Effects of Accents: Algorithmic Bias in Speech Recognition AI

Rhett Lavender, Arjun Mahesh, Rebekah Northrup, Shana (Vy) Tran

DATA 120
Ethics of AI Fall 2024
Prof. Neil Gaikwad
TA Name: Patrick Tong

Question and Background

Our study explores the problem of possible sampling and algorithmic bias in voice recognition systems, specifically focusing on how these biases affect fairness for native Chinese speakers across many domestic dialect zones. We will investigate how these biases might impact their user experience, while simultaneously considering the broader social implications of voice-based AI assistants in the context of AI ethics and societal equity. More specifically, we will compare an automatic speech recognition system's accuracy for native Chinese speakers speaking English to its accuracy for native English speakers. We will be using Whisper by OpenAI as our ASR system for testing. Additionally, we will use demographic parity as a fairness metric to evaluate the system's overall performance in terms of inclusivity and possible intersectionality issues due to the nature of the samples from a Chinese-speaking-English dataset.

Course Integration

Philosophical:
- Fairness and non-discrimination/ Equitable regardless of linguistic background
- Respect for cultural diversity
- Fair representation, user privacy, transparency and accountability/ Intersectionality

Legal:
- GDPR/ U.S. Anti-Discrimination laws

Technological:
- Pre-trained model for automatic speech recognition
- Compared native English speakers and Chinese-accented English speakers
- Limitation: The model's overrepresentation of dominant accents in training data, leading to significantly better performance for commonly represented speech patterns while struggling to accurately decipher, and or transcribe, thicker or underrepresented accents such as those found in the Chinese-accented dataset.
- Code generated by ChatGPT

Synthesis:
- Fairness and inclusivity in voice-assisted AI
- Evaluation audits against the AI system and evaluation of multiple demographic performance metrics
- Ensuring model accuracy remains balanced and unbiased across all accent groups

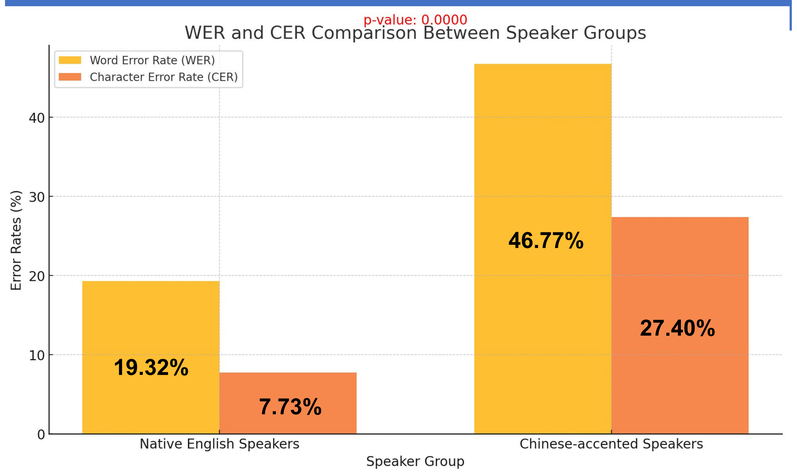
Literature:
- Debajyoti Pal wrote the paper "User Experience with Smart Voice Assistants: The Accent Perspective" and published it in 2019.
  - This finding implies that although to both the user groups, the VAs are usable the user satisfaction is much more with the native English speakers.

Analysis/Data/Method

Our system utilizes data from two different datasets designed for speech recognition software: one that contains English-speaking samples with thick Chinese accents and the other of native English-speaking samples.

- Mozilla Common Voice: Free database for speech recognition software that includes 3,384 speech samples of native English Speakers.
  - Out of the participants that listed their gender, 70% were female speakers
- 100,000 colloquial English sentences recorded by 3,691 Chinese-accented English speakers from Chinese-speaking dataset
  - 66% of the sample were female speakers.
- Key features: Accent, Age, Gender, Locale, and Variant
- Target variable: Actual sentence spoken
- Word Error Rate (WER) & Character Error Rate (CER): Used to measure the model's overall accuracy in transforming the audio samples into their corresponding target variable

Results



T-statistic: -144.03
P-value: 0.0000
Given the highly statistically significant p-value, we have sufficient evidence to reject the null hypothesis and suggest that there is a significant difference in WER/CER between native English speakers and Chinese-accented speakers. The p-value was calculated using a two-sample t-test.

Design Recommendations

Key components of the model: Existing ASR model (Whisper), two distinct linguistic datasets, and evaluation metrics including WER/CER to quantify transcription accuracy and demographic parity to assess fairness across groups.
Addressing fairness: Enabling an unbiased evaluation of the model's capacity to accurately transcribe audio from diverse demographic groups, particularly in assessing disparities in WER/CER between native English and Chinese speakers.
- Incorporated considerations for fairness and societal impacts by including diverse datasets to evaluate demographic parity, using WER/CER as an accuracy measure to identify transcription disparities between the two groups, and exploring possible issues of intersectionality (gender and demographic) to understand compounded effects on transcription accuracy and potential bias.

Hypothesis test: Statistically analyze the results and confirm any observed disparities.
- Focus on quantifying the differences in transcription accuracy (via WER/ CER) between the two datasets.
- Aim to assess whether demographic disparities exist and to quantify their significance.
- Null hypothesis: There is no significant difference in WER/CER between native English speakers and Chinese native speakers.

Conclusion/What did you Learn

The study found significant differences in transcription accuracy (measuring using Word Error Rate and Character Error Rate) between native English speakers and Chinese-accented English speakers.This highlights possible presence of algorithmic bias in the voice recognition system Whisper by OpenAI, most likely caused by underrepresentation of non-native English data in training sets, further indicating need for fairness and inclusivity of all groups.
Furthermore, our study found that gender was also a factor in the accuracy of the model. For the native English speaker dataset, average WER for female speakers was 20.65% while the male speakers had an average WER of 13.38%. In addition, female speaker's average CER was 13.38% and male speakers' average CER was only 3.82%. Similarly, for the Chinese ESL speakers, female speakers had an higher average WER of 56.52% and CER of 40.22% compared to male speakers who had an average WER of 37.14% and CER of 25.98%. This could possibly be due to harms caused by overrepresentation of an underrepresented group (women). Furthermore, these disparities highlight a more nuanced need to address harms of intersectionality.