Effects of Accents: Bias in Voice-Based AI

Rhett Lavender, Data Science B.S., Junior
Arjun Mahesh, Computer Science B.S., Sophomore
Rebekah Northrup, Statistics and Analytics B.S., Data Science B.A., Sophomore
Shana Tran, Computer Science B.A., Data Science B.A., Junior

PROJECT TYPE

Computational Modeling

KEYWORDS

- Algorithmic Bias
- Quality of Service
- Fairness Metrics
- Human-AI Interaction

Contribution Statement:

· **Rhett Lavender:** Contributed to research regarding automatic speech recognition systems and their uses, data preprocessing and cleaning, performance evaluation and analysis of the selected model, drafting the Evaluation and Results, Data Source and Preprocessing, and Course Integration sections of the manuscript, final revisions throughout the manuscript, and the writing and running of Python code. Additionally, I acknowledge that I have read the entire report, am familiar with its content, and am able to answer any questions related to the report's findings and conclusions.

· **Arjun Mahesh:** Contributed to doing research, I focused on examining the ethical dimensions of this project alongside analyzing the associated data. I dedicated time to researching and writing about the ethical and legal considerations, as well as exploring the technological limitations that might affect the project's scope. Additionally, I reviewed the data, summarized key findings, and provided interpretations to ensure a comprehensive understanding of the project's implications. Additionally, I acknowledge that I have read the entire report, am familiar with its content, and am able to answer any questions related to the report's findings and conclusions.

· **Rebekah Northrup:** Contributed to designing the overall structure of the project, researching literature sources, implementing mock code for model testing, designing methods, running hypothesis tests, creating visualizations, writing the poster, and revising the conclusion. Additionally, I acknowledge that I have read the entire report, am familiar with its content, and am able to answer any questions related to the report's findings and conclusions.

· **Shana Tran:** Contributed to designing the structure of the project, collecting and analyzing the Chinese-speaking-English dataset, interpreting the results of this dataset, drafting the introduction + data source and preprocessing + conclusion of the manuscript, and revising the final content of the report. Performed intellectual contributions on the conceptual direction of the project, critical analysis on intersectionality, theoretical input on demographic parity as a fairness metric, and possibilities of over-representational harm to female speakers. Performed practical tasks of coding, statistical analysis, manuscript formatting of the content, and initiating group assignments + updates of projects. Additionally, I acknowledge that I have read the entire report, am familiar with its content, and am able to answer any questions related to the report's findings and conclusions.

DECLARATION OF AI ASSISTANCE

This project utilized OpenAI's ChatGPT 4o model to assist in formulating general code structure and the debugging process. Specifically, the model provided insights on efficient algorithm design, recommended approaches for structuring code, and solutions to resolve errors during debugging. All outputs from the model were critically reviewed and adapted as necessary to ensure their accuracy and relevance to the project. Our input prompts and sample output are listed in the appendix.

ABSTRACT

This project presents the evaluation of Whisper by OpenAI, a speech-to-text transcription and speech recognition model to investigate the impact of various Chinese accents on the overall fairness of automatic speech recognition (ASR) systems. We utilized a dataset from Mozilla's Common Voice project and NexData to evaluate performance disparities between native English speakers and Chinese-accented English speakers by comparing each group's accuracy scores. In our context, accuracy scores are calculated from the Word Error Rate (WER) and Character Error Rate (CER). Additionally, we explored the influence of gender on these performance metrics, with the potential to identify algorithmic bias related to these features.

Our findings predict significant accuracy differences between the two linguistic groups, highlighting potential biases in the ASR model. We explore the intersectionality of demographic factors, race, and gender, to understand how these biases may compound. The study integrates ethical principles, including fairness and inclusivity, while adhering to legal standards like GDPR and anti-discrimination laws. Results will inform strategies to mitigate algorithmic bias, improve model performance across linguistic diversity, and promote equitable user experiences.

By addressing disparities in ASR systems, this work contributes to broader AI fairness and AI-human interaction goals by emphasizing the need for inclusive data sampling, transparency, and ethical design. Our findings aim to foster advancements in voice recognition technology, ensuring accessibility and equity for all users.

1. INTRODUCTION

Our study explores the problem of possible sampling and algorithmic bias in automated speech recognition (ASR) models, 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 speech recognition systems in the context of AI ethics and societal equity. More specifically, we will determine the ASR model's accuracy for Chinese-accented English speakers and native English speakers by comparing the word error rate and character error rate between both groups. Additionally, we will use demographic parity as a fairness metric to evaluate the ASR's overall performance in terms of inclusivity.

Some key contextual points for understanding this problem include differences in accents and dialects that may affect the quality of service, the population that is being represented in the dataset to train these voice-based assistants (as limited data on underrepresented dialects can be unfair and reinforce biases), and that voice-based AI systems typically perform best on standard English which consequently results in technological exclusion for users with diverse linguistic backgrounds.

To address the broader problem of sampling and algorithmic bias in ASR models, we will focus specifically on Whisper by OpenAI, a speech-to-text transcription and speech recognition model widely implemented by platforms like Microsoft. We plan to run two datasets through Whisper: one featuring standard English speakers and one featuring Chinese-accented English speakers. Our goal in comparing the accuracy results for this ASR model is to investigate and evaluate its performance across diverse linguistics backgrounds in speech recognition. By identifying potential disparities, we aim to raise awareness of the need to improve fairness in voice recognition systems. Our efforts seek to mitigate these biases, enhance recognition accuracy, and ultimately promote greater inclusivity and user satisfaction in recognition technologies.

This research is important because it highlights the need to account for scenarios in which inclusivity and fairness principles are violated, especially when it comes to human-AI interaction which has significant implications for accessibility and trustworthy technology for all users. It could potentially impact the AI industry by prompting the adoption of more concrete fairness metrics and careful consideration of ethics in model training, such as the incorporation of increasingly diverse datasets and testing protocols that ensure fair service across different dialects.

Our approach relates to AI fairness and ethics in the following ways: It addresses the concept of algorithmic bias, which occurs when AI systems favor certain demographics over others, inadvertently excluding users with thicker accents and dialects. In addition, it considers fairness in data sampling and model training, ensuring that voice-based AI is user-friendly and effective for a broader population. Data sampling and model training impact outcomes by improving inclusivity and the trustworthiness of technologies built to benefit human needs.

Key fairness/societal considerations in our design include: Reducing sampling bias by using a dataset with broad dialect representation. Mitigating algorithmic bias through model adjustments that improve recognition for underrepresented accents. Ensuring transparency by documenting the AI system's design, training, and performance across linguistic groups to maintain user trust and accountability.

2. COURSE INTEGRATION

2.1 Philosophical Aspect

From a deontological standpoint, we believe we must ensure fairness and non-discrimination in our system. We will strive to create a framework that serves every user equitably, regardless of linguistic background. Additionally, we draw on virtue ethics, focusing on values like inclusivity, fairness, and respect for cultural diversity. Some key ethical considerations that we must take into account when designing our system include fair representation, user privacy, and transparency and accountability. To ensure fair representation, we will include accent-diverse datasets that reflect a wide range of linguistic backgrounds. This approach helps prevent the overrepresentation of mainstream accents and reduces the risk of biased model behavior. To uphold fairness and non-discrimination principles, we will implement fairness metrics across different accent groups and regularly assess these metrics to identify and address potential biases.

2.2 Legal Aspect

We will adhere to provisions of the GDPR, which outline standards for data collection and protection, as well as various U.S. anti-discrimination laws that ensure services are accessible without discrimination based on race, national origin, or linguistic traits. To follow these regulations, we will ensure that all speech data is anonymized and processed in a secure environment. Our design will also include clear documentation and explanations about how the system works and the data it uses. Potential legal challenges may arise in the form of bias and discrimination claims. If our system were to demonstrate significant biases against certain linguistic groups, it could be investigated under anti-discrimination laws.

2.3 Technological Aspect

For the dataset, our dataset includes a diverse range of voice samples with various accents and dialects. This selection was made to address sampling bias and representation bias by ensuring that the model is trained on a broad spectrum of linguistic variations. We utilized an automated speech recognition system, applying it to datasets with native English speakers and Chinese-accented English speakers. A notable limitation observed was the model's overrepresentation of male-speaking standard English, leading to significantly better performance for commonly represented speech patterns and gender while struggling to accurately decipher, and or transcribe, thicker or underrepresented accents such as those found in the Chinese-accented dataset. This highlights the critical issue of algorithmic bias in voice recognition systems which impacts our goals of inclusivity and fairness of these technologies. To help address this, we might look at other models as well to help generate other examples that better represent underrepresented accents. This will potentially help the model generalize across diverse accents, enhancing its fairness and inclusivity.

2.4 Synthesis

In our project, the philosophical, legal, and technological aspects intersect to address fairness and inclusivity in voice-based AI. Philosophically, the concept of fairness is crucial, as we aim to ensure that users with diverse or non-mainstream accents receive equal treatment and recognition accuracy in the AI, which aligns with our fairness goals. Legally, our system aims to comply with regulations like the GDPR and anti-discrimination policies by protecting user data privacy and ensuring that no group is disadvantaged. Technologically, the implementation of the ASR system combined with our diverse datasets featuring native English speakers and Chinese-accented English speakers and fairness evaluation techniques such as demographic parity analysis, plays a central role in addressing fairness. Our design achieves a balance of all these aspects by properly upholding the fairness principles within the technical framework and diverse dataset. Legal concerns are also considered through our privacy measures and transparency on how the data will be used within the AI. Potential tradeoffs include increased computational efficiency from the large dataset and compromises in recognition performance for highly underrepresented groups, as achieving perfect accuracy across all accents/dialects and between genders is challenging.

Our system aims to adopt a holistic fairness approach by combining data diversity, and other computational techniques to address accent-based biases. Fairness is further enhanced through continuous evaluation audits of the ASR system and evaluation of demographic performance metrics, ensuring that model accuracy remains balanced and unbiased across diverse groups.

3. LITERATURE REVIEW

Debajyoti Pal wrote the paper “User Experience with Smart Voice Assistants: The Accent Perspective” and published it in 2019. The study investigated how a specific population of voice-assisted AI users have different experiences based on their English accent availability. This study aims to identify if there exist any differences between these two groups of users concerning the overall usability and the satisfaction received after using the voice assistants. They focused on this because, from an end-user perspective, very little is known about the usability, acceptability, satisfaction, and usage pattern of the voice assistants between two different groups of users. The main results of their study were that there were no significant differences in usability between the two user groups. They also found that the same was not true for satisfaction levels. They could not find any significant difference between the two groups of users. This indicates the positive outcomes of many previous researches on the voice recognition of voice-assisted systems. Furthermore, regarding satisfaction levels in using these systems, a noticeable difference was seen between the two user groups. Overall, the native speakers had a higher satisfaction rate than their counterparts. This finding implies that although voice-assisted systems are accessible to both user groups, there is a higher user satisfaction with native English speakers. These findings inform our conceptual extension by providing insights and fueling the discussion of how voice-assisted AI systems should be designed and evaluated to ensure they are user-friendly, trustworthy, ethical, and beneficial for all humans (and a big part of that is the accuracy in its speech recognition model).

4. COMPUTATIONAL MODELING PROJECT

4.1 Data Source and Preprocessing

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 standard English-speaking samples.

Common Voice is a crowdsourcing project started by Mozilla to create a free database for speech recognition software. We are using one of the database's smaller datasets to conduct our analysis for standard English speakers. After cleaning and preprocessing some data, we have 3,384 speech samples with which we can work. The data preprocessing step consisted of removing samples with missing values and filtering the data to show only audio recordings done in English. Some key features include accent, age, gender, locale, and variant. For the accent feature, we chose to only select samples with an associated "United States English" accent, assuming that this would give us clear, easy-to-understand English audio samples to use on the ASR. Our target variable in this case will be the actual sentence spoken. We can use Word Error Rate, a common speech recognition metric, to measure the ASR's overall accuracy in transforming the audio samples into their corresponding target variables. This dataset will intentionally underrepresent non-native English speakers, as we seek to assess model performance on standard English speakers and compare metrics to Chinese English speakers. It is also important to note that Common Voice obtains its speech recordings through donations. In other words, individuals must volunteer to read sentences and have their voices recorded and put into a dataset. There do not seem to be any sort of quality issues with the recordings, but there are a fair number of missing values in the dataset. Feature engineering was not necessary when working with this dataset.

After conducting preliminary research, we found that thick Chinese accents when speaking English seemed to have the highest error rate with English speech recognition software. We, therefore, chose a Chinese-speaking-English data source because it was collected to ideally improve the recognition effect of speech recognition systems on Chinese-speaking English. This is most relevant to our interest as it contains a range of audio samples of individuals from different regions in China with strong dialects speaking English on a variety of topics. The dataset contains 100,000 colloquial English sentences recorded by 3,691 Chinese-speaking English individuals. Key features include the environment/location the recordings took place in, race, and gender in the aims of the target variable: effective speech recognition. The environment/location is a key feature to reduce any noisy data that the ASR might collect. The demographic of this dataset is, evidently, Chinese individuals. Gender is particularly important with 66% of the samples being females. This representation in the dataset is significant to note when addressing potential gender biases as voice recognition software often has a strong representation of male voices. However, this could counterintuitively contribute to sampling bias, potentially leading to an overrepresentation of women which may ultimately harm their accuracy in speech recognition detection. There doesn't seem to be any quality issues with the dataset considering that all samples are in the same format, the data was recently collected (7 months ago), and there are no significant issues in missing/incomplete/duplicate data. We don't think data preprocessing is necessary with this particular dataset since all the samples are diverse and would further enrich our study. Feature engineering was also not necessary when working with this dataset.

4.2 Model Design and Methodology

Key components of our system include an ASR model (Whisper by OpenAI), two distinct datasets – one comprising audio files from Native English speakers and the other from Chinese-accented English speakers – Word Error Rate (WER) and Character Error Rate (CER) as evaluation metrics to quantify transcription accuracy, and demographic parity to assess fairness across groups. We have chosen a speech-to-text transcription and speech recognition model because it provides a standardized and robust method for comparing transcription performance across diverse accents.

This addresses fairness by 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 and non-native speakers. We’ve incorporated considerations for fairness and societal impacts by including diverse datasets representing accents and gender to evaluate demographic parity, using WER/CER as an accuracy measure to identify transcription disparities between both groups. In addition, we will explore the intersectionality of gender and accents to understand compounded effects on transcription accuracy and potential biases.

Our evaluation process involves using the ASR model for transcription, as the focus is on analyzing its performance rather than modifying the model. We validated our model by examining transcription accuracy (via WER/CER) across the two datasets and testing for significant differences in performance metrics. Additionally, we performed a hypothesis test to statistically analyze the results and confirm any observed disparities. The statistical analysis will focus on quantifying the differences in transcription accuracy (via WER/CER) between the two datasets. In addition, a statistical hypothesis test was conducted to determine whether the difference in WER between the two groups was significant. The hypotheses are as follows: the null hypothesis suggests that there is no significant difference in WER between Native English speakers and Chinese speakers, and the alternative hypothesis is that there is a significant difference in WER between the two groups. The statistical findings will be interpreted to assess whether demographic disparities exist and the extent of these disparities in transcription accuracy by quantifying their significance. As a result, the outcomes of this analysis will inform our understanding of potential algorithmic biases present in the model.

4.3 Evaluation and Results

We will assess the model’s performance by using two different metrics commonly used in the evaluation of speech recognition and natural language processing systems: word error rate (WER) and character error rate (CER). These measures take into account three primary types of errors – substitutions, deletions, and insertions. Formally, WER is calculated using the following formula:

$$WER = \frac{Substitutions + Deletions + Insertions}{\#\ of\ Words\ in\ Reference}$$

CER is calculated using a very similar formula, as it examines the same types of errors as WER:

$$CER = \frac{Substitutions + Deletions + Insertions}{\#\ of\ Characters\ in\ Reference}$$

For both metrics, a value of zero represents a perfect transcription of the speech sample. In other words, whatever words are spoken into the ASR system are output in text form without any error whatsoever. It is important to note that neither measure is bounded above by any value. If we had a reference of “hello” and candidate “bye-bye”, our WER would be 2.0 because there are two errors in the candidate and just one word in the reference. So, in general, we want the model’s WER and CER to be as close to zero as possible.

The dataset containing the speech samples of native English speakers was given to the model to transcribe (refer to Algorithm 1), and the mean word error rate among all samples was calculated to be approximately 19.32 percent. The mean character error rate was approximately 7.73 percent. When the model was given the speech samples of Chinese-accented speakers to transcribe (refer to Algorithm 2), the mean word error rate was approximately 46.77 percent, and the mean character error rate was approximately 27.40 percent. Keeping in mind that the model was tested on a relatively small number of samples of both native English speakers and Chinese-accented speakers, we can safely conclude that there is a significant difference in the Whisper ASR system's transcription accuracy for native English speakers and Chinese-accented English speakers. It appears that the voice recognition system is disproportionately negatively impacting those with Chinese accents speaking the English language. These calculated error rates highlight the possible presence of sampling and algorithmic bias in the Whisper ASR system, most likely caused by the underrepresentation of non-native English speakers' data in training sets, further indicating the need for rigid fairness measures and general inclusivity of all groups in data collection.

Our study also included conducting a two-sample t-test to compare the Word Error Rate (WER) and Character Error Rate (CER). The goal of the hypothesis test was to evaluate whether there was a statistically significant difference in the transcription accuracy of the Whisper ASR system for these two groups. The null hypothesis was that there is no difference in WER and CER between native English speakers and Chinese-accented English speakers. The alternative hypothesis was that there was a significant difference in WER and CER between the two groups. The p-value and t-statistic were calculated to assess the hypothesis. The results are as follows: t-statistic was -144.03 and the p-value was 0.0000. This large negative value of the t-statistic indicates that the mean WER and CER for Chinese-accented speakers are significantly higher than those for native English speakers. A p-value of 0.0000 indicates strong evidence to reject the null hypothesis in favor of the alternative hypothesis.

Our study additionally found that gender was a factor in the accuracy of the model. For the samples of native English speakers, the mean word error rate for female speakers was 20.65 percent, while male speakers had a mean word error rate of 13.38 percent. Female samples in this dataset also had an average character error rate of 13.38 percent, while male speakers' average character error rate was just 3.82 percent. Similarly, for the Chinese-accented speakers, females had a higher average WER of 56.52 percent and a higher average CER of 40.22 percent compared to male speakers who had an average WER of 37.14 percent and CER of 25.98 percent. We believe this discrepancy in error rates based on gender might be due to the overrepresentation of an underrepresented group (females). These disparities emphasize the need for a more nuanced approach to automatic speech recognition systems, possibly one that takes into account the question of intersectionality.

Table 1. Comparison of WER and CER among native English speakers and Chinese-accented speakers, by gender

Metric	M-NE	F-NE	M-CA	F-CA
WER	0.1338	0.2065	0.3714	0.5652
CER	0.0382	0.1338	0.2598	0.4022

Where M-NE represents male native English speakers, F-NE represents female native English speakers, M-CA represents male Chinese-accented English speakers, and F-CA represents female Chinese-accented English speakers.

5. DISCUSSION AND RECOMMENDATIONS

Challenges faced: collecting and curating sufficiently diverse datasets to minimize sampling bias and ensure inclusivity and addressing the imbalance in error rates across gender and linguistic groups, which highlights limitations in current speech recognition models. Our work has highlighted the need for further investigations into the intersectionality of linguistic accents and gender in speech recognition accuracy and the development of more inclusive training datasets that adequately represent underrepresented groups. In consideration of ethics, we addressed ensuring fairness by incorporating diverse datasets and evaluating demographic parity to address algorithmic bias and upholding user privacy through data anonymization and secure processing environments in our design. If implemented, our model could improve speech recognition accuracy for underrepresented linguistic and gender groups. It could also promote greater inclusivity and trustworthiness in AI systems, leading to more equitable user experiences. In contrast, possible negative consequences may include overgeneralization and overrepresentation of certain groups. Overgeneralization may happen where improvements for one underrepresented group may not generalize to others. A risk of skewing results from the overrepresentation of certain groups in the dataset exists. We plan to mitigate these by continuously updating datasets to reflect broader diversity and regularly auditing model performance across multiple demographic dimensions. Our model has implications for bridging technological accessibility gaps for users with diverse linguistic backgrounds and encouraging ethical AI practices in the development of voice-based technologies to ensure fairness and inclusivity. Based on our findings, we recommend the following policy changes: mandating the use of fairness metrics like demographic parity in evaluating AI models and requiring transparency and public documentation of datasets used for training speech recognition systems. Furthermore, for future AI systems in this domain, we suggest incorporating multi-accent and multi-gender balanced datasets from the early stages of development for training and establishing standardized fairness evaluation protocols to assess the inclusivity of speech recognition systems.

6. CONCLUSION

Our system for evaluating potential algorithmic bias in Whisper by Open AI, a speech-to-text transcription and speech recognition model, demonstrates that inclusivity and fairness metrics must be accounted for in human-AI interaction technologies, especially for those that involve speech recognition (such as voice-assisted AI). The most significant aspects are our highly statistically significant findings of WER and CER compared between both groups and between genders of each group affecting the overall accuracy of the ASR model. This project advances AI fairness by illustrating an important algorithmic flaw of an ASR model and underscores the importance of the inclusion of diverse datasets and acknowledging intersectionality to ensure that a wide range of societal groups are well-represented in developing fair and ethical AI models. Moving forward, it’s crucial to consider the implementation of ethical frameworks and fair practices in every step when designing fair AI models since any misstep can have detrimental effects and future implications on AI-human interaction, as seen by our study.

7. ACKNOWLEDGEMENT

We give special thanks to Neil Gaikwad and Guansen Tong for guidance related to this project.

8. APPENDIX

ALGORITHM 1: Evaluating Whisper with Native English speaking dataset

```python
import pandas as pd
import whisper
from jiwer import wer, cer
import json
# Load Whisper model
model = whisper.load_model("base")
# Initialize results list
results = []
# Process each file in the DataFrame
for index, row in us_english.iterrows():
    file_path = row['full_path']
    ground_truth = row['sentence']
    if not os.path.exists(file_path):
        print(f"File not found: {file_path}, skipping.")
        continue
    print(f"Processing {file_path}...")
    # Transcribe the audio file
    transcription = model.transcribe(file_path)["text"]
    # Calculate WER and CER
    current_wer = wer(ground_truth, transcription)
    current_cer = cer(ground_truth, transcription)
    # Append results
    results.append({
        "file": file_path,
        "transcription": transcription,
        "ground_truth": ground_truth,
        "wer": current_wer,
        "cer": current_cer,
    })
# Save results to a JSON file
with open("subset_results.json", "w") as f:
    json.dump(results, f, indent=4)
print("Processing complete. Results saved to 'subset_results.json'.")
```

ALGORITHM 2: Evaluating Whisper with Chinese-accented English dataset

```
!git clone
https://github.com/Nexdata-AI/593-Hours-Chinese-Speaking-English-Speech-Da
ta-by-Mobile-phone.git ## get dataset
!pip install git+https://github.com/openai/whisper.git
!pip install ffmpeg
!apt-get install ffmpeg ## install whisper AI
import whisper
import os
model = whisper.load_model("base")
# Path to the folder containing .wav files
wav_folder =
'/content/593-Hours-Chinese-Speaking-English-Speech-Data-by-Mobile-phone'
# List all .wav files in the folder
wav_files = [f for f in os.listdir(wav_folder) if f.endswith('.wav')]
# Loop through and transcribe each audio file
```

```
for wav_file in wav_files:
   file_path = os.path.join(wav_folder, wav_file)
   print(f"Transcribing {wav_file}...")
   # Perform transcription
   result = model.transcribe(file_path)
   # Print transcription
   print(f"Transcription for {wav_file}:")
   print(result['text'])
   print("=" * 80)  # Separator for readability
!pip install jiwer
import os
from jiwer import wer, cer
# Path to dataset
wav_folder = '/content/593-Hours-Chinese-Speaking-English-Speech-Data-by-Mobile-phone'
# Updated Transcriptions generated by Whisper
transcriptions = {
   "T0055G0032S0125.wav": "Hey, hold out a cave in the slowfield for light.",
   "T0055G0009S0148.wav": "끝에 beat punished 그래서 너무 두abilities",
   "T0055G0003S0009.wav": "Her tricks had fallen in, making her look old.",
   "T0055G0007S0001.wav:": "No one could know why he did like that.",
   "T0055G0064S0018.wav": "Attack from above.",
   "T0055G0030S0021.wav": "No me, okay, I'm in sleep.",
   "T0055G0023S0004.wav": "The box is cute of the tree.",
   "T0055G0002S0001.wav:": "오늘도",
   "T0055G0066S0004.wav": "He asked them to discontinue flights over the island.",
    "T0055G0033S0001.wav": "I can lend this book to you. I had it date two weeks ago."
}
# Load ground truth text files
ground_truth = {}
for txt_file in os.listdir(wav_folder):
    if txt_file.endswith('.txt'):
        with open(os.path.join(wav_folder, txt_file), 'r', encoding='utf-8') as f:
            ground_truth[txt_file.replace('.txt', '.wav')] = f.read().strip()
# Debugging: Check loaded files
print("Ground truth files loaded:", ground_truth)
print("Transcriptions keys:", transcriptions.keys())
print("Ground truth keys:", ground_truth.keys())
output_file = "wer_cer_results.txt"
# Open a file to save the results
with open(output_file, 'w', encoding='utf-8') as f:
    for filename, generated_text in transcriptions.items():
        if filename in ground_truth:
            reference_text = ground_truth[filename]
            error_rate_wer = wer(reference_text, generated_text)
            error_rate_cer = cer(reference_text, generated_text)
            # Write results to the file
            f.write(f"WER for {filename}: {error_rate_wer:.2%}\n")
            f.write(f"CER for {filename}: {error_rate_cer:.2%}\n")
            f.write(f"Reference: {reference_text}\n")
            f.write(f"Transcription: {generated_text}\n")
            f.write("=" * 80 + "\n")
            # Print results for verification
            print(f"WER for {filename}: {error_rate_wer:.2%}")
            print(f"CER for {filename}: {error_rate_cer:.2%}")
            print(f"Reference: {reference_text}")
            print(f"Transcription: {generated_text}")
            print("=" * 80)
        else:
            f.write(f"Ground truth not found for {filename}\n")
            f.write("=" * 80 + "\n")
            print(f"Ground truth not found for {filename}")
print(f"All WER and CER results saved to {output_file}")
import os
from jiwer import wer, cer
# Path to your dataset
wav_folder = '/content/593-Hours-Chinese-Speaking-English-Speech-Data-by-Mobile-phone'
# Updated Transcriptions generated by Whisper
transcriptions = {
   "T0055G0032S0125.wav": "Hey, hold out a cave in the slowfield for light.",
   "T0055G0009S0148.wav": "끝에 beat punished 그래서 너무 두abilities",
   "T0055G0003S0009.wav": "Her tricks had fallen in, making her look old.",
   "T0055G0007S0001.wav:": "No one could know why he did like that.",
   "T0055G0064S0018.wav": "Attack from above.",
   "T0055G0030S0021.wav": "No me, okay, I'm in sleep.",
   "T0055G0023S0004.wav": "The box is cute of the tree.",
   "T0055G0002S0001.wav:": "오늘도",
    "T0055G0066S0004.wav": "He asked them to discontinue flights over the island.",
    "T0055G0033S0001.wav": "I can lend this book to you. I had it date two weeks ago."
}
# Load ground truth text files
ground_truth = {}
for txt_file in os.listdir(wav_folder):
    if txt_file.endswith('.txt'):
        with open(os.path.join(wav_folder, txt_file), 'r', encoding='utf-8') as f:
            ground_truth[txt_file.replace('.txt', '.wav')] = f.read().strip()
# Variables to calculate total WER and CER
total_words = 0
total_chars = 0
total_word_errors = 0
total_char_errors = 0
output_file = "wer_cer_results_with_totals.txt"
# Open a file to save the results
with open(output_file, 'w', encoding='utf-8') as f:
    for filename, generated_text in transcriptions.items():
        if filename in ground_truth:
            reference_text = ground_truth[filename]
            # Calculate WER and CER
            word_error_rate = wer(reference_text, generated_text)
            char_error_rate = cer(reference_text, generated_text)
            # Calculate individual word and char counts
            word_count = len(reference_text.split())
            char_count = len(reference_text.replace(" ", ""))  # Exclude spaces
            # Aggregate totals
            total_words += word_count
            total_chars += char_count
            total_word_errors += word_error_rate * word_count
            total_char_errors += char_error_rate * char_count
            # Write individual results to the file
            f.write(f"WER for {filename}: {word_error_rate:.2%}\n")
            f.write(f"CER for {filename}: {char_error_rate:.2%}\n")
            f.write(f"Reference: {reference_text}\n")
            f.write(f"Transcription: {generated_text}\n")
            f.write("=" * 80 + "\n")
            # Print individual results for debugging
            print(f"WER for {filename}: {word_error_rate:.2%}")
            print(f"CER for {filename}: {char_error_rate:.2%}")
        else:
            f.write(f"Ground truth not found for {filename}\n")
            print(f"Ground truth not found for {filename}")
    # Calculate overall WER and CER
    total_wer = total_word_errors / total_words if total_words > 0 else 0
    total_cer = total_char_errors / total_chars if total_chars > 0 else 0
    # Write total results to the file
    f.write("\nOverall Results:\n")
    f.write(f"Total WER: {total_wer:.2%}\n")
    f.write(f"Total CER: {total_cer:.2%}\n")
    print("\nOverall Results:")
    print(f"Total WER: {total_wer:.2%}")
```

```
   print(f"Total CER: {total_cer:.2%}")

print(f"All WER and CER results, including totals, saved to
{output_file}")

gender_metadata = {

   "T0055G0032S0125.wav": "female",

   "T0055G0009S0148.wav": "male",

   "T0055G0003S0009.wav": "female",

   "T0055G0007S0001.wav": "male",

   "T0055G0064S0018.wav": "male",

   "T0055G0030S0021.wav": "female",

   "T0055G0023S0004.wav": "male",

   "T0055G0002S0001.wav": "female",

   "T0055G0066S0004.wav": "male",

   "T0055G0033S0001.wav": "female"

}

import os

from jiwer import wer, cer

# Path to your dataset

wav_folder =
'/content/593-Hours-Chinese-Speaking-English-Speech-Data-by-Mobile-phone'

# Updated Transcriptions generated by Whisper

transcriptions = {

   "T0055G0032S0125.wav": "Hey, hold out a cave in the slowfield for
light.",

   "T0055G0009S0148.wav": "끝에 beat punished 그래서 너무 두abilities",

   "T0055G0003S0009.wav": "Her tricks had fallen in, making her look
old.",

   "T0055G0007S0001.wav": "No one could know why he did like that.",

   "T0055G0064S0018.wav": "Attack from above.",
   "T0055G0030S0021.wav": "No me, okay, I'm in sleep.",
   "T0055G0023S0004.wav": "The box is cute of the tree.",
   "T0055G0002S0001.wav": "오늘도",
   "T0055G0066S0004.wav": "He asked them to discontinue flights over the
island.",
   "T0055G0033S0001.wav": "I can lend this book to you. I had it date two
weeks ago."
}
# Gender metadata
gender_metadata = {
   "T0055G0032S0125.wav": "female",
   "T0055G0009S0148.wav": "male",
   "T0055G0003S0009.wav": "female",
   "T0055G0007S0001.wav": "male",
   "T0055G0064S0018.wav": "male",
   "T0055G0030S0021.wav": "female",
   "T0055G0023S0004.wav": "male",
   "T0055G0002S0001.wav": "female",
   "T0055G0066S0004.wav": "male",
   "T0055G0033S0001.wav": "female"
}
# Load ground truth text files
ground_truth = {}
for txt_file in os.listdir(wav_folder):
   if txt_file.endswith('.txt'):
       with open(os.path.join(wav_folder, txt_file), 'r',
encoding='utf-8') as f:
           ground_truth[txt_file.replace('.txt', '.wav')] =
f.read().strip()
```

```
# Variables for male and female WER/CER

male_total_words = female_total_words = 0

male_total_chars = female_total_chars = 0

male_word_errors = female_word_errors = 0

male_char_errors = female_char_errors = 0

# Calculate WER and CER by gender

for filename, generated_text in transcriptions.items():

   if filename in ground_truth and filename in gender_metadata:

       reference_text = ground_truth[filename]

       word_error_rate = wer(reference_text, generated_text)

       char_error_rate = cer(reference_text, generated_text)

       word_count = len(reference_text.split())

       char_count = len(reference_text.replace(" ", ""))  # Exclude spaces

       if gender_metadata[filename] == "male":

           male_total_words += word_count

           male_total_chars += char_count

           male_word_errors += word_error_rate * word_count

           male_char_errors += char_error_rate * char_count

       elif gender_metadata[filename] == "female":

           female_total_words += word_count

           female_total_chars += char_count

           female_word_errors += word_error_rate * word_count

           female_char_errors += char_error_rate * char_count

# Calculate overall WER and CER by gender

male_wer = male_word_errors / male_total_words if male_total_words > 0 else 0

female_wer = female_word_errors / female_total_words if female_total_words > 0 else 0
```

```
male_cer = male_char_errors / male_total_chars if male_total_chars > 0 else 0

female_cer = female_char_errors / female_total_chars if female_total_chars > 0 else 0

# Print results

print("\nOverall Results by Gender:")

print(f"Male WER: {male_wer:.2%}")

print(f"Male CER: {male_cer:.2%}")

print(f"Female WER: {female_wer:.2%}")

print(f"Female CER: {female_cer:.2%}")
```

REFERENCES

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88. https://doi.org/10.1080/02763869.2018.1404391

McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, *99*, 28–37. https://doi.org/10.1016/j.chb.2019.05.009

Mozilla Common Voice Project. (2024). Common Voice Delta Segment 19.0 [Database]. https://commonvoice.mozilla.org/en/datasets

Nexdata. (2024). English (China) Scripted Monologue Smartphone Speech [Database]. https://www.nexdata.ai/datasets/speechrecog/32?source=Github

Pal, D., Arpnikanondt, C., Funilkul, S., & Varadarajan, V. (2019). User experience with Smart Voice assistants: The accent perspective. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/icccnt45670.2019.8944754