ASR Benchmarking for AAVE

Evaluating and Improving Commercial ASR Performance on African American Vernacular English

MILDNESS AKOMOIZE

EECS, HOWARD UNIVERSITY, MILDNESS.AKOMOIZE@BISON.HOWARD.EDU

SAURAV K. ARYAL

EECS, Howard University, saurav.aryal@howard.edu

GLORIA WASHINGTON

EECS, HOWARD UNIVERSITY, GLORIA.WASHINGTON@HOWARD.EDU

Automatic speech recognition (ASR) systems are widely used in voice assistants, transcription services, and accessibility tools, yet prior research suggests they perform unevenly across dialects. This project investigates performance disparities in commercial ASR systems for African American Vernacular English (AAVE). We curated and transcribed over 200 hours of question–response style AAVE speech data and split it into training, validation, and test sets. Using an automated benchmarking pipeline, we evaluate systems including OpenAI Whisper, Amazon Transcribe, and Deepgram. Performance is measured using Word Error Rate (WER), with statistical analyses such as Welch's t-test and Shapiro–Wilk tests applied to assess significance and distributional assumptions. Preliminary findings indicate that several commercial systems exhibit elevated WER on AAVE speech relative to reported general benchmarks. To address this gap, we are fine-tuning models on the curated dataset and observing reductions in WER, though this phase remains ongoing. By combining systematic benchmarking, statistical rigor, and dataset development, this work contributes toward more equitable and representative speech recognition technologies

CCS CONCEPTS • Speech Recognition • Natural Language Processing • Artificial Intelligence

Additional Keywords and Phrases: Automatic Speech Recognition, African American Vernacular English, Word Error Rate, Algorithmic Bias, AI Fairness, Model Fine-Tuning, Speech Benchmarking

REFERENCES

[1] Allison Koenecke, Andrew Nam, Emily Lake, et al. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117

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

[3] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), Article 115, 1–35. https://doi.org/10.1145/3457607

[4] Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, 53–59. https://doi.org/10.18653/v1/W17-1606