# Trusting Synthetic Speech: Robustness, Fidelity, and the Risk of Bias in Hate Speech Transcription

**Anonymous ACL submission**

## Abstract

The rise of synthetic speech audio-based NLP tasks has raised critical questions about the *robustness*, *fidelity*, and *fairness*. This study will empirically examine the relationship between Text-to-Speech (TTS) and Speech-to-Text (STT) models using hate and non-hate speech data. Our evaluation focuses on three key dimensions: (1) *STT robustness*, assessing the accuracy and gender sensitivity of STT models when transcribing synthetic versus human audio; (2) *TTS synthetic audio fidelity*, examining human-likeness and model preference through annotator evaluations and processing speed analysis; and (3) *Impact on hate speech classification*, quantifying how STT and TTS combinations affect downstream toxicity predictions. Our findings show that synthetic audio, especially from Microsoft Edge TTS, outperforms human audio in both transcription accuracy and consistency. WhisperX-Align (extended based on OpenAI's Whisper model) emerges as the most *robust* STT model across tasks, although some systems exhibit notable gender and domain-specific biases. We recommend Microsoft Edge TTS as a high *fidelity* benchmark and SpeechT5 as a human proxy for perceptual evaluation, while highlighting the need for *bias* aware deployment in sensitive applications, such as hate speech detection. The implementation code is publicly available at `https://anonymous.4open.science/r/Can-AI-Replace-Human-Speech-D0EF/`.

## 1 Introduction

The rapid expansion of social media has enabled users worldwide to disseminate their views and ideas at an unprecedented scale. Initially designed to foster connection and learning, these platforms have become breeding grounds for extremist ideologies. Early approaches to detecting harmful content mainly relied on text-based classification methods (Lee and Ram, 2024; Qian et al., 2018). However, as multimodal content, particularly hateful memes and videos, has become more common, research has expanded to include visual (Chen and Pan, 2022; Lee et al., 2021) and audio (An et al., 2024; Atanu et al., 2023; Imbwaga et al., 2024) hate speech detection. Audio-based detection, the latest frontier in the "trilogy of hate," now encompasses text, image, and audio modalities. A significant barrier in audio classification is the lack of large-scale, annotated datasets for domain-specific or toxic speech.

To address this, researchers have turned to Text-to-Speech (TTS) models to generate synthetic audio from existing text-based hate speech datasets, leveraging publicly available textual corpora (Waseem and Hovy, 2016; Ocampo et al., 2023a,b). While synthetic audio offers a scalable alternative to human recordings, it raises critical questions:

1. **STT Robustness:** How well do current STT models transcribe synthetic audio compared to human speech across different voice types and genders? We found that synthetic voices, particularly those from Edge-TTS and SpeechT5, consistently outperform human recordings in terms of accuracy. However, gender bias—especially favoring female voices—emerges across models, with SpeechT5 showing the highest disparity.

2. **TTS synthetic audio fidelity:** Which TTS models best simulate human speech from the perspective of human perception? Through a human-likeness ranking study and transcription speed analysis, we show that SpeechT5 is rated most human-like, while Edge-TTS provides the most efficient and consistent transcriptions. We also reveal that STT systems generally process female voices faster than male ones, introducing potential biases in data generation workflows.

3. **Impact on Hate Speech Classification:** How do combinations of TTS and STT models influence the accuracy of toxicity classification? Our results show that using synthetic audio improves classification performance compared to human speech. Nevertheless, hate samples are more prone to transcription errors, and specific TTS-STT pairings—such as VITS with DeepSpeech—exacerbate these issues.

Together, these findings provide a comprehensive benchmark for evaluating TTS and STT models in both technical and ethical dimensions. We conclude with practical recommendations on model selection, emphasizing the trade-off between transcription fidelity and bias in high-stakes applications like hate speech detection.

## 2 Experiment Settings

Before we investigate the findings, we will provide some background information about each model that will be used throughout this paper. The methods used for data collection and classification. In addition to the technique for audio creation, transcription, and normalization. All the following methods and results were run on a research server. Multiple clusters were used, each with 24 CPU cores and 64 GB of RAM.

### 2.1 TTS Models

TTS is the technology that transforms written text into spoken audio, allowing systems to communicate with users through synthetic speech. We will be using the following TTS models, through out our paper.

- **VITS** (Kim et al., 2021) with `p225` for female voice and `p229` for male voice.
- **SpeechT5** (Ao et al., 2022), `clb` for female voice and `bdl` for male voice.
- **Edge TTS**[1]. We used `Aria` for female voice and `Christopher` for male voice.

These models were chosen for two key reasons: their ease of use and widespread adoption, and their support for both male and female voices. The latter was essential to evaluate whether STT models exhibit inherent gender bias.

[1] https://github.com/rany2/Edge-TTS

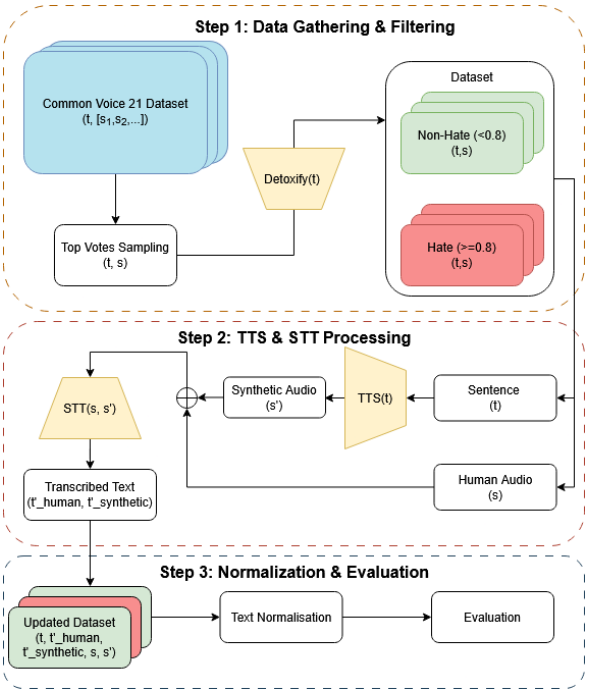


Figure 1: Overview of the data processing pipeline.

### 2.2 STT Models

STT refers to the ability of a system to convert spoken language into written text, enabling machines to interpret and process audio input. Below is a list of STT models used in this paper:

- **WhisperX** (Bain et al., 2023) we utilize both the standard `WhisperX` model and its aligned variant `WhisperX-Align`.
- **Vosk**[2]. In this study, we employ both the `Vosk-Small` and `Vosk-Giga` models.
- **DeepSpeech**[3]. We utilize both the standard `DeepSpeech` model and its variant enhanced with the language scorer `DeepSpeech-Scorer`.

### 2.3 Data Processing Pipeline

#### 2.3.1 Data Gathering & Filtering

The Mozilla Common Voice 21.0 dataset is a large, multilingual speech corpus for automatic speech recognition, crowdsourced from volunteers globally. Version 21.0 includes over 20,000 hours of validated speech across more than 70 languages. Each entry in the dataset consists of a unique `sentence ID`, the spoken `sentence`, the name of

[2] https://alphacephei.com/vosk/
[3] https://github.com/mozilla/DeepSpeech

the audio file, the speaker's gender, and the number of upvotes/downvotes the recording received.

To classify sentences as *hate* or *non-hate*, we used the **toxicity score** provided by the Detoxify model (Hanu and Unitary team, 2020). Detoxify provides several (e.g., identity attack, sexually explicit), we focus on the general toxicity score, as it aligns best with evaluating harmful or hateful speech. we chose the 'original' model, which is based on BERT. This model was chosen because other models yielded worse results, likely due to domain mismatch, as many were primarily trained on social media data.

To set an appropriate toxicity threshold for labelling hate content, we manually evaluated 50 hate and 50 non-hate sentences. We tested a range of threshold scores (0.6 - 0.9) and found that a threshold of 0.8 yielded the best performance in correctly identifying hateful content. These results can be found in Table 1.

Table 1: Performance metrics for Detoxify at varying toxicity thresholds (manual annotation)

| Threshold | Precision | Recall | F1 Score |
|---|---|---|---|
| 0.6 | 80.00 | 96.00 | 87.27 |
| 0.7 | 86.54 | 90.00 | 88.24 |
| 0.8 | 93.33 | 84.00 | 88.42 |
| 0.9 | 92.59 | 50.00 | 64.94 |

For the non-hate data entries, we included only those sentences that had both male and female audio recordings. When multiple recordings existed for a given gender, we selected the one with the highest number of upvotes to prioritise audio quality. In contrast, for the hate entries, where the number of available samples was more limited, we relaxed the gender-pairing requirement and selected only a single high-quality recording per sentence, again based on upvotes. Table 2 summarises the number of sentences included in the hate and non-hate categories.

Table 2: Number of sentences in *hate* and *non-hate*.

| Type | Male | Female | Unique Sentences |
|---|---|---|---|
| Non-Hate | 24,536 | 24,536 | 24,536 |
| Hate | 692 | 334 | 924 |

#### 2.3.2 TTS & STT Processing

For each sentence, denoted as $t$, we generate the synthetic audio $s'$ by parsing it through the previously selected TTS models. As the STT models require the audio file to be in the 'wav' format, we convert all audio files from 'mp3' into 'wav' by using ffmpeg[4]. Then, for all the audio files (synthetic $s'$ and human $s$), we process them through the selected STT, resulting in $t'_{\text{synthetic}}$ and $t'_{\text{human}}$.

#### 2.3.3 Text Normalisation

To minimise transcription bias, particularly variations in how different models transcribe numbers, we convert all numerals into their word representations (e.g., 42 $\rightarrow$ forty-two) using the inflect library[5]. Following this, we normalise the transcribed text by converting all characters to lowercase, removing terminal punctuation marks(".", ",", "?" and "!"), and eliminating any extra spacing around symbols. This normalisation process ensures a fair and consistent basis for comparing the outputs of different STT models.

## 3 Experimental Results

The evaluation metrics used in this paper can be found in appendix D.

### 3.1 STT Robustness

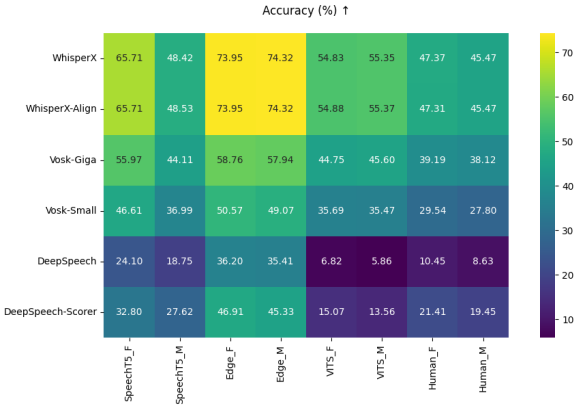


Figure 2: STT robustness on **non-hate** speech: Comparison of transcription Accuracy across TTS models and human audio (by gender).

In this section, we introduce a two-stage robustness analysis to provide a comprehensive understanding of how resilient STT models are to variations in *voice type*, *gender*, and *semantic content*.

**First**, we assess the global robustness of STT models when transcribing synthetic audio compared to real human audio. Formally, let $t$ denote the ground truth sentence and $s_{\text{human}}$ the corresponding human-spoken audio. We proceed as follows:

1. **TTS Generation:** Apply a text-to-speech (TTS) model $\mathcal{T}$ to generate synthetic audio

[4] https://ffmpeg.org/
[5] https://pypi.org/project/inflect/

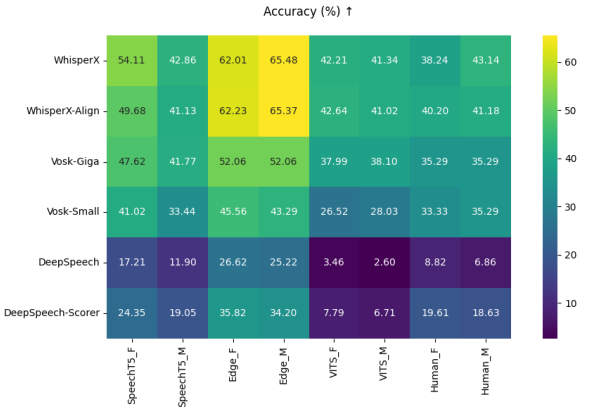


Figure 3: STT robustness on **hate** speech: Comparison of transcription Accuracy across TTS models and human audio (by gender).

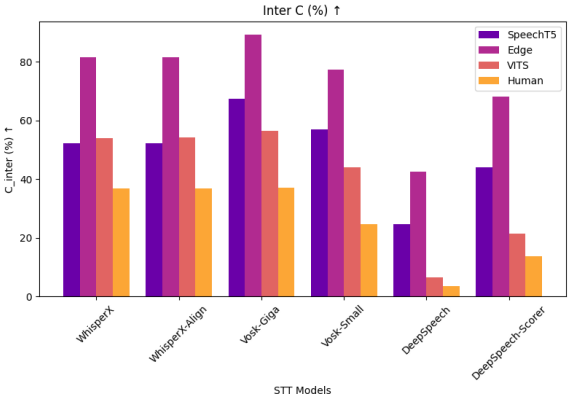


Figure 4: Inter-gender consistency on **non-hate** speech: Measuring STT model sensitivity to gender variation using Inter_C

from the text:

$$s_{\text{synthetic}} = \mathcal{T}(t).$$

2. **STT Transcription:** Apply the same speech-to-text (STT) model $\mathcal{S}$ to both audio samples:

$$t_{\text{human}} = \mathcal{S}(s_{\text{human}}), \quad t_{\text{synthetic}} = \mathcal{S}(s_{\text{synthetic}})$$

3. **Similarity Comparison:** Measure the similarity between the transcriptions and the ground truth using an evaluation metric $D(\cdot, \cdot)$ (i.e., Transcription Accuracy, Absolute Character Distance, and WER):

$$D_{\text{human}} = D(t, t_{\text{human}}), D_{\text{synthetic}} = D(t, t_{\text{synthetic}})$$

We then compare $D_{\text{human}}$ and $D_{\text{synthetic}}$ to evaluate the robustness of the STT model when transcribing synthetic versus real human speech. The experimental results are shown in Figs. 2 and 3 with additional information in appendix, which we will detail the analysis in Section 3.1.1 and 3.1.2, respectively.

**Then**, we examine how STT performance varies across speaker gender for each audio type. Formally, we proceed as follows:

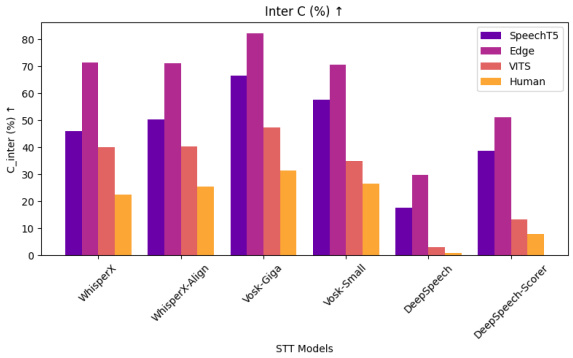


Figure 5: Inter-gender consistency on **hate** speech: Measuring STT model sensitivity to gender variation using Inter_C

1. **TTS Generation:** Apply a text-to-speech (TTS) model $\mathcal{T}$ to generate synthetic audio from the text for both male and female voices:

$$s^{\text{male}}_{\text{synthetic}} = \mathcal{T}_{\text{male}}(t), \quad s^{\text{female}}_{\text{synthetic}} = \mathcal{T}_{\text{female}}(t)$$

2. **STT Transcription:** Apply the same speech-to-text (STT) model $\mathcal{S}$ to the human audio and both synthetic audios:

$$t^{\text{male}}_{\text{human}} = \mathcal{S}(s_{\text{human}}), \quad t^{\text{female}}_{\text{human}} = \mathcal{S}(s_{\text{human}}),$$

$$t^{\text{male}}_{\text{synthetic}} = \mathcal{S}(s^{\text{male}}_{\text{synthetic}}), \quad t^{\text{female}}_{\text{synthetic}} = \mathcal{S}(s^{\text{female}}_{\text{synthetic}})$$

3. **Similarity Comparison:** Measure the similarity between the transcriptions and the ground truth using an evaluation metric $D(\cdot, \cdot)$:

$$D^{\text{male}}_{\text{human}} = D(t, t^{\text{male}}_{\text{human}}), \quad D^{\text{female}}_{\text{human}} = D(t, t^{\text{female}}_{\text{human}}),$$

$$D^{\text{male}}_{\text{synthetic}} = D(t, t^{\text{male}}_{\text{synthetic}}), \quad D^{\text{female}}_{\text{synthetic}} = D(t, t^{\text{female}}_{\text{synthetic}})$$

We then compare $D^{\text{male/female}}_{\text{human}}$, $D^{\text{male/female}}_{\text{synthetic}}$ to evaluate the robustness of the STT model when transcribing real versus synthetic speech and investigate any performance gaps related to speaker gender. Figs. 4 and 5, which we will detail the analysis in Section 3.1.1 and 3.1.2, respectively.

### 3.1.1 Non-Hate

**STT Robustness with Synthetic vs. Human Audio** Fig. 2 presents the global robustness of STT models across synthetic and human audio in the non-hate speech domain, with heatmaps displaying sentence-level accuracy, absolute character difference, and WER.

Performance on human speech varies significantly across STT models, ranging from 10.45% / 8.63% to 47.37% / 45.47% for female and male voices, respectively. This variation highlights that even real human audio poses challenges for

some STT models, particularly DeepSpeech and DeepSpeech-Scorer.

**Edge-TTS emerges as the most robust and balanced TTS model**, outperforming both SpeechT5 and human voices across all STT models. Its performance remains consistently high, with relatively low gender bias, indicating strong generalizability and clarity across genders. This makes Edge-TTS a strong candidate for use as a synthetic benchmark.

Across the STT models, **WhisperX-Align consistently achieves the best results**, regardless of TTS voice or gender, and leads in transcription accuracy, character difference, and WER. It demonstrates a **slight preference for female voices** in non-hate samples, aligning with previously reported gender bias trends.

**Inter-Gender Robustness Analysis** Fig. 4 shows the inter-gender robustness of STT models across synthetic and human audio in the non-hate domain, focusing on sentence-level agreement (InterC), absolute character difference, and WER across male and female voices for the same content.

We observe that **synthetic voices consistently outperform human voices** in terms of inter-gender transcription consistency. Human audio yields the highest character-level differences across genders, mainly when processed by DeepSpeech and DeepSpeech-Scorer, indicating that these models are more sensitive to natural voice variations.

Interestingly, while **VITS** generally underperforms in global accuracy (Fig. 2), it achieves strong inter-gender consistency across STT models. In particular, **VITS paired with DeepSpeech** performs better than human audio, which contrasts with the global results. This suggests that although VITS may struggle with overall transcription accuracy, its generated male and female voices are more acoustically aligned, leading to higher cross-gender consistency.

**SpeechT5**, despite being highly rated for human-likeness and global accuracy, exhibits the **largest gender-based character differences**. This further reinforces concerns about its strong gender bias, already observed in Fig. 2.

Among STT models, **Vosk-Giga** delivers the best inter-gender consistency in this setting, outperforming even WhisperX-Align, which led in global metrics. This suggests that Vosk-Giga is less sensitive to pitch, timbre, or spectral variations introduced by gender shifts.

Finally, the WER results reinforce earlier observations: human audio again shows the most variation between genders, while synthetic voices, especially from Edge-TTS and VITS, are more stable. Overall, this highlights that certain synthetic voices offer not just better average performance, but also stronger consistency across gender variants.

### 3.1.2 Hate

**STT Robustness with Synthetic vs. Human Audio** Fig. 3 illustrates the performance of STT models when transcribing synthetic versus human audio in the hate speech domain.

Overall transcription performance declines across all TTS and STT combinations when transitioning from non-hate to hate speech samples. Human voice performance ranges from 8.82% / 6.86% to 40.20% / 43.14% (female / male), showing slightly reduced variability compared to the non-hate domain.

Despite the domain shift, **Edge-TTS** continues to outperform all other TTS models across all STTs and metrics, maintaining strong accuracy and low character and word error rates. This reinforces its status as the most robust and reliable TTS model across different content types.

Across STT models, **WhisperX-Align remains the top performer** in the evaluation metrics. However, in this domain, the gap between WhisperX-Align and other STTs, such as **Vosk-Giga**, narrows, indicating that more STT models can handle hate speech robustly if paired with strong TTS input.

Another key difference from the non-hate setting is the shift in gender preference. While most models favored female voices in non-hate speech, **hate samples see a partial reversal**: more STT models prefer male voices, and gender bias magnitudes are generally smaller. This may indicate that models perceive aggressive or emotionally charged prosody in male synthetic voices as more intelligible in hate contexts.

**Inter-Gender Robustness Analysis** Fig. 5 examines the inter-gender transcription consistency of STT models in the hate speech domain.

In line with global observations, we find that **all systems perform worse in the hate domain** compared to the non-hate setting.

**Edge-TTS** maintains its strong performance across all metrics. It continues to produce the most consistent outputs between male and female voices, suggesting a balanced acoustic profile across genders even under emotionally or semantically com-

plex content.

Among STT models, **Vosk-Giga** again provides the most stable inter-gender results, outperforming even WhisperX-Align in consistency across genders. This trend mirrors the findings from the non-hate domain. It suggests that Vosk-Giga may be more resilient to pitch and tonal variations introduced by gender, particularly in emotionally charged speech.

### 3.2 TTS Synthetic Audio Fidelity

While identifying the most accurate TTS model is important, it does not necessarily indicate which model best replicates human speech. Therefore, we conducted a dedicated evaluation to determine which TTS model serves as the most suitable substitute for human audio. This distinction is critical, especially if the most human-like model differs from the one that yields the highest transcription accuracy. Additionally, we examine whether listener preferences exhibit gender-based variation, and whether processing time differs across TTS-STT combinations, particularly in cases where gender bias may lead to faster processing. If such disparities exist, they may influence the practical choice of synthetic voices for large-scale data generation.

To evaluate human-likeness and clarity, we conducted a perceptual study involving seven annotators. Each participant was asked to rank non-hate synthetic audio samples on a scale from 1 to 3, where 1 indicates the most human-like and 3 the least. Annotators also indicated which gendered voice (male or female) was more understandable for each TTS model. Each model was represented by three audio samples per gender. The number of samples was determined based on preliminary testing, which showed that annotators were able to form reliable judgements with one to two examples, and three samples provided a good balance of confidence and coverage. Details about metrics used can be found in appendix D.1.

#### 3.2.1 Preferred TTS and Gender

Table 3: Average human-likeness rankings for TTS models by gender (lower is better)

| Gender | Edge (Rank) | SpeechT5 (Rank) | VITS (Rank) |
|---|---|---|---|
| Male | 2.43 | 1.71 | 1.86 |
| Female | 2.0 | 1.86 | 2.14 |

Among the evaluated TTS models, Edge-TTS was initially identified as the best-performing model based on *objective* metrics (Accuracy, WER and absolute character difference). However, to assess human-likeness, we refer to the *subjective* ratings presented in Table. 3, where **SpeechT5** was consistently rated as the most human-sounding across all voice genders. Notably, Edge-TTS was ranked as the least human-like model despite its objective performance.

Table 4: Annotator Preferences for Gendered Voices by TTS Model

| Preferred Gender | Edge (%) | SpeechT5 (%) | VITS (%) |
|---|---|---|---|
| Female | 42.86 | 71.43 | 71.43 |
| Male | 57.14 | 28.57 | 28.57 |

Additionally, the female voices exhibited a narrower rating range compared to male voices, indicating more consistent preferences or perceptions among listeners. The gap in ratings between SpeechT5 and VITS was smaller for the male voice, suggesting a more stable ranking order for male speakers.

Table. 4 further reveals that annotators showed a clear preference for female voices in both the SpeechT5 and VITS models, whereas male voices were preferred for Edge-TTS. Moreover, the magnitude of gender preference bias was more pronounced for SpeechT5 and VITS than for Edge-TTS, indicating stronger listener preferences aligned with gender for these models.

Overall, the experimental results demonstrate that STT models are generally robust and perform well across various TTS-generated samples. However, both the STT models and human annotators demonstrate biases when the transcriptions originate from specific TTS voices.

#### 3.2.2 Assessing Gender Bias in STT Transcription Speed

Finally, we investigate whether there are any gender-based biases in the *processing time* required by STT models. To evaluate this, we use a subset of 100 audio samples and measure the average transcription time for each STT–TTS–gender combination. This setup mirrors our transcription method used earlier.

Formally, let $\mathcal{S}$ denote a given STT model, and $\mathcal{T}$ a TTS model. We define the processing times for female and male voices as follows:

1. **Female Voice Transcription Time:**

$$t_{\text{female}} = \text{time}(\mathcal{S}(\mathcal{T}_{\text{female}}(t)))$$

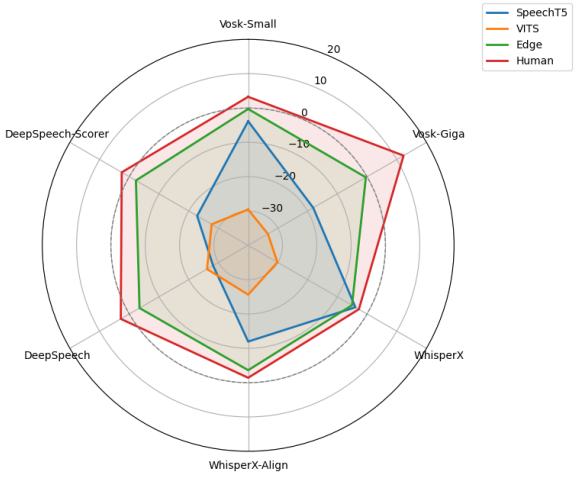


Figure 6: Percentage difference in average processing time between male and female voices across various TTS and STT model combinations. Negative values indicate faster processing for female voices, while positive values indicate faster processing for male voices. Results are based on 100 audio samples per configuration.

2. **Male Voice Transcription Time:**

$$t_{\text{male}} = \text{time}(\mathcal{S}(\mathcal{T}_{\text{male}}(t)))$$

3. **Relative Speed Difference:** We then compute the relative difference in processing speed between female and male samples as:

$$\text{SpeedDiff} = \frac{t_{\text{female}} - t_{\text{male}}}{t_{\text{male}}} \times 100,$$

where, SpeedDiff quantifies the percentage difference in processing time between female and male audio for a given STT–TTS pair, allowing us to identify potential gender-related latency biases in STT performance; and time$(\cdot)$ denotes the average time taken to transcribe 100 audio samples.

The results are presented in Fig. 6, and additional information is in the appendix E.2. As we can see, **all synthetic voice models exhibit a preference for female voices in terms of faster processing times**. In contrast, for human voice recordings, the preferred gender varies depending on the specific STT model used. Notably, the gender-based processing time differences are most pronounced for the VITS and SpeechT5 models, with disparities reaching up to 33.35%. In general, larger models tend to have longer processing times compared to their smaller counterparts. An exception is observed with DeepSpeech-Scorer, where enabling the language scorer unexpectedly resulted in faster processing.

Due to the lack of publicly available information regarding the gender distribution in the training data of these models, we can only hypothesise about the underlying causes of this observed bias. A possible reason could be that the model may have been overexposed to female voice data during training, resulting in fewer recognition paths for male voices, or has allowed the model to become attuned to specific vocal features more commonly present in female speech, and thus faster processing for female samples. Or differences in pre- and post-processing pipelines might favour specific acoustic characteristics that are more prevalent in female voices, making them easier for the model to handle.

### 3.3 Impact of STT and TTS on Hate Speech Classification

Modern audio-based classification pipelines often rely on transcripts generated by STT models. Therefore, it is critical to understand how different STT and TTS models influence the final classification results, particularly in tasks involving hate speech detection.

To evaluate this, we measure whether the classification of a sentence remains consistent before and after applying TTS and STT transformations. Let $c(t)$ be the classification outcome of the original sentence $t$, and let $t_j^{\text{transcribed}}$ be the transcript obtained by passing $t$ through a TTS model followed by an STT model. In experiments, we use Detoxify model (Hanu and Unitary team, 2020) as the classifier.

We define the classification preservation metric as follows:

1. **Original Classification:** Assign a class label to the baseline text input:

$$y_j^{\text{baseline}} = c(t_j)$$

2. **Transformed Classification:** Apply a TTS model $\mathcal{T}$ followed by an STT model $\mathcal{S}$, then classify the resulting transcription:

$$t_j^{\text{transcribed}} = \mathcal{S}(\mathcal{T}(t_j)), \quad y_j^{\text{transcribed}} = c(t_j^{\text{transcribed}})$$

3. **Classification Stability:** Compute the proportion of unchanged classification outcomes:

$$\text{Unchanged} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\left(y_j^{\text{transcribed}} = y_j^{\text{baseline}}\right),$$

where, $N$ is the total number of sentences evaluated, and $\mathbf{1}(\cdot)$ is the indicator function that returns 1 if the classification label remains unchanged (i.e., does not flip from hate to non-hate or vice versa), and 0 otherwise., and the metric Unchanged helps quantify the reliability of STT and TTS models in

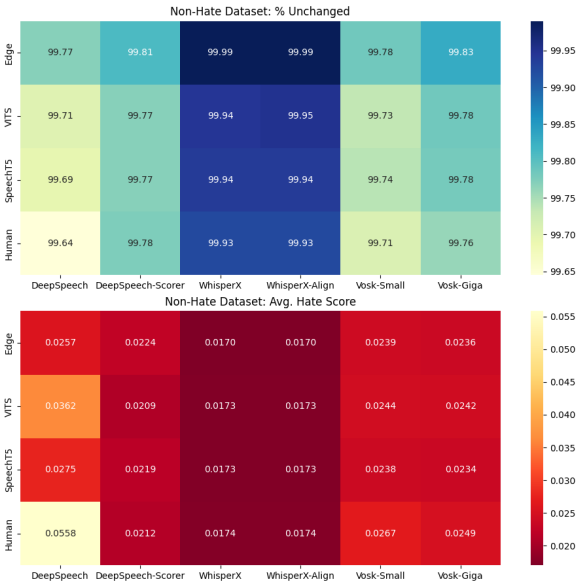


Figure 7: **Non-Hate** Dataset Detoxify Scores: Transcription accuracy and average hate speech score relative to the baseline

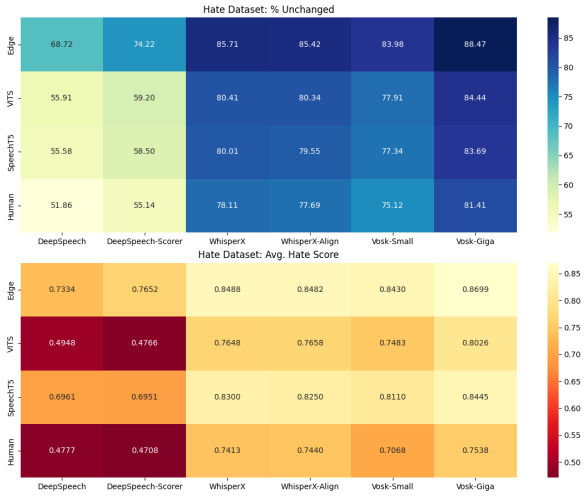


Figure 8: **Hate** Dataset Detoxify Scores: Transcription accuracy and average hate speech score relative to the baseline

preserving semantic integrity under classification tasks sensitive to lexical changes.

The results in Fig 7 & 8 show that, STT models generally perform worse with human speech than synthetic audio. The most significant degradation in classification accuracy (49.38%), was observed when combining DeepSpeech with a human hate speech sample. This suggests that transcription errors have a significant impact on downstream classification results.

The impact was less pronounced for non-hate samples, indicating that inaccuracies are less likely to elevate non-hateful content into the hate category falsely. This asymmetry highlights the greater vulnerability of hate speech samples to semantic distortion during transcription.

Among the TTS models evaluated, only Edge-TTS and SpeechT5 consistently held average toxicity scores exceeding the 0.8 threshold, reaffirming their ability to preserve critical lexical cues associated with hate speech. Overall, Edge-TTS emerged as the most reliable TTS model across both hate and non-hate datasets. At the same time, WhisperX-Align and Vosk-Giga were the top-performing STT models in preserving classification fidelity.

## 4 Conclusion

This study examined the robustness of STT and TTS models, focusing on inter-gender and intra-gender metrics, gender bias, and processing speed. The best-performing TTS model was Edge-TTS, consistently delivering top results across all categories, indicating strong compatibility with various STT models. SpeechT5 stood out as the most natural and human-like vocal output, making it a strong candidate for a human audio baseline. For STT models, WhisperX-Align was the top performer, exhibiting high accuracy and low deviations at both word and character levels. However, it showed a gender bias towards female voices, particularly with SpeechT5-generated audio. Despite this, its overall performance remains superior, while Vosk-Giga is recommended for scenarios requiring minimal gender bias. All STT models showed a tendency to favor synthetic voices over human audio.

The study also identified a general bias towards female voices, which affects evaluation metrics and processing speed. As it was found to be faster to process female spoken audio samples then male ones. This bias may stem from training data imbalances. Furthermore, STT model choice had a more significant impact on hate speech classification, with human audio samples causing more deviations than synthetic ones. It was also found that current STT models struggle to transcript hate audio correctly, requiring a fine tuned or better ASR model.

In conclusion, Edge-TTS is recommended as a high-performance TTS benchmark to estimate the theoretical upper bound of model performance. At the same time, SpeechT5 serves as an effective replacement for human baselines when human audio is unavailable. For transcription tasks, WhisperX-Align is recommended for its accuracy; however, researchers should also account for potential gender biases, particularly in hate speech classification.

## 5 Limitations

One main limitation of this study is the lack of gold-standard text transcriptions corresponding to the hate speech audio samples. This limitation resulted in a smaller and less comprehensive dataset of human-generated hate speech compared to the non-hate subset. Future work could address this by developing a larger, more representative dataset of hate speech audio with accurate ground-truth transcriptions.

Another challenge encountered was that some TTS models occasionally failed to synthesize specific audio files, requiring multiple attempts to generate a balanced dataset. Despite these efforts, inconsistencies and gaps remained, which were only a few audio samples. Additionally, some models exhibited "hallucination" behaviour during transcription, repeatedly generating the same word and thereby degrading transcription quality.

Lastly, platform and licensing restrictions limited our ability to use more advanced commercial models such as Google Voice or OpenAI's voice assistant, since generating audio with these tools would violate their terms of service. Consequently, future research in this area may be limited by the availability and ethical considerations surrounding cutting-edge speech synthesis technologies.

The use of synthetic voice would allow for the reduction of the requirement for audio hate speech. However, by doing so, the increase in fidelity or humanness of audio may enable actors to use it to create hate speech and attack people.

## 6 AI-Generated Content Acknowledgement

We acknowledge the use of AI in this paper. We used ChatGPT to assist in creating table titles and captions. It was also used to improve paragraphs, but a human then rewrote/modified all generated paragraphs. It was also used to find better ways to display the data, i.e, changing it from a table to a graph, creating Python code to display new data options, and for basic code structure and optimisation. The use of an AI-assisted spelling & grammar checker was also used in this paper.

## 7 Appendix

## A TTS and STT Model Information

- **VITS** (Kim et al., 2021) is a neural TTS model that unifies the training of the acoustic model and vocoder into a single framework. VITS integrates a variational autoencoder (Kingma et al., 2013), and normalising the flow. The addition of adversarial training, along with other methods, is used to create natural-sounding speech.

- **SpeechT5** (Ao et al., 2022) is a unified model for speech transcription, inspired by the work of the T5 (Text-To-Text Transfer Transformer) framework. It uses a shared encoder-decoder architecture, and enables a range of tasks, including speech recognition, text-to-speech, speech translation, voice conversion, speech enhancement, and speaker identification.

- **Edge TTS**[6] is a Python wrapper of the TTS service provided by Microsoft, which leverages techniques such as FastSpeech to produce audio and supports multiple languages.

- **WhisperX** (Bain et al., 2023) extends OpenAI's Whisper model (Radford et al., 2023) by enhancing timestamp alignment and improving transcription of longer audio recordings. These advancements are achieved through modifications to the Voice Activity Detection (VAD) component and the integration of forced phoneme alignment, resulting in more accurate word alignment. As a result, WhisperX provides more precise timestamps and higher-quality transcriptions. Making it a widely used transcription model in industry and research.

- **Vosk**[7] is an open-source STT toolkit that provides real-time transcription capabilities. It is built on top of the Kaldi automatic speech recognition framework (Povey et al., 2011). Vosk supports a range of model sizes to accommodate different performance and resource requirements.

- **DeepSpeech**[8] is an open-source implementation of the STT model proposed by Awni Hannun et al. (Hannun et al., 2014), developed by Mozilla. It is designed to function both with and without an external language scorer, which helps improve transcription accuracy by providing contextual guidance during the decoding process.

[6] https://github.com/rany2/Edge-TTS
[7] https://alphacephei.com/vosk/
[8] https://github.com/mozilla/DeepSpeech

## B Dataset

The common voice dataset is publicly available, at https://commonvoice.mozilla.org/en and is under Mozilla Public License 2.0.

## C Related Work

Automatic hate speech classification has been a longstanding research focus within the NLP community, with numerous studies exploring various classification frameworks and strategies. Notable contributions include the work of Tommaso Caselli et al. (Caselli et al., 2020) and Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang (Cao et al., 2020), which propose different models and architectures for detecting hate speech. These methods are particularly effective in identifying explicit hate speech in large, general-purpose datasets or within domain-specific contexts.

Beyond basic classification, recent research has also investigated methods to enhance the accuracy and robustness of hate speech detection. For example, Yadav et al. (Yadav et al., 2024) introduced an approach that enriches input data with post-level descriptions, enabling large language models to better detect implicit hate speech. In contrast, Lee and Ram (Lee and Ram, 2024) proposed augmenting classification inputs with physiological information rather than relying solely on textual macro- and micro-context. Both approaches represent innovative efforts to capture nuanced and indirect forms of harmful content that may otherwise evade detection by standard models.

Some prior research has focused on predicting user interactions related to hate speech. For example, Masud et al. (Masud et al., 2021) propose a case-based approach to predict the potential spread of hate speech and introduce a corresponding dataset to support this task. Additionally, Herodotou, Chatzakou, and Kourtellis (of Electrical and Electronics Engineers, 2021-4) developed a method for real-time aggression detection, specifically designed to efficiently classify large volumes of data simultaneously.

As the presentation of hate speech evolves, classification methods have also advanced to address new modalities. In particular, approaches that incorporate visual content have gained attention. For example, Lee R et al. (Lee et al., 2021) and Chen Y and Pan F (Chen and Pan, 2022) propose techniques that extract textual information alongside visual cues, such as identifying targeted individuals or extracting object tags, to improve hate speech detection in multimodal content.

Modern audio-based hate speech classification typically involves first transcribing audio recordings into text, which is then analysed using established text-based hate speech classifiers. Wu and Bhandary (Wu and Bhandary, 2020), uses the Google Cloud Speech-to-Text API to transcribe given YouTube videos, and then classify these being normal, racist, or sexist. This was done the help of blob for sentiment analysis. The work by Imbwaga, Chittaragi, and Koolagudi (Imbwaga et al., 2024), also transcribes given audio into text. But utilise whisper as there STT model. There paper looked more into why classify sentences as hate speech. And used multiple datasets and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) model to understand the models' predictions. Allowing for insights into models' classification decisions. Other possible implimentations of audio classification are audio to classification models. Examples such as the paper by An J et al. (An et al., 2024), where they use TTS to create their dataset. They showed two pipelines, one being similar to others, where audio in transcribed using Whisper and then classified with Bert, but the other pipeline, utilise wav2vec (Schneider et al., 2019) to classify audio, into hate or non hate.

## D Evaluation Metrics

The evaluation metrics cover multiple levels of granularity, including *sentence*-level, *word*-level, and *character*-level differences, as introduced in the following subsections.

### D.0.1 Transcription Accuracy (Radford et al., 2023)

It is computed at the sentence level to evaluate how often a transcription matches the reference text exactly. It is defined as the ratio of correct transcriptions to the total number of transcriptions:

$$\text{Accuracy} = \frac{\#\text{Correct Transcriptions}}{\#\text{Total Transcriptions}}, \quad (1)$$

where, #Correct Transcriptions refers to the number of transcriptions that perfectly match the reference sentence, and #Total Transcriptions denotes the total number of sentences processed by the model.

### D.0.2 Word Error Rate (WER) (Radford et al., 2023)

WER is a standard metric for evaluating the performance of transcription models. It quantifies the

difference between the predicted transcription and a reference transcript by measuring the minimum number of word-level edits (i.e., substitutions, deletions, and insertions) required to convert the prediction into the reference. Formally, WER is defined as:

$$\text{WER} = \frac{\#\text{substitutions} + \#\text{deletions} + \#\text{insertions}}{\#\text{words in reference}}. \quad (2)$$

A lower WER indicates a more accurate transcription, making it a widely adopted benchmark for comparing STT systems.

#### D.0.3 Absolute Character Difference

We introduce this metric to quantify the character-level deviation between the STT-generated transcript ($t'$transcript) and the reference baseline ($t$baseline). It is defined as:

$$\Delta^{\text{abs}}_{\text{char}} = \left|\text{len}(t'_{\text{transcript}}) - \text{len}(t_{\text{baseline}})\right|, \quad (3)$$

where $\text{len}(\cdot)$ denotes the number of characters in a given text. This metric captures the absolute difference in length between the transcription and the baseline, providing a simple yet effective measure of how much the output diverges in terms of textual content.

#### D.0.4 Inter-Gender Consensus

To assess consistency across genders, we introduce the Inter-Gender Consensus metric. This measures the proportion of cases in which the transcriptions generated for male and female speakers are identical for the same sentence. Formally, it is defined as:

$$C_{\text{inter}} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}(\text{transcript}^{j}_{\text{female}} = \text{transcript}^{j}_{\text{male}}), \quad (4)$$

where $N$ is the total number of sentence instances evaluated, $\mathbf{1}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise, and $\text{transcript}^{j}_{\text{female}}$ and $\text{transcript}^{j}_{\text{male}}$ are the transcriptions for the $j$th sentence by female and male speakers, respectively. This metric captures the level of agreement between male and female transcriptions for the same content, providing insight into potential gender-related variations in how models process speech. A lower consensus score may suggest that the model is more sensitive to gender-specific acoustic features or pronunciation differences.

### D.1 Human Like Metrics

To evaluate the perceptual results from human annotators, we conducted two analyses. First, we computed the average human-likeness ranking for each STT model using the following formula:

$$\text{AverageRanking}_{STT_i} = \frac{1}{N}\sum_{j=1}^{N} r_{ij}, \quad (5)$$

where $\text{AverageRanking}_{STT_i}$ is the average ranking of the $i^{\text{th}}$ STT model, $r_{ij}$ is the ranking score given by annotator $j$ to STT model $i$ and $N$ is the total number of annotators.

Next, to assess gender preference for each STT model, we calculated the proportion of times each gendered voice (male or female) was rated as more understandable. This is formalized as:

$$\text{PreferedGender}^{STT}_{g} = \frac{C_g}{\sum_{g' \in G} C_{g'}} * 100, \quad (6)$$

where $g \in G = \{\text{male}, \text{female}, \text{none}\}$ are the options for each STT model, $C_g$ is the total number of times each option $g$ was selected, and $\text{PreferedGender}_g$ is the proportion of preferences for category $g$.

## E Extra Results

### E.1 Global & Inter Gender Results

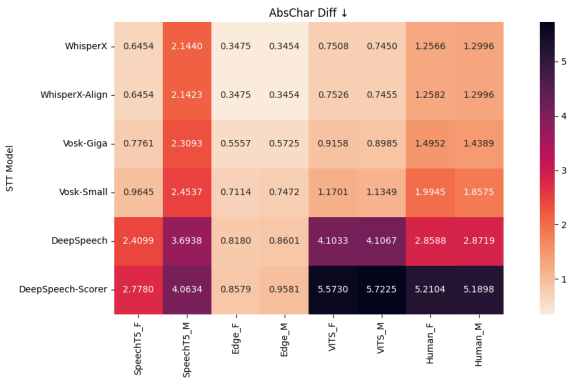


Figure 9: STT robustness on **non-hate** speech: Comparison of Absolute Character Difference across TTS models and human audio (by gender).

### E.2 Processing Speed Results

## F Annotator Prompt

Here is the following instruction given to the annotators

Please listen to the following audio files, for each gender and TTS model. And rank them from 1-3 where 1 is the most human like and 3 the least

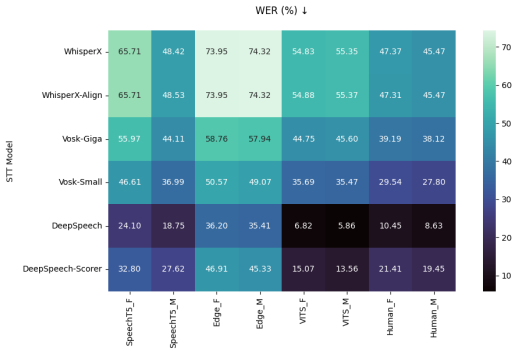


Figure 10: STT robustness on **non-hate** speech: Comparison of transcription and WER across TTS models and human audio (by gender).

Table 5: Average Time per Audio Sample (in Seconds) for different STT models for each given TTS model and voice option, running on 100 audio samples

| TTS Model (F/M) | Avg pre sample (sec) (F/M) | Difference (%) |
|---|---|---|
| **STT Model: Vosk-Small** | | |
| SpeechT5 | 0.8918 / 0.9276 | -3.86 |
| VITS | 0.6330 / 0.8987 | -29.56 |
| Edge | 0.8968 / 0.8993 | -0.28 |
| Human | 0.9714 / 0.9406 | 3.27 |
| **STT Model: Vosk-Giga** | | |
| SpeechT5 | 1.7589 / 2.1478 | -18.11 |
| VITS | 1.3254 / 1.9887 | -33.35 |
| Edge | 1.9310 / 1.9386 | -0.39 |
| Human | 2.7333 / 2.4341 | 12.29 |
| **STT Model: WhisperX** | | |
| SpeechT5 | 0.3660 / 0.3809 | -3.91 |
| VITS | 0.2497 / 0.3575 | -30.15 |
| Edge | 0.3453 / 0.3637 | -5.06 |
| Human | 0.3637 / 0.3740 | -2.75 |
| **STT Model: WhisperX-Align** | | |
| SpeechT5 | 1.0372 / 1.1782 | -11.97 |
| VITS | 0.7697 / 1.0350 | -25.63 |
| Edge | 1.0310 / 1.0698 | -3.63 |
| Human | 1.0735 / 1.0890 | -1.42 |
| **STT Model: DeepSpeech** | | |
| SpeechT5 | 1.9496 / 2.7153 | -28.2 |
| VITS | 1.5764 / 2.1352 | -26.17 |
| Edge | 2.2974 / 2.3796 | -3.45 |
| Human | 2.7654 / 2.6886 | 2.86 |
| **STT Model: DeepSpeech-Scorer** | | |
| SpeechT5 | 1.6336 / 2.1171 | -22.84 |
| VITS | 1.2177 / 1.6847 | -27.72 |
| Edge | 1.8056 / 1.8468 | -2.23 |
| Human | 2.2140 / 2.1591 | 2.54 |

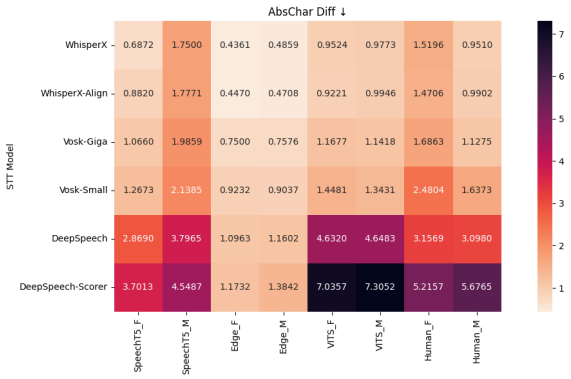


Figure 11: STT robustness on **hate** speech: Comparison of transcription Absolute Character Difference across TTS models and human audio (by gender).

human like. After please label which gender was easier to listen for each TTS model.

Also agree that we use this data for our research.

## References

Jinmyeong An, Wonjun Lee, Yejin Jeon, Jungseul Ok, Yunsu Kim, and Gary Geunbae Lee. 2024. An investigation into explainable audio hate speech detection. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*,

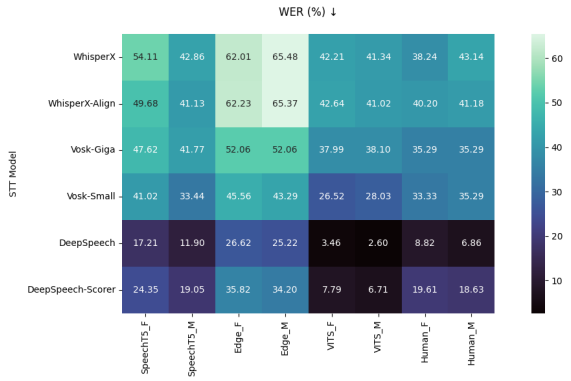


Figure 12: STT robustness on **hate** speech: Comparison of WER across TTS models and human audio (by gender).

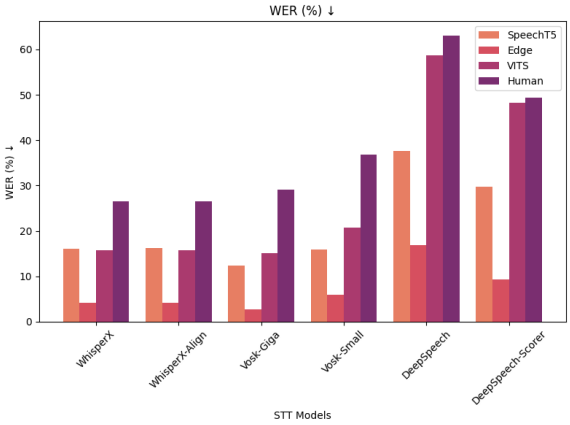


Figure 13: STT robustness on **non-hate** speech: Comparison of transcription WER across TTS models and human audio (by gender).

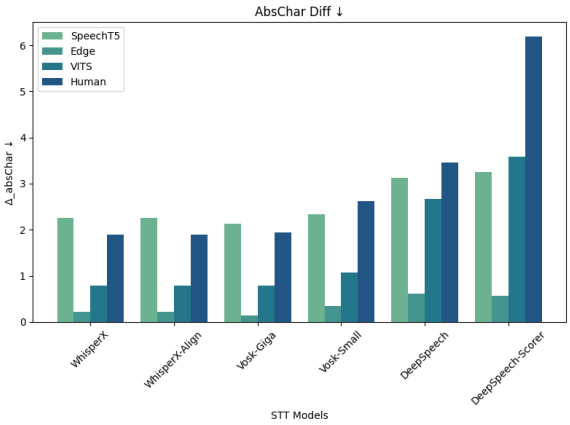


Figure 14: STT robustness on **non-hate** speech: Comparison of transcription Absolute Character Difference across TTS models and human audio (by gender).

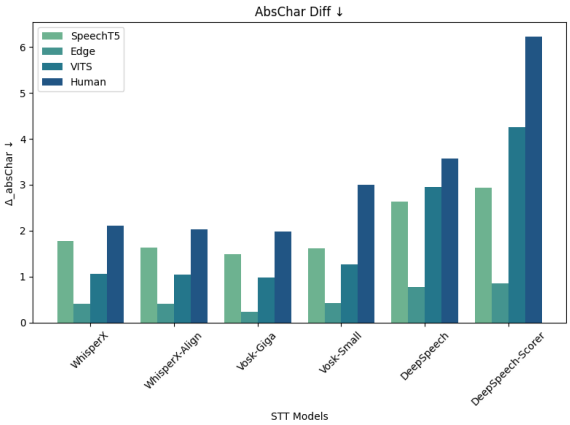


Figure 15: STT robustness on **hate** speech: Comparison of transcription Absolute Character Difference across TTS models and human audio (by gender).

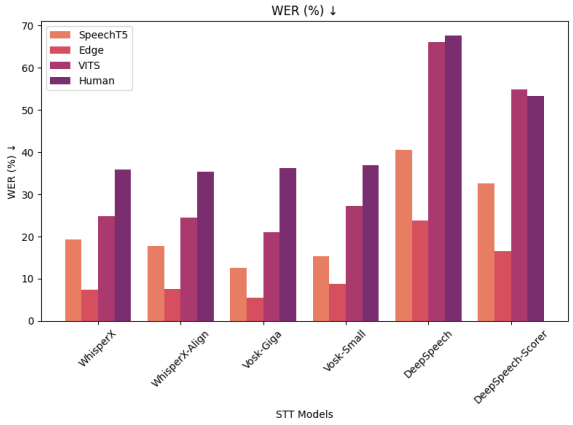


Figure 16: STT robustness on **hate** speech: Comparison of transcription WER across TTS models and human audio (by gender).

pages 533–543, Kyoto, Japan. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *Preprint*, arXiv:2110.07205.

Mandal Atanu, Roy Gargi, Barman Amit, Dutta Indranil, and Naskar Sudip. 2023. Attentive fusion: A transformer-based approach to multimodal hate speech detection. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 720–728.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *Preprint*, arXiv:2303.00747.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, pages 11–20.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *Plos one*, 17(9):e0274300.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *Preprint*, arXiv:1412.5567.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Joan L Imbwaga, Nagaratna B Chittaragi, and Shashidhar G Koolagudi. 2024. Explainable hate speech detection using lime. *International Journal of Speech Technology*, 27(3):793–815.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *CoRR*, abs/2106.06103.

Diederik P Kingma, Max Welling, and 1 others. 2013. Auto-encoding variational bayes.

Kyuhan Lee and Sudha Ram. 2024. Deep learning for hate speech detection: A personality-based approach. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1667–1671.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147.

Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. 2021. Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter. In *ICDE*, pages 504–515.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023a. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *ACL 2023-61st Annual Meeting of the Association for Computational Linguistics*, pages 2758–2772. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023b. An in-depth analysis of implicit and subtle hate speech messages. In *EACL 2023-17th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2023, pages 1997–2013. Association for Computational Linguistics.

Institute of Electrical and issuing body. Electronics Engineers, author. 2021-4. Catching them red-handed: Real-time aggression detection on social media. In *2021 IEEE 37th International Conference on Data Engineering (ICDE) /*, 2021 IEEE 37th International Conference on Data Engineering (ICDE), Piscataway, New Jersey :. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, and 1 others. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Jing Qian, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. *arXiv preprint arXiv:1804.03124*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *2020 international conference on computational science and computational intelligence (CSCI)*, pages 585–590. IEEE.

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-bart: Leveraging toxicity attributes for explanation generation of implicit hate speech. *arXiv preprint arXiv:2406.03953*.