CSE 5249

VoiceQuity

Project Final Report

Team
Nikhil Pavan Kanaka
Srishti Ginjala

May 1, 2024

Contents

1 Abstract **1**

2 Introduction **1**

3 Literature Review **2**

4 Methodology and Results **3**
4.1 Model Overview 3
4.2 Dataset 3
4.3 Approach Overview 4
4.3.1 Baseline Accuracy Evaluation 4
4.3.2 Advancing to Real-World Simulations 6

5 Discussion **14**

6 Implications **15**

7 Conclusion **15**

1 Abstract

The advancement of Automatic Speech Recognition (ASR) systems has revolutionized interaction with technology across diverse applications, including smart homes and robotics. However, the widespread deployment of ASRs necessitates rigorous assessment to ensure non-discriminatory performance across varied user demographics and operational environments. This project, VoiceQuity, addresses the critical challenge of inherent biases in voice-activated technologies and their impact on users of different ages and genders. We introduce a comprehensive testing framework that includes a novel set of metamorphic transformations to simulate a wide range of environmental conditions. By applying these transformations across a comprehensive collection of voice utterances that represent a diverse demographic spectrum, we systematically evaluate the discrepancies in voice transcription accuracy. Our findings reveal significant disparities, with overall recognition accuracy dropping to 17.91% post-transformation from an initial 67.34%, highlighting the pronounced effect of simulated biases. Furthermore, this investigation enables us to pinpoint specific scenarios where ASR systems fail to deliver equitable results, thereby allowing for targeted enhancements to improve fairness. The project not only contributes to the technical field by advancing our understanding of ASR system limitations but also proposes actionable insights for developing more robust and inclusive voice recognition technologies. By ensuring equitable performance, VoiceQuity aims to make digital interaction accessible and reliable for all users, irrespective of their age, gender, or the complexity of their environment.

2 Introduction

Automatic Speech Recognition (ASR) systems, such as Whisper, have revolutionized the way we interact with technology, enabling efficient and accurate speech-to-text processing. These systems are trained on large datasets of transcribed audio recordings, utilizing machine learning algorithms to learn patterns and relationships between spoken language and text. The training process typically involves optimizing the model's performance on a validation set, with the goal of achieving high accuracy and robustness.

The evaluation of ASR systems is typically done under controlled laboratory conditions, using standardized datasets and metrics such as Word Error Rate (WER) or Character Error Rate (CER). However, these idealized testing conditions may not accurately reflect real-world scenarios, where speech can be affected by various factors such as background noise, accents, dialects, and speaking styles.

Recent studies have highlighted the potential biases of ASR systems against certain demographic groups, including women and older adults. These biases can lead to reduced accuracy and unfair outcomes, perpetuating existing social inequalities. For instance, ASR systems may struggle to recognize speech from individuals with non-standard accents or speaking styles, resulting in reduced accuracy and potential misinterpretation.

To simulate real-world conditions and uncover potential biases, metamorphic transformations can be applied to the test datasets. These transformations aim to mimic the variations and noise present in real-world speech, such as adding background noise, modifying pitch and tone, or introducing hesitations and filler words. By testing the ASR system on these transformed datasets, we can gain a deeper understanding of its robustness and potential biases under diverse conditions.

In this project, we aim to investigate the accuracy and bias of the Whisper ASR system on

datasets with varying gender and age demographics, under both idealized and real-world simulated conditions. By applying metamorphic transformations, we seek to uncover potential biases and explore strategies for mitigating them, ultimately contributing to the development of more inclusive and robust ASR systems.

3 Literature Review

The influence of demographic factors on the accuracy and fairness of Automatic Speech Recognition (ASR) systems has garnered significant attention in recent research. Various studies have explored how elements such as environmental conditions, dataset characteristics, and speaker demographics impact the performance of these technologies. This literature review synthesizes key findings from notable research efforts in the field, examining both the technical challenges and the methodologies proposed to address these disparities. By analyzing how demographic factors such as age, gender, and dialect influence ASR accuracy, these studies collectively contribute to the ongoing development of more equitable and effective speech recognition technologies.

G. Ceidaite & L. Telksnys (2015)[1] investigated the influence of environmental conditions and training datasets on the accuracy of speech recognition. They highlighted that the environment and the dataset size are significant factors affecting accuracy, suggesting that demographic factors like the dialect or language of the environment might also play a role.

D. Gillick (2010)[2] explored how conversational word usage can predict speaker demographics, such as age and gender, showing that these factors can be discerned from speech patterns with a reasonable degree of accuracy.

Leda Sari et al. (2021)[3] addressed fairness in Automatic Speech Recognition (ASR), presenting a method to train ASR systems to minimize performance discrepancies across demographic groups by counterfactually modifying demographic attributes during training. This approach aimed to ensure consistent recognition accuracy irrespective of demographic factors.

G. Fenu et al. (2020)[4] discussed the performance disparities in deep speaker recognition systems based on sensitive attributes like gender. They found that balancing demographic representation in training datasets could mitigate these biases, leading to fairer and more accurate recognition across different demographic groups (Fenu, Medda, Marras, & Meloni, 2020).

Rajan, S., Udeshi, S., & Chattopadhyay, S. (2021)[5] introduced AequeVox, an automated framework for evaluating the fairness of Automatic Speech Recognition (ASR) systems across different demographic groups. AequeVox simulates various environments to test ASR effectiveness and uses fault localization to identify specific words that demonstrate bias in recognition. Their findings highlighted significant disparities, with non-native English, female, and Nigerian English speakers experiencing substantially higher error rates compared to native English, male, and UK Midlands speakers (Rajan, Udeshi, & Chattopadhyay, 2021).

These studies clearly demonstrates that demographic factors significantly influence the performance of ASR systems. From environmental impacts to dataset diversity and algorithmic fairness, the breadth of research underscores the complexity of achieving high accuracy and fairness in speech recognition technologies. The studies highlight the need for robust, inclusive training datasets and the implementation of novel methodologies such as counterfactual modifications to demographic attributes during training. These approaches are critical in minimizing biases and ensuring that ASR systems perform equitably across

all user groups.
Overall, this body of work not only advances our understanding of the technical challenges in ASR but also guides future innovations towards more just and effective implementations.

4 Methodology and Results

4.1 Model Overview

OpenAI's Whisper is an end-to-end neural network-based automatic speech recognition (ASR) model designed to convert spoken language into text. The version utilized in this project was "Whisper Base," a medium-sized variant of the model which offers a balance between performance and computational efficiency. Whisper Base is pre-trained on a diverse dataset comprising thousands of hours of multilingual and multitask supervised data collected from the web. This extensive training enables the model to recognize and transcribe speech across various languages and dialects.

4.2 Dataset

The project utilized the "Biometrics Visions and Computing (BVC) Gender & Age from Voice Dataset"[6] to assess the performance of the OpenAI Whisper model in transcribing voice recordings and examining potential biases related to gender and age. This dataset comprises voice utterances from a diverse group of 526 individuals, each contributing between one to five voice recordings, leading to a robust collection suitable for detailed analysis. The gender distribution within the dataset includes 336 male and 190 female participants, providing a significant sample size to analyze gender-based differences in transcription accuracy. Notably, we have ground truth data available to validate against, ensuring a high degree of accuracy in our analysis.

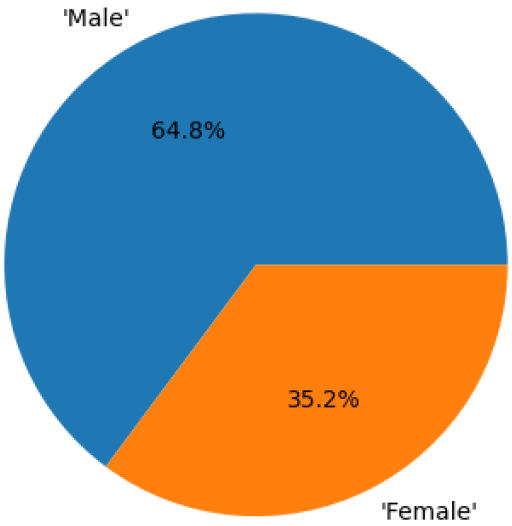

Figure 1. Gender distribution

Each voice recording in the dataset is uniquely identified, enabling precise tracking and analysis across different experiments. The dataset is well-balanced in terms of genders

and age groups, though the specific age group distribution is uniformly anonymized to maintain participant privacy. This segmentation allows for nuanced analysis of age-related variations in voice recognition accuracy.

The recordings are primarily in English, with a minor portion in other languages. The quality of recordings varies, including both studio-quality clips and more casual settings, which introduces realistic variability in audio quality.

4.3 Approach Overview

We employed two iterative approaches to evaluate the transcription accuracy of the OpenAI Whisper model across various demographics and simulated real-world conditions.
Baseline Accuracy Evaluation involves an initial assessment of the model's performance in transcribing audio across different genders and ages under controlled conditions to establish a baseline for accuracy.
Following the baseline evaluation, Simulation of Real-World Scenarios introduces metamorphic transformations to the data. This step is designed to mimic real-world scenarios that may not be well-represented in the initial dataset, allowing us to examine the model's robustness and uncover any latent biases. Together, these approaches provide a comprehensive analysis of the model's performance and its implications in practical applications.

4.3.1 Baseline Accuracy Evaluation

In the initial phase of our project, we aimed to establish a baseline for the transcription accuracy of the OpenAI Whisper model by evaluating its performance across different demographic groups defined by gender and age. To achieve this, we utilized the BVC dataset, which is composed of audio recordings from a diverse group of speakers, each labeled with demographic information that accurately reflects their gender and age.

Approach

We processed the audio files from the BVC dataset using the Whisper model to generate transcriptions. Each transcription was then compared to the corresponding ground truth data provided with the dataset. The evaluation metric was accuracy, which we calculated by measuring the similarity between the Whisper-generated text and the ground truth using the word-level Levenshtein distance. This metric assesses text similarity by quantifying the minimum number of single-character edits required to change one word into another, thereby evaluating both word accuracy and sequence alignment.

To understand the model's performance across different demographics, we segregated the dataset into various groups based on the recorded demographic attributes:
Gender: Male, Female
Age Groups: Continuous values from ages 15 to 35
This segregation allowed us to assess whether there were any initial discrepancies in model performance when transcribing voices from different genders and age groups.

Results

The initial evaluation of the OpenAI Whisper model using the BVC dataset provided insightful results regarding its transcription accuracy across different demographic groups defined by gender and age. The overall accuracy achieved was 67.34%.

$$ASR_{Err}(👨💬) \approx ASR_{Err}(👩💬)$$

Figure 2. Baseline Error Comparison

Accuracy by Gender

The results displayed a moderate variance in accuracy between genders:
Female: 69%
Male: 64%

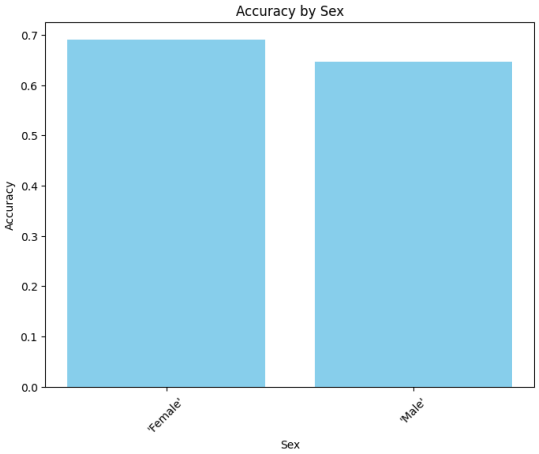

Figure 3. Sex vs Accuracy

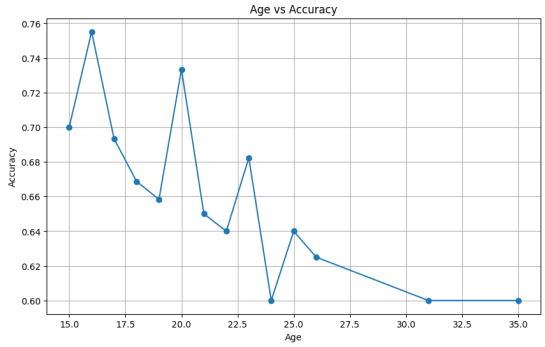

Figure 4. Age vs Accuracy

These differences suggest some variation in how well the model handles audio from male versus female speakers, although the disparity is not significantly large. This indicates a generally balanced performance across genders with a slight advantage in favor of female voices.

Accuracy by Age

The accuracy across different age groups varied, showing a range of performance that seemed to slightly favor younger voices. The accuracy for specific age groups is as follows:

Ages 15-20 showed relatively higher accuracy, with the highest being 73.33% at age 20. A gradual decrease in accuracy was observed in the early twenties, with ages 22 and 24 recording lower scores of 64% and 60%, respectively. There is a noticeable drop for age groups 26, 31, and 35, all scoring around 60-62.5%. The age-related data indicates that the Whisper model tends to perform better on younger voices, which may be due to clearer articulation or recording quality factors inherent in the dataset.

Discussion

The results indicate that while there is some variation in accuracy between different genders and age groups, the overall disparity is not pronounced. This suggests that the Whisper

model, under controlled conditions without real-world audio complexities, provides a fairly uniform performance across the evaluated demographic categories. The slight variations observed are within an acceptable range, indicating no significant bias at this stage of testing.

These findings set the stage for further tests involving metamorphic transformations, which aim to simulate more realistic and challenging audio scenarios. This next step will be crucial to verify if the initial apparent lack of bias holds under conditions that more closely resemble real-world usage.

4.3.2 Advancing to Real-World Simulations

In the second phase of our project, we aimed to simulate real-world scenarios to evaluate how the Whisper model's performance would vary in less controlled environments compared to the baseline accuracy established in Baseline Accuracy Evaluation.

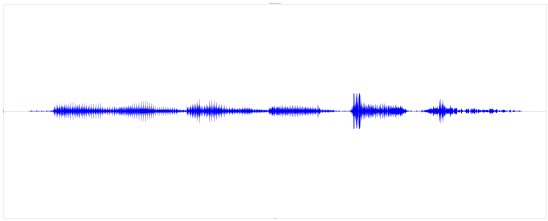

Figure 5. Original waveform

This phase involved introducing metamorphic transformations to the audio data from the BVC dataset, mimicking common distortions and variations that occur in real-life audio recordings. The objective was to uncover any latent biases or performance drops when the model faced more challenging conditions.

Metamorphic Transformations

We applied a series of metamorphic transformations to the original audio files to simulate real-world acoustic phenomena. These transformations included:

Uniform Noise:
Adding uniform noise across the audio spectrum to simulate background noise.
The audio is subjected to a transformation where uniform noise is added to the waveform. The noise is characterized by a standard deviation of 0.1, affecting the entire audio sample uniformly.

```
def apply_uniform_noise(y, noise_range=0.1):
return y + np.random.normal(0, noise_range,
    y.shape)
```

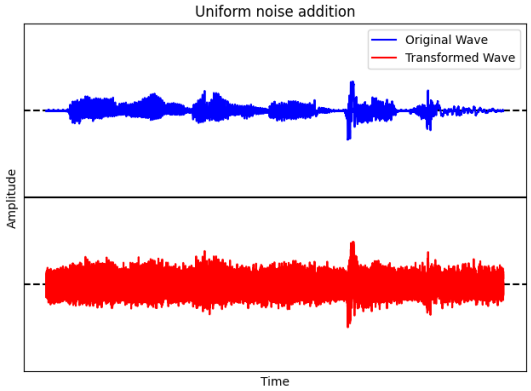

Figure 6

The transcription model exhibits significant variability in accuracy across different ages when subjected to uniform noise, suggesting potential age-related biases. Notably, accuracy is especially low for younger ages, with a minimum accuracy of 0.0 observed for ages 31 and 35, and peaks modestly at age 21 with an accuracy of 0.2375. The accuracy appears to fluctuate without a clear trend across the remaining ages.

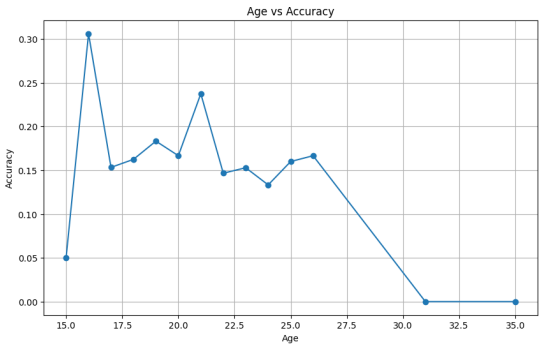

Figure 7

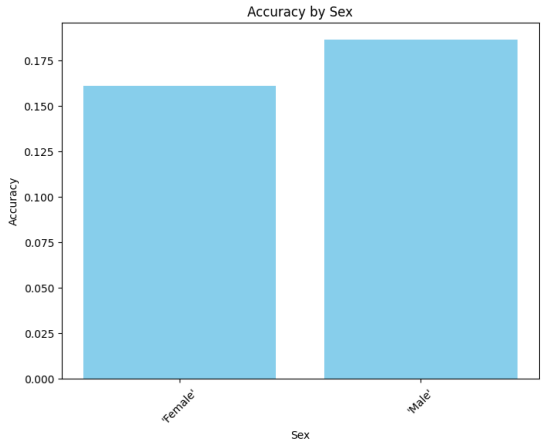

Figure 8

In terms of sex-based differences, the model shows a slightly higher accuracy for male voices (0.1864) compared to female voices (0.1610), which could imply a bias favoring male speech in noisy environments. The overall accuracy of the model under these conditions is 0.1712, highlighting a general decline in performance due to the added noise, which could affect its reliability in real-world applications where background noise is common.

Frame Drop: Intentionally dropping frames from the audio to mimic packet loss in digital communications.
The audio is subjected to selective muting where ten segments, each 1250 samples long, are silenced throughout the waveform. These muted segments are evenly distributed across the entire length of the audio, excluding the very beginning and end, to simulate the loss of data that might occur in real-world audio processing scenarios.

```
def apply_frame_drop(y, num_clips=10,
    clip_length=1250):
y_transformed = copy.deepcopy(y)
total_samples = len(y)
clip_points = np.linspace(0, total_samples,
    num_clips + 2)[1:-1]
for point in clip_points:
    start = int(point)
    end = min(int(point + clip_length),
        total_samples)
    y_transformed[start:end] = 0
return y_transformed
```

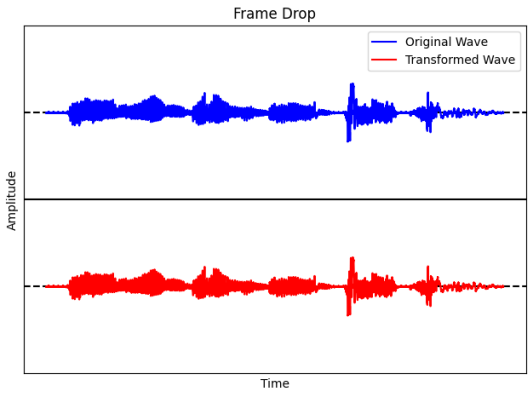

Figure 9

In our investigation of the transcription model's performance under conditions simulating real-world audio loss, we observed distinct variations in transcription accuracy across different age groups and sexes. The model's accuracy showed notable discrepancies among age groups, reaching its highest at 69.4% for individuals aged 16 and its lowest at 45% for those aged 15. Such variability underscores potential age-related biases that merit further scrutiny. Meanwhile, the transcription accuracy for female voices averaged 62.7%, slightly exceeding the 59.9% accuracy observed for male voices, suggesting a modest but consistent advantage for female speech under these testing conditions.

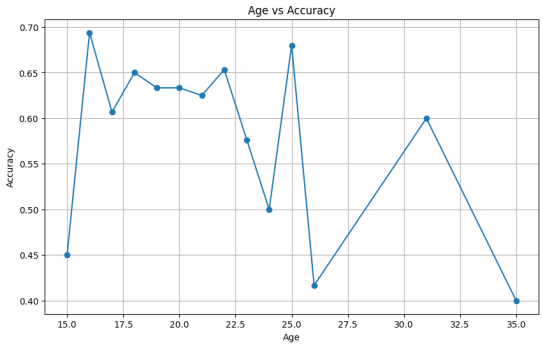

Figure 10

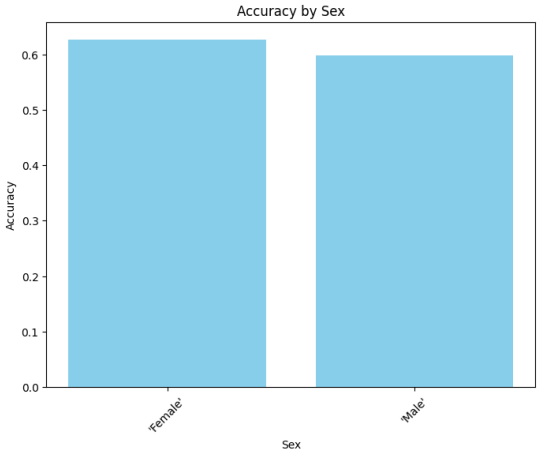

Figure 11

The overall accuracy rate of the transcription model was determined to be approximately 61.6%, indicating a resilient yet imperfect performance amidst the simulated audio disruptions.

Frequency Scaling: Modifying the frequency characteristics of the audio to represent different recording qualities.
The audio is subjected to a transformation where its frequency is scaled down by a factor of 0.5. This manipulation effectively reduces the audio's speed and pitch, simulating scenarios where speech patterns may vary significantly from the standard recording conditions.

```
def apply_frequency_scaling(y,
    scaling_factor=0.5):
    return np.interp(np.arange(0, len(y),
        scaling_factor), np.arange(0, len(y)
        ), y)
```

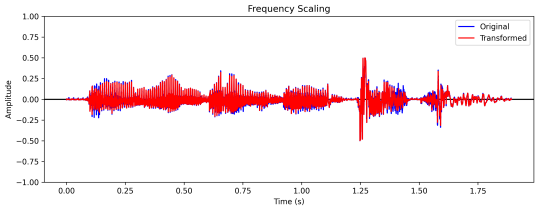

Figure 12

The transcription model demonstrates notable variability in accuracy across different ages when exposed to scaled-down frequency transformations, indicative of potential age-related biases. Accuracy was particularly low for younger ages, with a minimum accuracy of 0.0 observed for age 31, whereas it peaked at age 15 with an accuracy of 0.4. This fluctuation in accuracy across different ages suggests a lack of consistent response to altered auditory conditions. In terms of sex-based differences, the model exhibited a higher accuracy for female voices, achieving an accuracy of 0.436 compared to only 0.130 for male voices. This could indicate a bias favoring female speech patterns under modified frequency conditions.

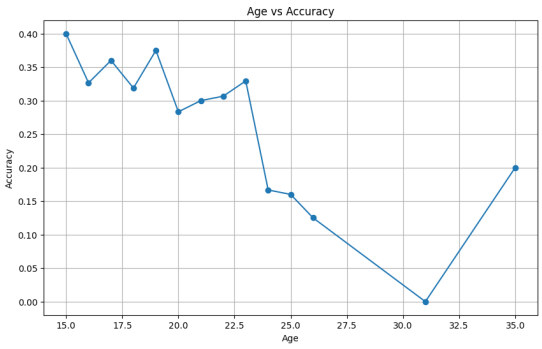

Figure 13

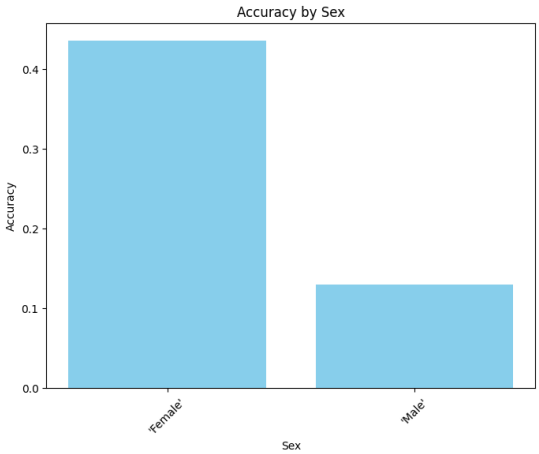

Figure 14

Overall, the model achieved an accuracy of 0.314, underscoring a significant drop in performance that could compromise its reliability in real-world scenarios where audio quality varies.

Amplitude Modification: Altering the amplitude of the audio signal to simulate varying speaker volumes.
The audio waveform is scaled down to 25% of its original amplitude. This adjustment is intended to test the transcription system's ability to accurately process and recognize speech from audio signals that are significantly quieter than typical levels.

```
def apply_amplitude_modification(y,
    scale_factor=0.25):
    return y * scale_factor
```

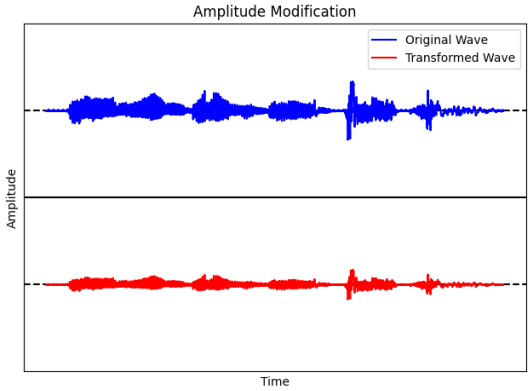

Figure 15

The transcription model's performance varies significantly when analyzing accuracy across different ages after subjecting the audio to a metamorphic transformation where the amplitude is scaled down to 25% of its original level. This simulation of quieter audio conditions reveals potential age-related biases in the model. The model achieves the highest accuracy for ages 16 and 20, with values of 0.7347 and 0.7167 respectively, but the accuracy decreases for ages 22 and 24, recording lower values of 0.6400 and 0.5667. This indicates that the model's ability to handle speech from younger individuals might be less robust compared to other age groups. Moreover, the performance differences between male and female voices are noticeable; the accuracy for female voices is approximately 0.6948, which is notably higher than the 0.6384 observed for male voices under these conditions. This could suggest a potential favoritism towards female speech in quieter environments.

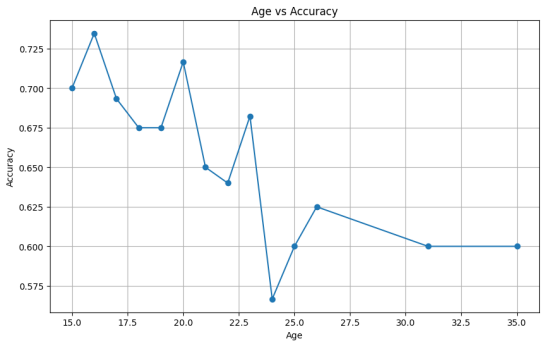

Figure 16

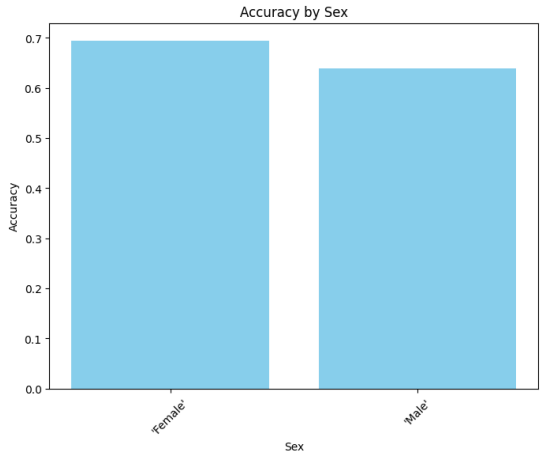

Figure 17

The overall accuracy under these altered amplitude conditions is 0.6723, indicating a general decline in model performance, which may impact its reliability in real-world applications where such variations in speaker volume are common.

Clipping: Clipping the audio signal peaks to mimic distortion typically found in overly loud recordings.
The audio is subjected to a transformation where each waveform sample is constrained between -0.025 and 0.025. This process ensures that the audio signal's amplitude does not exceed these specified bounds, effectively compressing extreme values.

```
def apply_clipping(y, CIPPING_VAL=0.025):
    y_transform = [min(CIPPING_VAL, elem)
        for elem in y]
    y_transform = [max(-1*CIPPING_VAL, elem)
         for elem in y_transform]
    y_transform = np.array(y_transform)
    return y_transform
```

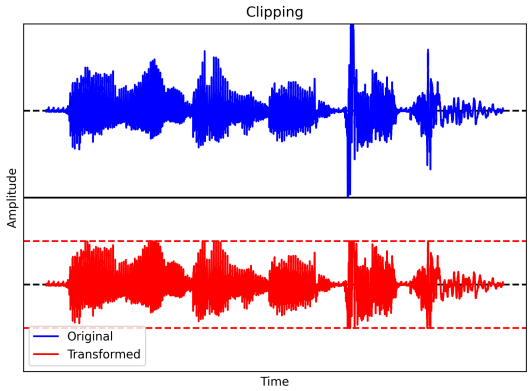

Figure 18

The transcription model demonstrates significant variability in accuracy across different age groups when subjected to audio transformations intended to simulate real-world distortions, such as clipping the audio signal peaks. Notably, the accuracy for younger age groups (15-26) generally hovers above 50%, with a peak at age 16 (65.31%) and the lowest at age 26 (33.33%). However, a substantial drop in accuracy is observed in the older age group, with ages 31 and 35 exhibiting lower accuracies of 40% and 20%, respectively. This suggests potential age-related biases, particularly under conditions of audio distortion. When analyzing sex-based differences, the model exhibits higher accuracy for female voices (53.18%) compared to male voices (49.15%), contradicting common biases favoring male speech.

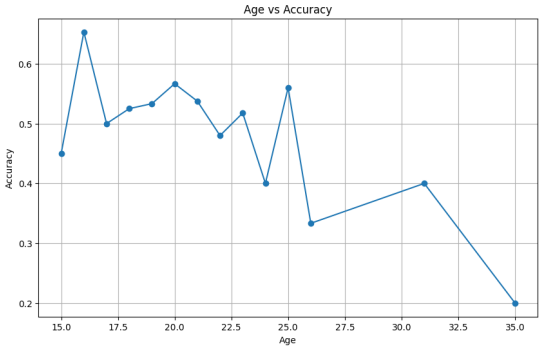

Figure 19

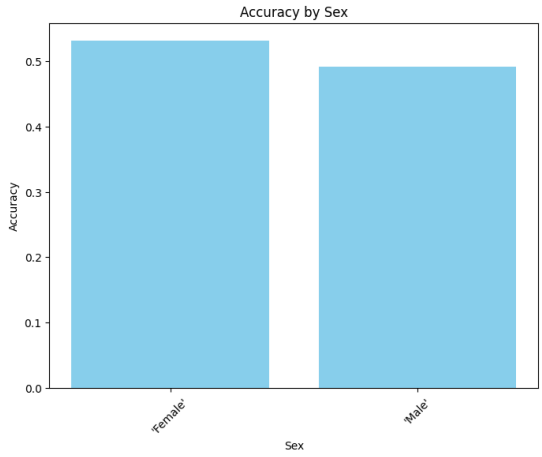

Figure 20

The overall accuracy of the model under these transformation conditions is 51.58%, indicating a moderate decline in performance, which could impact the model's reliability in practical, noisy environments.

High Pass Filtering: Applying a high pass filter to remove lower frequencies, simulating recordings made with low-quality microphones.
A high-pass filter with a cutoff frequency of 5000 Hz is applied to the audio signal. This filter allows frequencies higher than 5000 Hz to pass through while attenuating frequencies below this threshold, effectively reducing lower frequency components in the audio waveform.

```
def apply_high_pass(y, cutoff=5000,
    sample_rate=22050):
    from scipy.signal import butter, lfilter
    b, a = butter(1, cutoff / (0.5 *
        sample_rate), btype='high', analog=
        False)
    return lfilter(b, a, y)
```

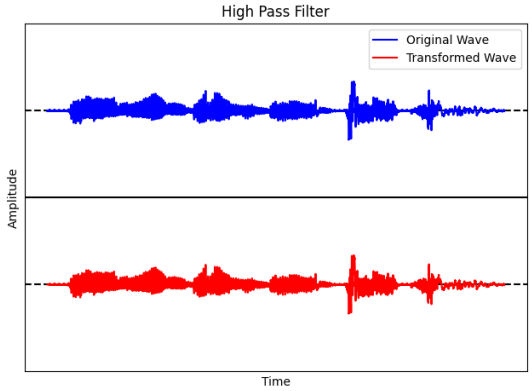

Figure 21

The transcription model exhibits significant variability in accuracy across different ages when subjected to a high-pass filter, simulating recordings made with low-quality microphones. Notably, accuracy is especially low for younger ages, with a minimum accuracy of 0.4 observed for age 35, and peaks at age 19 with an accuracy of 0.7. The accuracy appears to fluctuate without a clear trend across the remaining ages.

In terms of sex-based differences, the model shows a slightly higher accuracy for female voices (0.6891) compared to male voices (0.6215), which could imply a bias favoring female speech in noisy environments.

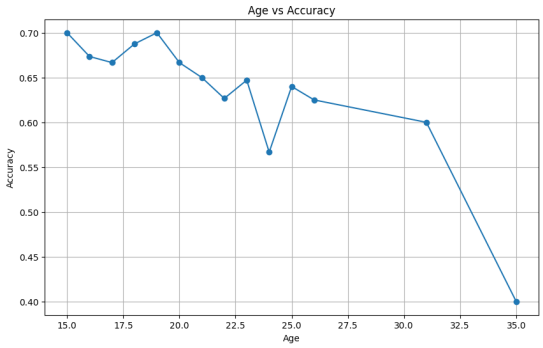

Figure 22

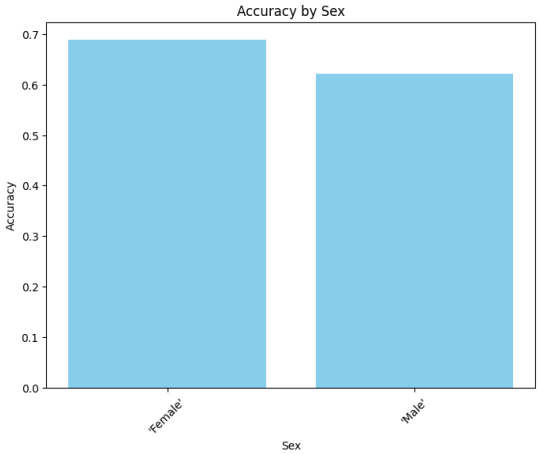

Figure 23

The overall accuracy of the model under these conditions is 0.6553, highlighting a general decline in performance due to the added noise, which could affect its reliability in real-world applications where background noise is common.

Low Pass Filtering: Applying a low pass filter to eliminate higher frequencies, representing the effect of an obstructed microphone.
The audio is subjected to a low-pass filter transformation where frequencies above 5000 Hz are attenuated, using a digital Butterworth filter with a normalized cutoff frequency. This process is applied to ensure that only frequencies below the cutoff are retained, enhancing focus on lower frequency components of the audio.

```
def apply_low_pass(y, cutoff=5000,
    sample_rate=22050):
    from scipy.signal import butter, lfilter
    b, a = butter(1, cutoff / (0.9 *
        sample_rate), btype='low', analog=
        False)
    return lfilter(b, a, y)
```

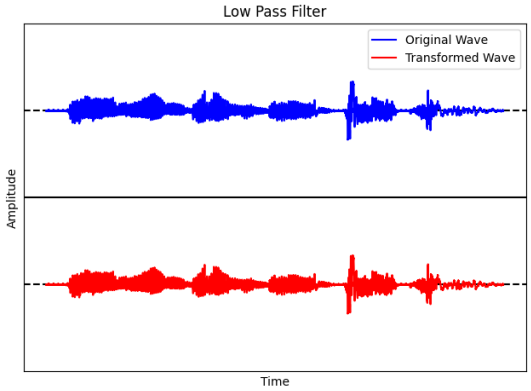

Figure 24

The transcription model exhibits significant variability in accuracy across different ages when subjected to a low-pass filter transformation, suggesting potential age-related biases. Notably, accuracy is especially low for younger ages, with a minimum accuracy of 0.6 observed for ages 31 and 35, and peaks moderately at age 16 with an accuracy of 0.7551. The accuracy appears to fluctuate without a clear trend across the remaining ages.

In terms of sex-based differences, the model shows a slightly higher accuracy for female voices (0.6873) compared to male voices (0.6469), which could imply a bias favoring female speech in noisy environments.

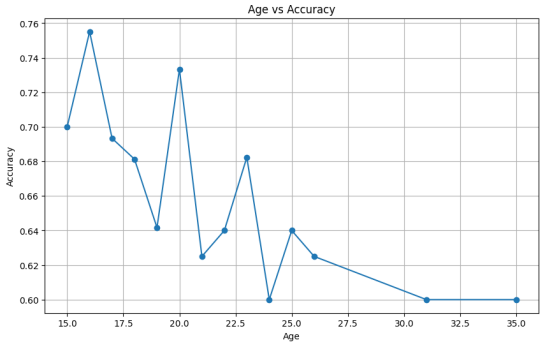

Figure 25

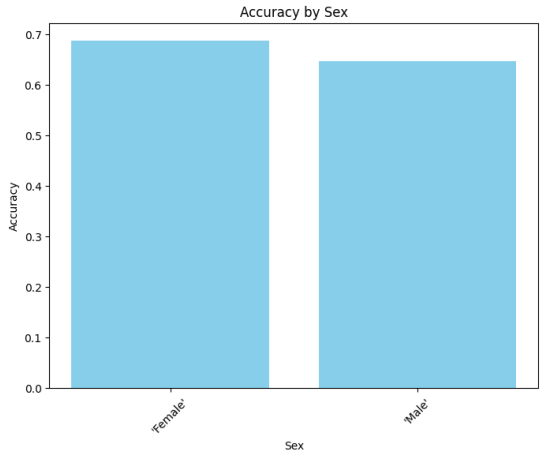

Figure 26

The overall accuracy of the model under these conditions is 0.667, highlighting a general decline in performance due to the added noise, which could affect its reliability in real-world applications where background noise is common.

After applying the transformations, we processed the modified audio files through the Whisper model to generate transcriptions. These transcriptions were then compared to the original ground truth data, using the same accuracy metrics as in Approach 1. This allowed us to directly compare the effects of real-world conditions on transcription accuracy.

5 Discussion

The analysis of automated speech recognition (ASR) accuracy following various metamorphic transformations provides critical insights into the performance disparities influenced by gender and age. Our study uncovered several key patterns in ASR performance degradation under conditions such as low pass filtering, clipping, uniform noise, high pass filtering, amplitude modification, frame dropping, and frequency scaling.
Each of these transformations was designed to test the robustness of ASR systems under different types of audio signal degradation, simulating real-world scenarios that might affect users differently based on their demographic characteristics.

Gender-Based Analysis

The results indicate that there is a considerable difference in ASR accuracy between genders, particularly in the frequency scaling transformation, where males showed significantly lower accuracy compared to females. This suggests that frequency scaling altering the pitch and timbre of the audio may disproportionately affect male voices, which typically have lower fundamental frequencies. This could be due to the inherent characteristics of the ASR systems that may have been trained or optimized primarily on voice samples that do not adequately represent the lower frequency ranges.

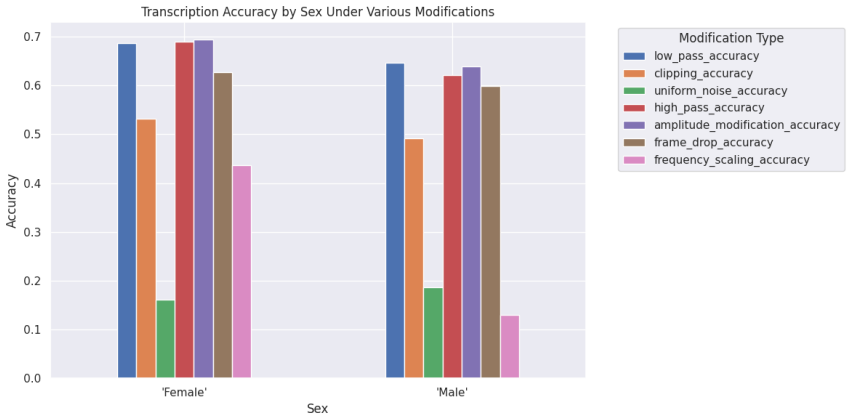

Figure 27

Additionally, other transformations like clipping and uniform noise also show notable disparities. Although both genders experience accuracy reductions, the drop is more pronounced in males for clipping, which suggests that sudden changes in signal amplitude affect male voices more adversely. This could potentially be attributed to the dynamic range of male voices being broader, thus making them more susceptible to distortion when the amplitude is clipped.

Age-Based Analysis

Across age groups, the ASR system's performance also varied significantly, particularly in younger (below 20 years) and older (above 50 years) age groups. These groups showed the highest drop in accuracy under uniform noise conditions, which could be linked to the speech characteristics like clarity, speed, and pitch variation that differ notably in these age brackets. Younger voices tend to have higher pitches and faster speech rates, which might be less accurately captured when background noise is present.

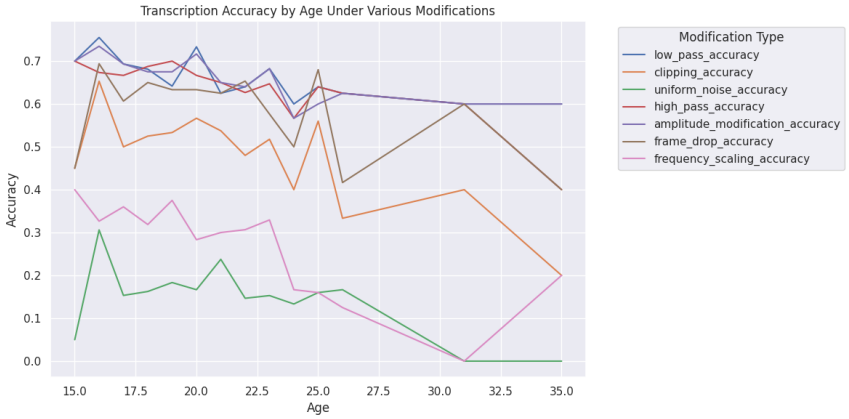

Figure 28

Frequency scaling again shows as a critical transformation where discrepancies are observed, indicating that pitch and timbre variations are not uniformly handled across different ages. The lesser accuracy in older age groups could reflect a lack of robustness in current ASR systems to changes in voice pitch and timbre, which naturally vary with aging due to physiological changes in vocal cords.

6 Implications

Theoretical

These findings have profound implications for the development of more inclusive ASR technologies. They highlight the necessity for training these systems on a more diverse set of voice samples that include a wide range of genders, ages, and other demographic factors. The significant disparities observed suggest that current systems might be employing algorithms that inherently favor certain voice types over others, thereby limiting their effectiveness and accessibility.

Practical

From a practical standpoint, our study suggests several areas for targeted enhancements. ASR developers can use these insights to refine their algorithms, focusing particularly on transformations that showed the highest disparities. For instance, improving noise handling capabilities, especially for uniform noise, could drastically increase the usability of ASR systems in real-world environments where background noise is common. Similarly, enhancing the system's ability to handle frequency variations could make the technology more accessible and equitable for all users, regardless of age or gender.

7 Conclusion

In conclusion, this detailed analysis not only sheds light on the specific areas where ASR technologies falter in terms of equity and performance but also provides a roadmap for future improvements. By addressing these disparities, technology developers can ensure that voice-activated systems are truly inclusive, providing equitable access and performance for all users across the spectrum of human diversity.

References

[1] G. Ceidaite and L. Telksnys. "Analysis of Factors Influencing Accuracy of Speech Recognition". In: *Elektronika Ir Elektrotechnika* 105 (2015), pp. 69–72. DOI: 10.5755/J01.EEE.105.9.9180 (page 2).

[2] D. Gillick. "Can conversational word usage be used to predict speaker demographics?" In: (2010), pp. 1381–1384. DOI: 10.21437/Interspeech.2010-421 (page 2).

[3] Leda Sari, M. Hasegawa-Johnson, and C. Yoo. "Counterfactually Fair Automatic Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3515–3525. DOI: 10.1109/taslp.2021.3126949 (page 2).

[4] G. Fenu et al. "Improving Fairness in Speaker Recognition". In: *Proceedings of the 2020 European Symposium on Software Engineering* (2020). DOI: 10.1145/3393822.3432325 (page 2).

[5] Sai Sathiesh Rajan, Sakshi Udeshi, and Sudipta Chattopadhyay. "AequeVox: Automated Fairness Testing of Speech Recognition Systems". 2022. arXiv: 2110.09843 [cs.LG] (page 2).

[6] Ogechukwu Iloanusi et al. "Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca". In: Nov. 2019, pp. 175–179. DOI: 10.1109/ISCMI47871.2019.9004306 (page 3).

[7] Overleaf. "Learn LaTeX in 30 minutes". URL: https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes.

[8] Scott Pakin. "The comprehensive LaTeX symbol list". 2020. URL: http://tug.ctan.org/info/symbols/comprehensive/symbols-a4.pdf.

[9] Joseph Wright. "`siunitx` – A comprehensive (SI) units package". 2022. URL: https://ctan.org/pkg/siunitx.