Dialect Bias in Commercial ASR Systems: A Statistical Fairness Evaluation Using Error and Semantic Similarity Metrics

Kennedy Gregg, Saurav K. Aryal, PhD, Gloria Washington, PhD,
AI4PC Lab, Institute for Human Centered Artificial Intelligence

INTRODUCTION

Automatic Speech Recognition (ASR) systems are widely used in applications such as virtual assistants, transcription services, and accessibility tools. However, these systems often exhibit **performance disparities across dialects**, raising concerns about fairness and equity.

This study evaluates whether commercial ASR systems perform differently on:

- Standard English (Common Voice dataset)
- African American Vernacular English (AAVE)

Key Question:
Do ASR systems perform worse on AAVE due to dialect differences rather than noise or other factors?

METHODS

This study presents a systematic pipeline for evaluating dialect bias in commercial Automatic Speech Recognition (ASR) systems. The approach integrates multi-condition speech inputs, transcription outputs, error analysis, semantic similarity evaluation, and statistical testing to assess performance disparities.

The system consists of the following major steps:

1. Speech Input Collection:

~ 200 hours of audio samples were used from two sources: Standard English (Common Voice) [3] and African American Vernacular English (AAVE) [2}, We controlled for noise and duration by combining noise characteristics from AAVE with duration properties from Common Voice.

2. Error Metric Evaluation:
Each speech sample was processed through eight commercial ASR systems, generating predicted transcripts for each input. **Generated transcripts were compared against human ground truth transcripts using standard error metrics & semantic similarity metrics:**

- Word Error Rate (WER)
- Character Error Rate (CER)
- Levenshtein Distance
- BERT Cosine Similarity
- Euclidean Distance
- Jaccard Similarity

METHODS

3. Statistical Testing:
Performance differences across speech conditions were evaluated using two-sample t-tests (Student's or Welch's based on variance), with Levene's test for variance assessment and Benjamini–Hochberg correction for multiple comparisons.

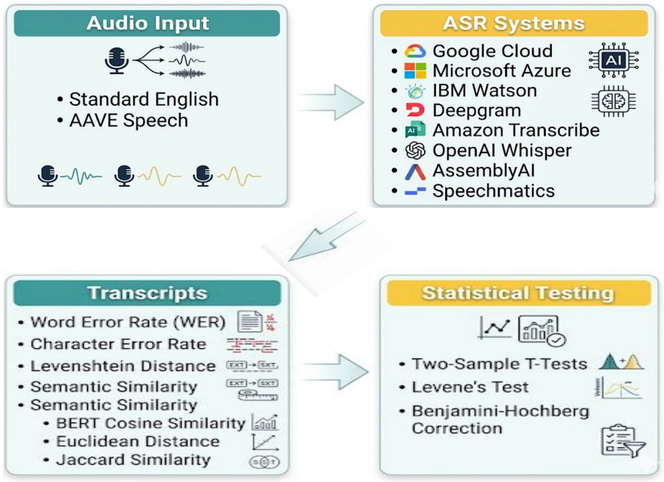
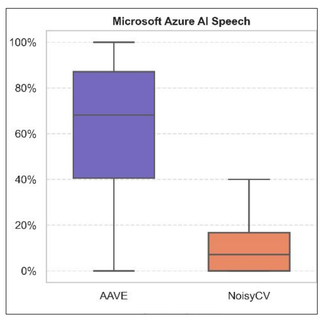


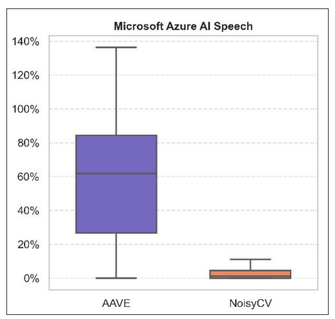
Figure 1: System Pipeline for Dialect Bias Evaluation in ASR Systems.

RESULTS

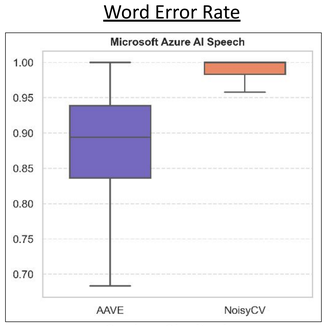
African American Vernacular English vs. Common Voice (Noisy Speech Conditions)



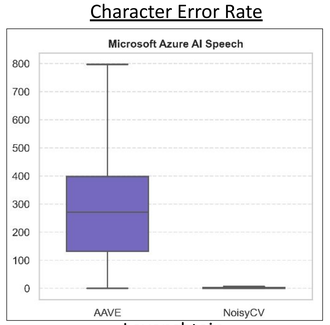
Word Error Rate



Character Error Rate



Cosine Similarity



Levenshtein

RESULTS

All statistical tests show significant disparities between African American Vernacular English and noisy Common Voice speech across all evaluated ASR systems and metrics (adjusted $p < .001$). Error-based and distance metrics (WER, CER, Levenshtein, Euclidean) are consistently higher for African American Vernacular English, while similarity metrics (cosine and Jaccard) are lower. These results demonstrate that performance differences are consistent, statistically significant, and reflect reduced transcription accuracy and semantic alignment for African American Vernacular English.

Company	metric	test_type	t	reject_H0
Microsoft Azure AI Speech	Word Error Rate	Welch	74	TRUE
Microsoft Azure AI Speech	Character Error Rate	Welch	65	TRUE
Microsoft Azure AI Speech	Cosine SImilarity	Welch	-51	TRUE
Microsoft Azure AI Speech	Levenshtein	Welch	59	TRUE

CONCLUSION

This study demonstrates that commercial ASR systems exhibit consistent and statistically significant performance disparities when processing African American Vernacular English compared to noisy Common Voice speech. **Across all evaluated systems and metrics, African American Vernacular English results in higher error rates and lower semantic similarity, indicating reduced transcription accuracy and meaning preservation.**

These findings suggest that **performance** gaps are driven more by dialectal variation than by noise, highlighting a critical limitation in current ASR systems. Ensuring equitable **performance** requires the development of dialect-aware evaluation methods and more diverse training data to improve robustness across linguistic variation.

REFERENCES

[1] Manu Edavakandam, "From Audio to Words: A Python Guide to Measuring Transcription Accuracy,"Medium,2023.
https://medium.com/@manuedavakandam/from-audio-to-words-a-python-guide-to-measuring-transcription-accurracy-f9dd9e70651f

[2] Johnson, J., Williams, L., Nias, J., Aryal, S. K., & Washington, G. (2025, April). Centering black voices: Lessons learned and reflections from a large-scale AAVE data collection at a historically black university. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-7).

[3] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), 4211–4215.

ACKNOWLEDGEMENTS

This work was supported in part by a Google Award (#956968). The views expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsor.