from the Speech Accent Archive, 60 per group, each a reading of the same 69-word elicitation passage. Table 1 summarizes the two groups. American speakers were selected on birthplace within the United States and English as first language; Indian speakers on birthplace within India and an Indian language

as first language. Every recording was scored against the same reference transcript, so the denominator of every WER calculation in this paper is 69. Because the Archive is read speech, results reflect a controlled and relatively favorable condition for ASR; spontaneous speech would likely widen the gaps reported here.

Table 1. Speaker groups drawn from the Speech Accent Archive.

Group	n	Female / Male	Age range	First language(s)
American English	60	25 / 35	18–60	English
Indian English	60	28 / 32	18–60	Hindi, Telugu, Tamil, Marathi

Models. Six systems were evaluated on the identical 120 clips. Four are Whisper checkpoints of increasing size: tiny (39M parameters), base (74M), small (244M) and medium (769M). All were run with the language forced to English and the task set to transcription, so that no error arises from language misidentification. Two are architecturally distinct baselines: Meta’s MMS-1B-all, a wav2vec 2.0 encoder with a CTC head (Pratap et al. 2024), and NVIDIA’s Parakeet-TDT-0.6B-v3, a FastConformer with a token-and-duration transducer (NVIDIA 2025). Whisper ran locally; MMS and Parakeet ran in Google Colab. For every clip and model, the raw transcript was stored in a CSV alongside the reference so that all statistics could be recomputed offline.

Worked example: word error rate.

WER counts the minimum number of substitutions (S), deletions (D) and insertions (I) needed to turn the model’s output into the reference, divided by the reference length N :

$$\text{WER} = \frac{S + D + I}{N} \quad (1)$$

One American clip on whisper-tiny dropped two words and made no other error, so $S = 0$, $D = 2$, $I = 0$ and $\text{WER} = 2/69 = 2.90\%$. A clip transcribed perfectly scores 0. WER can exceed 100% when insertions outnumber reference words, which happened once in this study.

Worked example: mean and standard error. Per-clip WERs are averaged within each group. On whisper-tiny the 60 American clips have mean 4.52% and sample standard deviation 4.77%; the 60 Indian clips have mean 8.86% and standard deviation 11.44%. The standard error of each mean is $s/\sqrt{n}$:

$$\text{SE}_{\text{US}} = \frac{4.77}{\sqrt{60}} = 0.62 \text{ pp}, \quad \text{SE}_{\text{India}} = \frac{11.44}{\sqrt{60}} = 1.48 \text{ pp} \quad (2)$$

These are the values in parentheses in Table 3. The larger Indian SE reflects the heavier tail of that distribution rather than a smaller sample.

Worked example: gap and ratio. The absolute gap is the difference in group means, reported in percentage points (pp) to distinguish it from the WERs themselves, which are percentages of reference words:

$$\text{Gap} = 8.86 - 4.52 = 4.35 \text{ pp}, \quad \text{Ratio} = \frac{8.86}{4.52} = 1.96\times \quad (3)$$

A gap in percentage points can hold steady while the ratio rises, which is exactly what Figure 2 shows as model size increases.

Worked example: significance. Per-clip WER is bounded at zero and right-skewed, so a t -test is inappropriate. The Mann-Whitney U test was used instead, with the one-sided alternative that Indian WER is greater. All 120 values are pooled and ranked; tied values receive the average of the ranks they span. On whisper-tiny the Indian ranks sum to $R = 4,289.5$, giving

$$U = R - \frac{n_1(n_1 + 1)}{2} = 4,289.5 - 1,830 = 2,459.5 \quad (4)$$

Under the null hypothesis U has mean $n_1 n_2 / 2 = 1,800$ and, after correcting for ties, standard deviation 189.2.

The standardized statistic is

$$z = \frac{2,459.5 - 1,800}{189.2} = 3.49, \quad (5)$$

and the one-sided upper-tail probability is $p = 2.5 \times 10^{-4}$. SciPy's implementation, which applies the same tie correction, returns the identical value. The same procedure yields every p -value in Table 3.

Results

Whisper produces higher word error rates for Indian English at every model size. Across 120 recordings of the same 69-word passage for different speakers and accents, Whisper transcribed American-accented English more accurately than Indian-accented English in all four model sizes tested (Figure 1). On whisper-tiny, mean WER was 4.52% (SE 0.62) for American speakers and 8.86% (SE 1.48) for Indian speakers. On whisper-base, 3.26% (SE 0.51) versus 7.34% (SE 0.56). On whisper-small, 1.52% (SE 0.27) versus 4.76% (SE 0.53). On whisper-medium, 1.09% (SE 0.21) versus 5.65% (SE 1.68). Every comparison was significant by one-sided Mann-Whitney U test ($p = 2.5 \times 10^{-4}$, 3.3×10^{-8} , 1.9×10^{-7} , and 9.5×10^{-8} respectively).

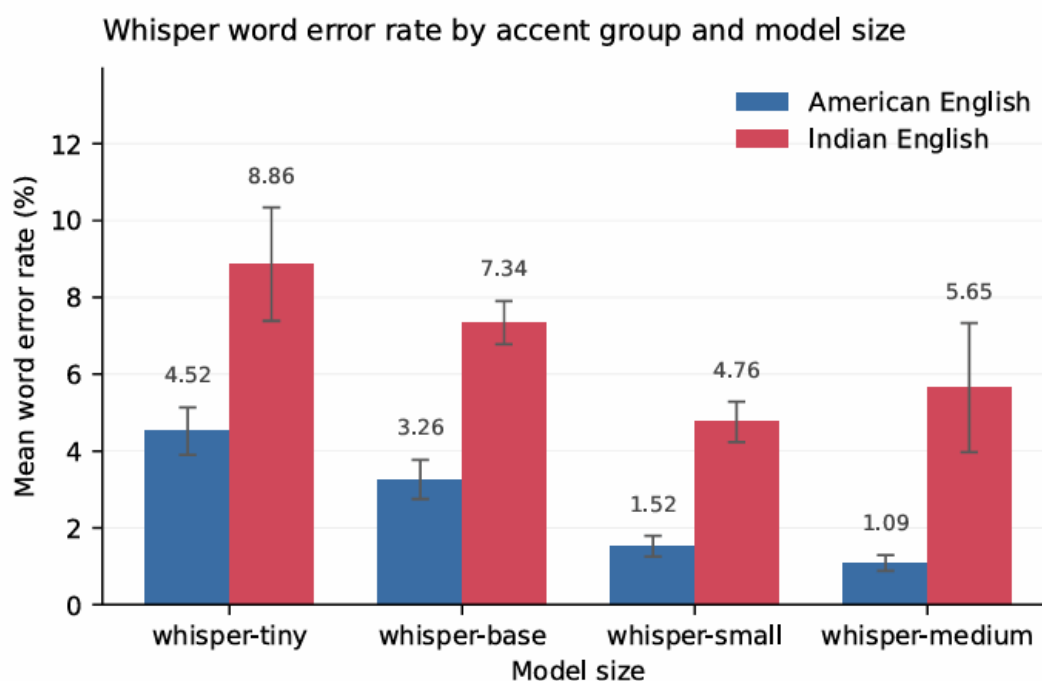


Figure 1. WER by accent group across Whisper model sizes. Bars show mean per-clip WER; error bars show one standard error ($n = 60$ per group).

Scale reduces absolute error but does not close the gap. Moving from tiny to medium cut American WER by 76% and Indian WER by 36%. The absolute gap, however, remained essentially flat across the entire range, at 4.35, 4.08, 3.24, and 4.57 percentage points for tiny, base, small, and medium. Because the American baseline fell faster than the gap, the disparity ratio widened monotonically with scale, from 1.96 $\times$ at tiny to 2.25 $\times$ at base, 3.13 $\times$ at small, and 5.20 $\times$ at medium (Figure 2). A tenfold increase in the number of parameters improved recognition for both groups but delivered that improvement unequally, leaving Indian speakers relatively further behind than they began.

The distribution of errors is more unequal than the means suggest. At whisper-medium, 34 of 60 American recordings of different speakers were transcribed with zero errors, against 16 of 60 Indian recordings. The Indian group also carried a heavier upper tail: 90th-percentile WER at medium was 3.04% for American

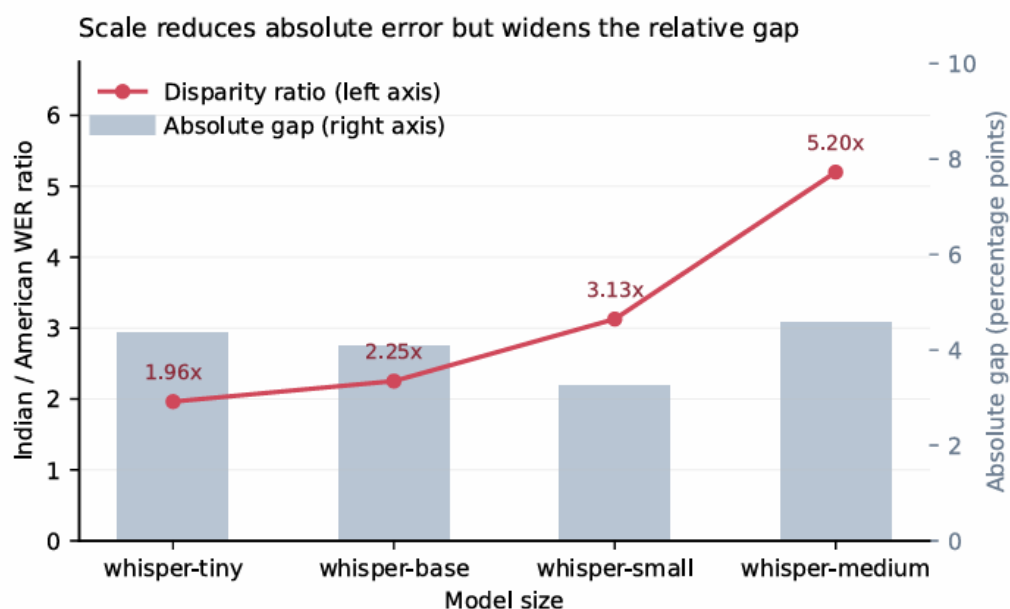


Figure 2. Disparity ratio as a function of model size. The line shows Indian WER divided by American WER; the bars show the absolute gap in percentage points.

speakers and 8.70% for Indian speakers. Two Indian clips failed catastrophically, one reaching 85.5% WER at tiny and another 101.5% at medium, the latter exceeding 100% because the model inserted more words than the reference contained. No American clip in any condition exceeded 21%. Whisper-medium logged 80 insertions on Indian audio against 10 on American audio, indicating that the model's failure mode on unfamiliar acoustics is hallucination rather than omission. Excluding the single hallucinated clip lowers Indian medium WER from 5.65% to 4.03%, which does not eliminate the gap but shows how much of the headline mean is carried by a small number of severe failures (Figure 3).

A large share of the measured gap is orthographic, not acoustic. Inspecting substitution pairs revealed that the three most frequent errors on Indian audio at every model size above tiny were “six” transcribed as “6”, “five” as “5”, and “three” as “3” (Table 2). These are formatting decisions, not misrecognitions: the model heard the word correctly and rendered it as a numeral. The behavior was strongly asymmetric. At whisper-small, 30 of 60 Indian clips contained digit output against 0 of 60 American clips; at base, 41 versus 6; at medium, 32 versus 0. Re-scoring with numerals mapped back to words reduced Indian WER at small from 4.76% to 2.66% and at medium from 5.65% to 3.38%, leaving American figures unchanged. The gap narrowed but survived, to 1.75× at small ($p = 0.025$) and 1.77× at tiny ($p = 0.003$). The remaining substitutions were genuinely phonetic, clustering on voiceless fricatives and consonant clusters: “slabs” heard as “slaps” (9 occurrences at medium), “snake” as “snack”, “frog” as “frock”, and “snow” as “nose”.

This result carries a methodological consequence beyond this study. Standard WER penalizes orthographic convention identically to genuine error, so an ASR system that merely formats non-native speech differently will register as less accurate. Any audit reporting raw WER without number normalization will overstate the

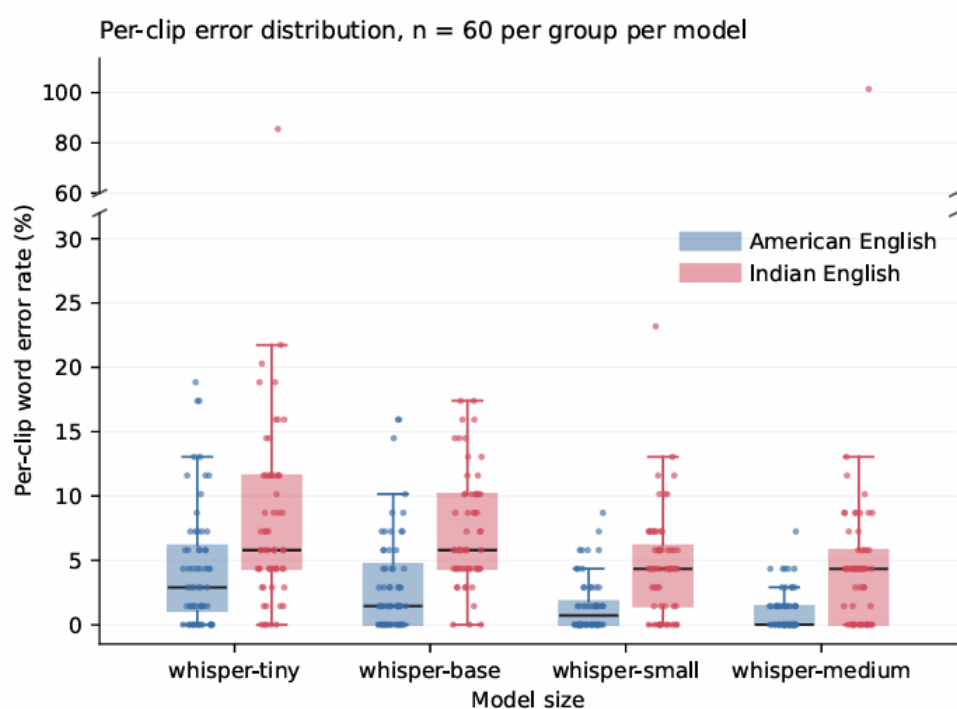


Figure 3. Per-clip WER distribution by group and model size. Boxes span the interquartile range; points are individual clips. The axis is broken so that the two catastrophic Indian clips remain visible.

Type	Reference	Transcribed as	Indian	American
Orthographic	five	5	32	0
Orthographic	six	6	32	0
Orthographic	three	3	30	0
Phonetic	slabs	slaps	9	0
Phonetic	snow	no	6	6
Phonetic	we	we'll	5	3
Phonetic	snake	snack	3	1
Phonetic	her	a	3	0

Table 2. Most frequent substitution pairs by group on whisper-medium, counted as clips affected out of 60 per group. Pairs occurring at least three times on Indian audio are shown. Orthographic errors are cases where the model rendered a correctly recognized number word as a numeral. acoustic disparity, and the direction of that overstatement is not neutral.

Table 3. Mean per-clip word error rate by accent group and Whisper model size, in percent, with standard error in parentheses. $n = 60$ recordings per group at each model size. Gap is the absolute difference in percentage points, Ratio is Indian WER divided by American WER, and p is from a one-sided Mann-Whitney U test.

Model	American	Indian	Gap (pp)	Ratio	p
tiny	4.52 (0.62)	8.86 (1.48)	4.35	1.96×	2.5×10^{-4}
base	3.26 (0.51)	7.34 (0.56)	4.08	2.25×	3.3×10^{-8}
small	1.52 (0.27)	4.76 (0.53)	3.24	3.13×	1.9×10^{-7}
medium	1.09 (0.21)	5.65 (1.68)	4.57	5.20×	9.5×10^{-8}

The pattern generalizes across architectures. Two additional systems were evaluated on the identical corpus (Figure 4). Meta’s wav2vec-based model produced 5.97% WER for American English and 9.13% for Indian English, a ratio of $1.53 \times$ ($p = 2.6 \times 10^{-5}$). NVIDIA’s Parakeet performed best overall, at 0.92% and 1.74% respectively, a ratio of $1.89 \times$ ($p = 0.046$). Neither model emitted numerals in any transcription, so their gaps are unadjusted and fully acoustic. Parakeet’s near-perfect American performance and its 31 error-free Indian transcriptions out of 60 show that low absolute error and a persistent relative gap are compatible: the best-performing system in the study still misrecognized Indian English at roughly twice the rate of American English.

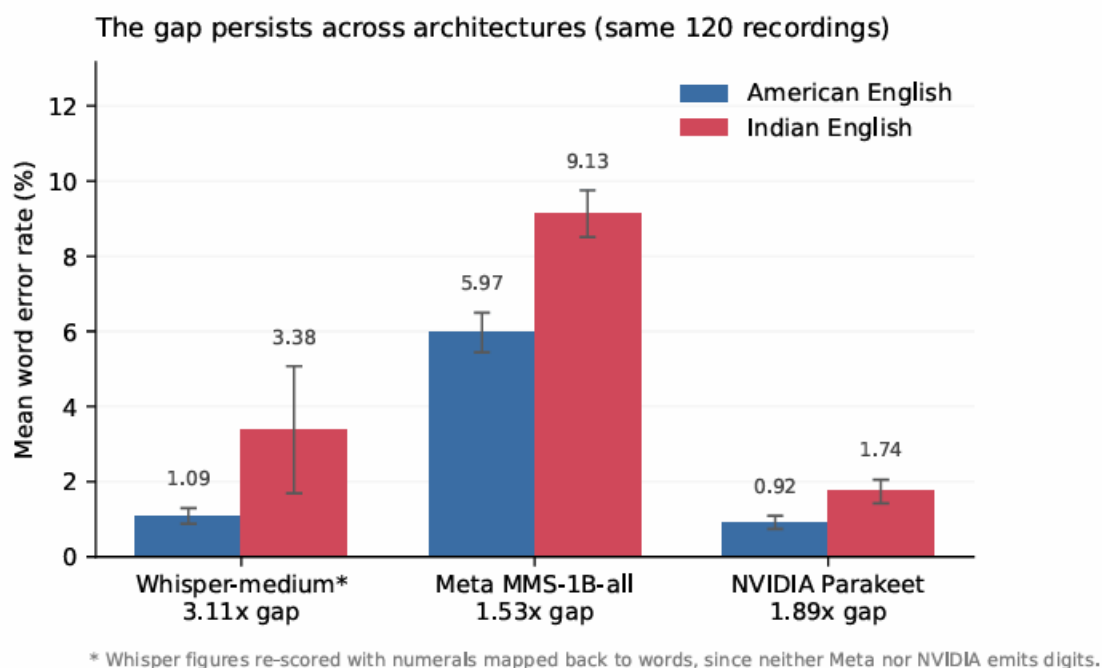


Figure 4. Cross-architecture comparison, Whisper versus Meta versus NVIDIA, on the same 120 recordings. Whisper-medium is shown after numeral normalization so that all three systems are scored on the same orthographic convention.

Discussion

The clearest result is that scale reduces error without redistributing it. Every increase in Whisper’s parameter counts lowered WER for both groups, yet the absolute gap stayed close to four percentage points and the ratio widened from 1.96× to 5.20×. The cause is visible in the error breakdown: American clips converge toward zero errors as models grow, while Indian clips retain a residual layer of accent-specific substitutions and, at medium, a heavier tail of insertions. Because Whisper’s decoder is autoregressive, a single acoustically unfamiliar clip can trigger runaway generation, which is how one recording reached 101.5% WER. This is consistent with Koenecke et al.’s (2020) finding that the disparity originates in acoustic modeling, and it extends that finding to a system whose authors explicitly claimed accent robustness as a benefit of scale (Radford et al. 2023).

Parakeet’s performance warrants separate treatment because its absolute numbers are so low: 0.92% for American speakers, 1.74% for Indian speakers, with 37 and 31 of 60 clips respectively transcribed with no errors at all. Three properties of the model can explain this. First, it was trained on roughly 670,000 hours of audio, about equal to Whisper’s corpus, but with punctuation and capitalization applied at training time and no numeral output, so its transcripts already conform to the orthographic conventions that WER rewards (NVIDIA 2025). Second, its transducer decoder emits at most a bounded number of tokens per audio frame, so it cannot hallucinate the way an autoregressive decoder can; its worst Indian clip scored 11.6% against Whisper-medium’s 101.5%. Third, the Speech Accent Archive is short, read, studio-quality speech, which is the condition under which any ASR system performs best, and the passage has been public since 2015, so it cannot be ruled out that recordings from it appear in the pseudo-labeled portion of Parakeet’s training data. What matters for this study is that none of these advantages closed the gap. Parakeet’s ratio of

1.89× is almost identical to whisper-tiny's, and its one-sided p -value of 0.046, while marginal, sits on a floor effect where half the clips score zero and the test has little room to discriminate. A system can be near-perfect on the dominant accent and still misrecognize Indian English at close to twice the rate.

Two implications follow. The number-normalization finding shows that a substantial share of the raw Whisper gap is a formatting choice rather than a recognition failure, and any audit that reports WER without inverse text normalization will overstate acoustic bias, in a direction that is not neutral across groups. And Parakeet shows that the ratio is the more durable measure of fairness than the absolute gap: as systems improve, absolute gaps shrink toward zero for everyone, but the relative burden on the under-represented group can persist unchanged. Fairness claims about ASR should therefore be evaluated on disaggregated, content-matched audio, and the choice between Whisper and a transducer-based system for Indian English users is, on this evidence, less a question of which is unbiased than of which fails more gracefully.

Conclusion

This paper asked whether scaling an ASR system resolves accent bias or merely lowers the error rate around it. For Whisper, the answer is the latter. Across 120 recordings of an identical passage, Indian-accented English was transcribed less accurately than American-accented English at every model size tested, and the difference was statistically significant in each case. Increasing the parameter count twentyfold reduced error for both groups, but it reduced American error faster, so the relative burden on Indian speakers rose from roughly double to roughly fivefold. Scale, on this evidence, is not a remedy for accent bias; it is a mechanism that can deepen it.

Three further findings sharpen that conclusion. A substantial share of Whisper's raw gap is orthographic, arising from the model's tendency to render spoken numbers as numerals for Indian speakers and as words for American speakers, and any audit that omits text normalization will misstate the size of the acoustic disparity. Once that share is removed, a smaller gap remains, concentrated on voiceless fricatives and consonant clusters and therefore acoustic in origin. And the gap is not an artifact of one architecture: Meta's MMS and NVIDIA's Parakeet reproduced it on the same audio, with Parakeet demonstrating that a system can approach zero error on the dominant accent while still misrecognizing Indian English at nearly twice the rate.

The study's scope is deliberately narrow. It examines read speech, a single passage, sixty speakers per group, and one under-represented accent, and it cannot say how the gap behaves in spontaneous conversation or across the regional varieties of Indian English. What it can say is that the disparity is real, measurable with open models on a consumer laptop, and invisible in the aggregate figures by which these systems are currently judged. If a word error rate is to mean anything for the second-largest English-speaking population in the world, it has to be reported for them.

References

- Faria, Arlo. 2006. "Accent Classification for Speech Recognition." In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, edited by Steve Renals and Samy Bengio, 3869:285–293. Lecture Notes in Computer Science. Berlin: Springer. https://doi.org/10.1007/11677482_25.
- Graham, Calbert, and Nathan Roll. 2024. "Evaluating OpenAI's Whisper ASR: Performance Analysis across Diverse Accents and Speaker Traits." *JASA Express Letters* 4 (2): 025206. <https://doi.org/10.1121/10.0024876>.
- Huang, Xuedong, and Li Deng. 2010. "An Overview of Modern Speech Recognition." In *Handbook of Natural Language Processing*, 2nd ed., edited by Nitin Indurkha and Fred J. Damerau, 339–366. Boca Raton, FL: Chapman / Hall/CRC.

Ike, Chinaemere, Seth Polsley, and Tracy Hammond. 2022. "Inequity in Popular Speech Recognition Systems for Accented English Speech." In *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces*, 66–68. New York: Association for Computing Machinery. <https://doi.org/10.1145/3490100.3516457>.

Javed, Tahir, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra. 2023. "Svarah: Evaluating English ASR Systems on Indian Accents." In *Proceedings of Interspeech 2023*, 5087–5091. <https://doi.org/10.21437/Interspeech.2023-2588>.

Koenecke, Allison, 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. <https://doi.org/10.1073/pnas.1915768117>.

NVIDIA. 2025. "Parakeet-TDT-0.6B-v3." Accessed September 19, 2026. <https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3>.

Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, et al. 2024. "Scaling Speech Technology to 1,000+ Languages." *Journal of Machine Learning Research* 25

(97): 1–52. <https://jmlr.org/papers/v25/23-1318.html>.

Radford, Alec, 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*, 202:28492–28518. *Proceedings of Machine Learning Research*. PMLR. <https://proceedings.mlr.press/v202/radford23a.html>.

Sailaja, Pingali. 2009. *Indian English*. Dialects of English. Edinburgh: Edinburgh University Press.

Vaessen, Nik. 2023. "JiWER: Similarity Measures for Automatic Speech Recognition Evaluation." Accessed September 19, 2026. <https://github.com/jitsi/jiwer>.

Weinberger, Steven H. 2015. "Speech Accent Archive." George Mason University. Accessed September 19, 2026. <http://accent.gmu.edu>.

Xiong, Wayne, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L. Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. "Toward Human Parity in Conversational Speech Recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (12): 2410–2423. <https://doi.org/10.1109/TASLP.2017.2756440>.

A Code and Interactive Results

Code repository. The audio clip lists, transcription scripts, per-model result CSVs and the figure-generation script are available on GitHub:

https://github.com/Arnavjagetiya/ASR_unicorn_labs

Interactive website. The results dashboard, the pipeline walkthrough and the in-browser WER calculator described in the Methodology are published at:

<https://asr-unicorn-labs.vercel.app>

WER calculator: <https://asr-unicorn-labs.vercel.app/wer-calculator>