Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Fair or Fare? Understanding Automated Transcription Error Bias in Social Media and Videoconferencing Platforms

Daniel J Dubois[1], **Nicole Holliday**[2], **Kaveh Waddell**[3,4], **David Choffnes**[1]
[1] Northeastern University, Khoury College of Computer Sciences
[2] Pomona College, Department of Linguistics and Cognitive Science
[3] Consumer Reports, Innovation Lab
[4] Stanford University, Institute for Human-Centered AI
d.dubois@northeastern.edu, nicole.holliday@pomona.edu, kavehw@stanford.edu, choffnes@ccs.neu.edu

Abstract

As remote work and learning increases in popularity, individuals, especially those with hearing impairments or who speak English as a second language, may depend on automated transcriptions to participate in business, school, entertainment, or basic communication. In this work, we investigate the automated transcription accuracy of seven popular social media and videoconferencing platforms with respect to some personal characteristics of their users, including gender, age, race, first language, speech rate, F_0 frequency, and speech readability. We performed this investigation on a new corpus of 194 hours of English monologues by 846 TED talk speakers. Our results show the presence of significant bias, with transcripts less accurate for speakers that are male or non-native English speakers. We also observe differences in accuracy among platforms for different types of speakers. These results indicate that, while platforms have improved their automatic captioning, much work remains to make captions accessible for a wider variety of speakers and listeners.

1 Introduction

Social media and videoconferencing platforms are increasingly becoming an important part of everyday life, and so is their functionality and need for accessibility. One of the recent accessibility features these platforms offer is the ability to automatically transcribe spoken material, which allows users to employ them for remote work, remote learning, social activities, or simply basic communication. As the adoption of these platforms increases and automatic transcriptions become a necessity, especially for those who are hard of hearing or who are working in their second language, it is important that the transcripts they generate are accurate, no matter the speaker's gender, race, or other characteristics.

Previous work has evaluated the automated transcription capability of common transcription libraries and voice assistants, *e.g.,* Koenecke et al. (2020); Lima et al. (2019); Feng et al. (2021); Meyer et al. (2020). These studies all discovered the presence of systematic *bias*, *i.e.,* a different level of accuracy, based on demographic characteristics of the speakers. Some limitations of these previous works are that they focus on platforms that are less popular than social media and videoconferencing platforms, or they use corpora with limited data or limited labels. Moreover, popular social media and videoconferencing platforms do not disclose what transcription models they use, making it challenging to perform large-scale analysis on them.

In this work, we are the first to study transcription biases in popular social media and videoconferencing platforms. In particular, we compare *transcription error* for seven platforms with respect to seven characteristics of their users: gender, age, race, first language, speech rate, F_0 (a primary correlate of what we perceive as pitch), and speech readability. Specifically, we address the following questions:

- Are certain groups subject to more transcription errors than others when using social media and videoconferencing platforms?
- What are the factors that correlate with these biases and how can companies mitigate them?

To answer these questions, we address two key challenges that limit prior work in this space: (*i*) the *lack of a suitable corpus*, *i.e.,* a corpus composed of multiple large natural speech samples with correct manually-curated transcripts from speakers having a diverse set of personal characteristics; (*ii*) the *lack of a scalable experimental methodology*, *i.e.,* a way for producing transcripts for a large number of speech samples that is tailored for social media and videoconferencing platforms, and that reduces the manual effort.

Specifically, we first curate a *new corpus*, named MonoTED, inspired by the existing TED-LIUM (Hernandez et al. 2018) (a simple collection of TED talks with transcripts that are sometimes machine-generated), but with new, accurate, manually-generated transcripts and comprehensive speaker labels (§3.1). This new corpus comprises 194 hours and 1.8M words of English monologues spoken by 846 TED talk speakers. Then, in §3.2 we describe our approach for performing transcription experiments in bulk across three social media platforms: YouTube, Facebook Video, Microsoft Stream; and four videoconferencing platforms: BlueJeans, Zoom, Google Meet, Webex. Finally, we measure the transcription error using the Word Error Rate metric (Woodard and Nelson 1982) (§3.3), and perform a regression analysis across all data dimensions to examine possible bias against types of speakers (§3.4).

The results of our analysis (§4) show the presence of significant bias, with higher word error rates for speakers

who are male and/or non-native English speakers, which confirms previous work in other automated speech recognition contexts (Koenecke et al. 2020). Surprisingly, we did not find evidence of racial bias, which contrasts with such prior work. We also discovered that social media platforms transcribe better than videoconferencing platforms, and that speakers that speak slower, faster, or with more common words and shorter sentences than the average have more transcription errors. We conclude the paper by discussing possible reasons that explain the biases we measured, possible mitigation strategies, and the feedback we obtained from the companies running the platforms after disclosing our findings to them (§5).

To promote reproducibility and support future research, we released at https://github.com/NEU-SNS/MonoTED:

1. MonoTED, our **new corpus**, with the new labels and TED transcripts (1.8M words, 194 hours, 846 speakers).
2. The 5922 platform-generated transcripts.
3. Software to compute and analyze the word error rate.

2 Context and Goals

Context and scope. We analyze the *transcription error bias* among popular social media and videoconferencing platforms offering automatic transcriptions, where transcription error bias is defined as any significant change in transcription errors when varying speaker characteristics.

Social media platforms are popular platforms that allow users to upload videos, and that automatically create downloadable transcripts of such videos. In this study we analyze YouTube[1], Facebook Video[2], and Microsoft Stream[3].

Videoconferencing platforms are popular platforms that allow users to create video meetings with other users and that automatically create downloadable transcripts of such meetings. In this study we analyze Zoom[4], BlueJeans[5], Webex[6], and Google Meet[7]. Note that the platforms we chose are not the only popular platforms, but other popular ones we considered did not offer the transcript download feature when we started our experiments (Mar. 2022). However, our methods generalize to other platforms that become available.

Goals. Our goal is to answer the following questions:

Q1: *What is the impact of speaker characteristics and the platform used on transcription errors?* This question entails analyzing the transcripts generated from a diverse set of speakers and platforms and determining if the transcription error bias correlates to any of them more than others. The motivation for this question is to determine if certain groups are more at risk of being discriminated with respect to the platform's ability to transcribe correctly.

[1]https://youtube.com (Accessed: 2024-04-11).

[2]https://facebook.com (Accessed: 2024-04-11).

[3]https://www.microsoft.com/en-us/microsoft-365/microsoft-stream (Accessed: 2024-04-11).

[4]https://zoom.us (Accessed: 2024-04-11).

[5]https://bluejeans.com (Accessed: 2024-04-11).

[6]https://webex.com (Accessed: 2024-04-11).

[7]https://meet.google.com (Accessed: 2024-04-11).

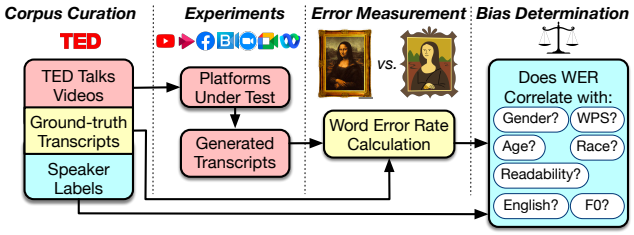


Figure 1: Methodology Overview.

Q2: *What are the factors that correlate with transcription error bias and how can we mitigate them?* First, we investigate whether speaker and speech characteristics correlations and other platform characteristics (*e.g.,* time spent to process the transcript) can explain the transcription error bias we observe when answering Q1. Second, we investigate how we can use what we have learned about the bias to mitigate it.

3 Methodology

We use four methods to determine transcription bias in videoconferencing and social media platforms (Fig. 1).

3.1 Corpus Curation

The purpose of corpus curation is to create a collection of spoken material organized into different units sharing similar characteristics. We refer to this collection of material as MonoTED, *i.e.,* the *corpus* of this study. To answer our research questions, we created a corpus that met the following requirements. First, each unit of spoken material must be an English monologue with a *ground-truth transcript* associated with it (*i.e.,* a transcript not automatically generated) that we take to be accurate. This allows us to avoid errors in disambiguating multiple speakers and gives us a baseline for measuring errors in transcripts generated by video-conferencing and social media platforms. The second requirement is the presence of labels that associate each corpus unit to the characteristics of the speaker (*e.g.,* gender). This allows us to compare the transcription error (and therefore measure bias) among types of speakers with respect to their labels (*e.g.,* for the *gender* label, to compare the error between female and male speakers).

Corpus Selection. To obtain a large set of English monologues with accurate transcripts, we considered several existing corpora typically used to train and evaluate speech recognition systems such as Mozilla Common Voice (Ardila et al. 2020), CORAAL (Kendall and Farrington 2018), among others (Nagrani, Chung, and Zisserman 2017; Panayotov et al. 2015; Meyer et al. 2020). However, these corpora do not satisfy our requirements, as discussed in §6.

To address this issue, we created a new corpus, MonoTED, inspired by TED-LIUM (Hernandez et al. 2018), *i.e.,* a collection of transcribed spoken material in which each unit is a TED talk[8]. A TED talk is an English talk of 18 minutes or less in which a speaker, typically a renowned

[8]https://www.ted.com/ (Accessed: 2024-04-11).

Characteristic	Value	# of speakers
Perceived gender	Male	548 (65%)
	Female	298 (35%)
English language	First (US)	478 (56%)
	First (non-US)	216 (26%)
	Second	152 (18%)
Perceived race	White	651 (77%)
	Black	77 (9%)
	Asian	104 (12%)
	Latino	14 (2%)

Table 1: Distribution of non-numeric labels in our corpus.

expert in their field, presents an idea. Although TED-LIUM and our corpus share similar spoken material, the transcripts we use in our corpus are different. After a manual analysis, we discovered TED-LIUM transcripts are often incorrect, with the frequent presence of words not properly transcribed (*e.g.,* frequent appearance of `<unk>` markers, meaning that a word was not transcribed). Another problem is that TED-LIUM transcripts often lack the proper markers to distinguish monologues from dialogues, and for recognizing singing and external videos. To overcome this problem we only use transcripts extracted from the TED website using Selenium[9], since they are not affected by such limitations. Note that TED talks and their transcripts are owned by TED and released under the Creative Common CC BY–NC–ND 4.0 International license.

To complete the selection of our corpus, we exclude TED talks that include speech that is not English monologue using the information from the new transcripts. First, we exclude *dialogues*, *i.e.,* talks whose transcripts include sentences starting with the name or initials of a speaker followed by a colon, since this does not occur for monologues. Then, we exclude talks having markers showing the presence of non-monologue audio: *e.g.,* "♪" and "`(music)`" representing singing or playing music, "`(video)`" representing a video playing, *etc.* Finally, to avoid having overrepresented speakers in our corpus, **we only consider one talk per speaker**, chosen randomly.

After our selection process, our final corpus has 846 talks/speakers, 194 hours of monologues, and 1.8M words.

Labeling Speaker Characteristics. To complete our curation step, we add, for each talk/speaker, labels representing individual characteristics. We select labels that are commonly used in previous work (see §6): four common demographic characteristics (gender, age, English language status, and race) and three non-demographic characteristics (speech rate, F_0, and speech readability).

Both TED-LIUM and the TED website provide information about the name of the speaker and the year of the talk, but they do not provide any demographic information. Therefore, we need a method to reliably derive it.

Labeling Gender and Age. We address this labeling problem by leveraging *automated web crawling*. First, we perform an automated Google search on each speaker to find their Wikipedia pages (if they exist). Once their page is found, we programmatically analyze the page for the presence of gender and birth year information. If no page is found, or if the result is ambiguous, we make the final determination manually. Note that our definition of gender represents the **perceived gender** (from the Wikipedia page authors or from this paper authors), and not necessarily the self-identified gender of the speaker. As a result, we refer to this variable as "Perceived Gender" in our analysis. Specifically, we label perceived gender based on the pronouns used in their Wikipedia page to refer to the speaker: if the pronouns are "she/her" the speaker is labeled as *female*, if the pronouns are "he/him/his," the speaker is labeled as *male*, and if "they/them/their" are used, the speaker is labeled as *gender-neutral*. Cases of missing or mixed pronouns (*e.g.,* due to missing or ambiguous Wikipedia page) are decided by the authors based on research about speakers as well as perception. By chance, we did not find any of the speakers to be identified or perceived as gender-neutral.

We also identify the year of birth using a similar crawling approach, but we assume the Wikipedia date of birth to be correct and not just perceived. Cases of missing or impossible (*e.g.,* a year in the future) birth years based on crawls are instead manually labeled to ensure correctness. Note that we were unable to obtain the year of birth for every speaker, but for those available, we could estimate their actual age as the difference between their talk year and their birth year.

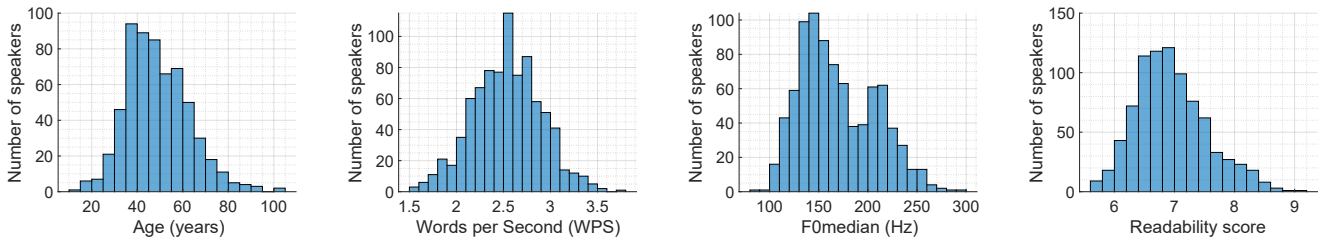
Labeling English Language Status and Race. Regarding English language status and race labels, we were unable to find a programmatic way to obtain this information; hence, we relied on a manual approach. Speakers whose biographical information was available on Wikipedia were classified by their place of residence and/or birth. However, the English language status of speakers for whom this information was not available were impressionistically coded by the team. To label English language status, we consider: (*i*) *native US English speaker*, (*ii*) *native non-US English speaker* (*e.g.,* British English), (*iii*) *non-native English speaker*. Since our preliminary analysis did not show differences between speakers of U.S. and other varieties of English, those categories are collapsed in the results.

To label race, we consider the following categories: (*i*) *White*, (*ii*) *Black*, (*iii*) *Asian*, (*iv*) *Latino*, following a simplified modification of the U.S. Census categories[10]. Similarly to the gender and English language status classifications, we rely on a combination of manual research and impressionistic coding by the research team, based on observations from their TED talk. As such, Race here is defined as a combination of available biographical information as well as impressionistic judgments by the research team. Similar to gender-labeling, we recognize that these may not reflect each speaker's self-identified race, so this characteristic is also referred to as "Perceived Race" in our analysis.

Coding perceived labels. When relying on impressionistic coding, we relied on three people on the research team: one author, and two non-authors, all experienced researchers and trained linguists, making decisions independently. Each label was annotated by two people, with a 95.6% rate of

[9]https://www.selenium.dev/ (Accessed: 2024-04-11).

[10]https://www.census.gov/2020census (Accessed: 2024-04-11).



(a) Age (b) Speech Rate (c) Median F_0 (d) Speech Readability

Figure 2: Histograms showing the distribution of numeric speaker/speech characteristics in our MonoTED corpus.

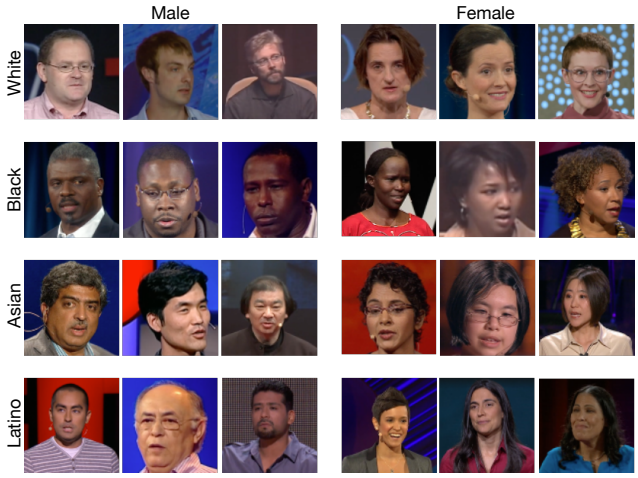


Figure 3: Examples of perceived gender and race.

agreement. The few cases of disagreement were handled using the third person as tiebreaker. While we acknowledge the limitations of impressionistic coding of such traits, previous research has demonstrated the value of such methods for datasets where aspects of speaker background may not be readily available (Reid 2010; Sims, Pirtle, and Johnson-Arnold 2020). For perceived gender, coders were instructed to evaluate speaker name and perceived gender-specific traits (*e.g.,* facial hair, tone of voice, *etc.*). For perceived race, we used a holistic approach based on speaker name, country of origin, skin/hair/eye color, *etc.* While we acknowledge that perceived race and actual racial identification of the speakers are distinct criteria, when humans engage in linguistic discrimination, they often do so without empirical knowledge; instead relying on the same types of perceptual characteristics that our coders employed here.

The final distribution of speaker characteristics in our corpus is reported in Table 1 and Fig. 2a. In Fig. 3 we show some examples of perceived gender and race.

Labeling Speech Characteristics. We assign to our corpus also the following non-demographic labels, to understand if any sources of transcription error bias can be explained by speech characteristics. The distribution of speech characteristics in our corpus is reported in Fig. 2b-d.

Speech rate. This label measures the speaking pace of the speakers as words per second (WPS). We calculate this by dividing the total number of words in the ground-truth transcript by the talk duration.

F_0. This label measures the fundamental frequency (Gerhard et al. 2003) for each talk, a primary correlate of what we perceive as pitch, estimated with the MATLAB `pitch` function (MathWorks, Inc. 2018). Since F_0 is a function with respect to time, we consider its mean and median value across the whole talk duration.

Speech readability score. This label measures the readability of the transcript of a talk estimated using the New Dale–Chall readability formula (Sternlicht and Edgar 1995). Lower (higher) readability values mean that the text is understandable by lower (higher) grade students. We chose the New Dale-Chall score with respect to other scores because it also considers the impact of lexical frequency (*i.e.,* the effect of common *vs.* non-common words).

3.2 Transcription Experiments

Once we have labeled our spoken material and obtained its correct transcripts, we process each talk on each platform and then download the auto-generated transcripts. For videoconferencing platforms, we play our spoken material during ad-hoc videoconferences, while for social media platforms we simply upload our spoken material directly. To avoid having to manually repeat this process for each talk, we combine the talks into groups, separated by an audio separator, so that we can process multiple talks in a single experiment and thus minimize the manual effort. More details on how we conducted experiments are discussed as follows.

Videoconferencing Platforms. The videoconferencing platforms we consider are: Zoom (ZM), BlueJeans (BJ), Webex (WX), and Google Meet (GM). We run the experiments for videoconferencing platforms on a Windows 11 computer. To play our corpus we use VLC media player[11]. To redirect the audio output of VLC to the audio input of videoconferencing applications without quality loss, we use VB Virtual Audio Cable [12], which creates and connects a virtual output device (seen as a speaker by VLC) to a virtual input device (seen as a microphone by videoconferencing applications). Regarding the Zoom, BlueJeans, and Webex platforms, before starting the experiments, we install the latest desktop version of their Windows app and start a pre-

[11]https://www.videolan.org/vlc/ (Accessed: 2024-04-11).

[12]https://vac.muzychenko.net/ (Accessed: 2024-04-11).

mium subscription that includes real-time automatic closed captioning. Regarding the Google Meet platform, since it only has a Web app, we use the latest version of the Chrome browser, with a Meet Transcript extension[13] installed, which is needed to download the transcripts after a meeting has ended. Once all the software is ready, we iteratively perform the following steps for each experiment: (1) we open the videoconferencing app, enable recording and closed captioning, and have a dummy user join the meeting (to prevent the meeting from being automatically ended after a certain time), (2) we use VLC to play a group of TED talks we want to transcribe, (3) we end the meeting and download the transcripts once they become available.

Social Media Platforms. The social media platforms we consider are: YouTube, Facebook Video, and Microsoft Stream. Since all the social media platforms allow their users to upload videos that will be transcribed automatically, we use a simpler approach. First, we upload the video containing the group of talks we want to transcribe, which is then transcribed by the platform automatically, without any other actions from us. Then, after waiting for the video to be processed, we download the transcript.

Grouping of Talks and Transcripts. Since our experiment methodology relies on several manual steps, we do not want to run the procedures above for each talk, as it would need to be manually repeated 846 times for each platform, making our methodology difficult to be used in practice. To streamline automated testing, we grouped multiple talks together by concatenating them and adding an audio separator between a talk and the next. This separator avoids interference between subsequent talks, and acts as a marker to separate transcripts for different talks. Since all the videoconferencing platforms we consider allow meetings up to 24 hours long, we were able to transcribe ≈96 talks per experiment, *i.e.,* each group of talks included up to 96 talks. Regarding social media platforms, they allowed us to transcribe videos up to 4 hours long, therefore we were able to use groups of talks comprising up to 16 talks each.

Dealing with Platform Issues. During our experiments we noticed that sometimes the platforms produce errors or incomplete transcripts with no apparent reasons. In such cases we simply repeat the transcription experiments until we get a complete transcript and no errors. We detect incomplete transcripts by performing an outlier analysis among platforms with respect to the length of the transcript and the word error rate (see §3.3). We then manually check these outliers to confirm if they are complete and can be kept as they are, or if they are incomplete and the experiments associated with them need to be repeated.

3.3 Transcription Error Measurement

To measure the *transcription error* between the generated transcripts and the correct ones, we use the *Word Error Rate* (WER), a widely used (Koenecke et al. 2020) error metric that compares a reference text (a ground-truth transcript) to a hypothesis text (an automatically generated transcript). WER measures the proportion of words in the reference text that have been mistranscribed in the hypothesis text and is formally defined as: $WER = \frac{S+D+I}{N}$, where S, D, and I are the number of substitutions, deletions, and insertions needed to transform the reference text into the hypothesis text, while N is the number of words in the reference text.

We measure the WER in two steps, a *preprocessing* step and a *WER calculation step*. In the preprocessing step we alter the reference and the hypothesis text to reduce the occurrence of false mistranscriptions due to the presence of equivalent, but different, words. Specifically, we (1) transform ordinal and cardinal integer numbers in textual form to their numeric forms (*e.g.,* "third" becomes "3rd" and "eleven" becomes "11"); (2) we transform the textual form of "percent" and the most popular currencies into their symbolic versions (*e.g.,* "5 percent" becomes "5%" and "2 dollars" becomes "$2"); (3) we make all the text lowercase (to ignore capitalization); (4) we replace punctuation such as commas, periods, dashes and apostrophes with spaces; (5) we remove text surrounded by brackets (*e.g.,* "(Applause)") since it is present in the reference text found in the TED website, but never actually spoken by a speaker. After the preprocessing step, we then calculate the WER directly using *JiWER* (Vaessen 2018).

3.4 Bias Determination

In this last step, we measure how the WER changes with each platform and each speaker characteristic, *i.e.,* the *bias* of the speaker characteristics with respect to the WER. We define as bias any statistically significant difference of WER with respect to a given speaker characteristic (Koenecke et al. 2020). For example, if a platform has a WER that is consistently lower for females with respect to males, such platform is exhibiting gender bias.

Since we have multiple person characteristics that are not independent from each other (*e.g.,* the WER of a speaker that is, for example, both Asian and female has effects on both the race bias and the gender bias), we need a statistical method that takes these correlations into account when determining the bias. To test for such interactions, we ran a Linear Mixed Effects Regression (LMER) model and follow-up omnibus testing with ANOVAs using the lme4 package (Bates et al. 2015). Initially our model included WER by platform, with interaction effects of Perceived Gender, Language, Perceived Race, Age, Words Per Second (WPS), median F_0, and readability score. The model also included a random effect of Speaker. Preliminary testing indicated that age did not contribute to the model, and it was also not significant in any ANOVAs that we ran, so age was ultimately dropped from the final model. Results of the model for WER by platform indicate significant differences in performance between platforms, as well as a number of interactions between platform and speaker characteristics. We then employ this model, as well as follow-up testing with ANOVAs, to predict how the WER varies with respect to any of the speaker characteristics, using a bootstrapped 95% confidence interval to ensure statistical significance.

[13] https://thanesh.dev/meet-transcript (Accessed: 2024-04-11).

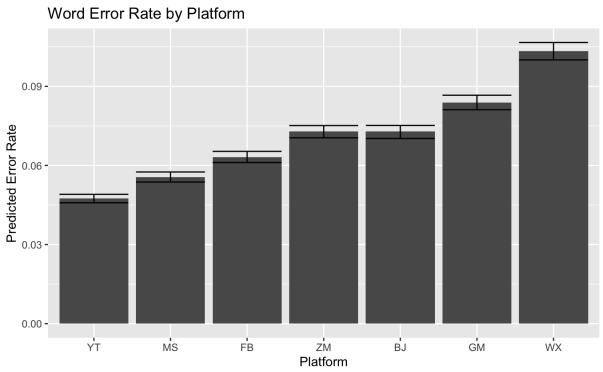


Figure 4: WER by Platform.

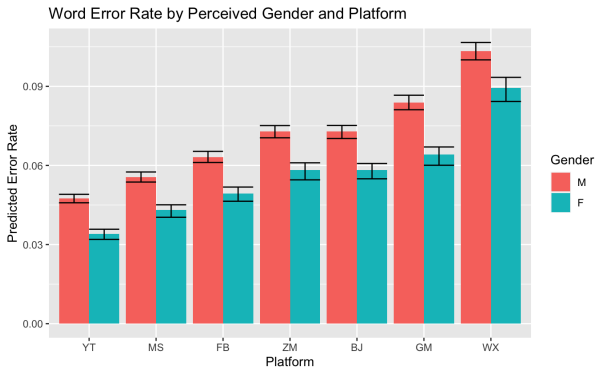


Figure 5: WER by Speaker Gender and Platform.

4 Results

In this section we compare WER among platforms, speaker characteristics, and speech characteristics.

4.1 Platform Comparison

Understanding baseline differences in WER between platforms (Fig. 4) is a useful starting point for better analyzing how WER interacts with aspects of speaker and speech characteristics, which will be discussed below.

Social Media Platforms. Results of the LMER model and omnibus testing with ANOVAs reveal significant differences among social media platforms, with YouTube (YT) consistently returning the lowest error rates and Facebook Video (FB) consistently returning the highest, with Microsoft Stream (MS) between the two.

Videoconferencing Platforms. Regarding videoconferencing platforms, Zoom (ZM) and BlueJeans (BJ) perform the best, with very similarly low WERs, while Webex (WX) consistently yields the highest WER. Finally, Google Meet (GM) performs better than WX, but worse than ZM and BJ.

Differences between platform type. Social media platforms consistently outperform video conferencing ones (Est=0.744, SE=0.18, $p<.001$). While this is not surprising since different types of platforms have different purposes, it is surprising to see large differences among different products from the same company (*i.e.,* YT vs. GM). One hypothesis for this difference is that the different types of platforms may employ different types of training data. In particular, though information about platforms' training data is generally proprietary, social media platforms may be more likely to be trained on single-channel or one-speaker datasets, while video conferencing platforms may be more likely to be trained on multi-channel or multi-speaker datasets. Recent research such as Wang et al. (2021) has shown differences in training for single-channel vs. multi-channel automatic speech recognition, and since our current corpus has only monologues, it is possible that social media platforms are better suited to this type of data.

4.2 Speaker Characteristics

Gender Bias. Figure 5 depicts mean WER by gender for the seven platforms. Among the comparisons we conducted for WER and aspects of speaker characteristics, the most consistent finding is a significantly higher WER for perceived male speakers than for perceived female speakers, and this holds across platforms and even when controlling for Perceived Race, Language, Words Per Second, and Platform (Est=0.389, SE=0.0701, $p<.001$). Overall mean WER for male speakers is .076 and for female speakers it is .059, with some variation by platform. For social media platforms, the largest difference between perceived male and perceived female speakers is for YT, where the difference is 0.021, while the smallest is for MS, where the difference is 0.015. For videoconferencing platforms, the largest difference is for GM (0.027), while the smallest one is for WX (0.011).

These results may be surprising given mixed results of previous work on gender in captioning (Tatman and Kasten 2017). However, other studies, including Adda-Decker and Lamel (2005) and Feng et al. (2021) have found that captioning systems perform better for female speakers than for male speakers. Adda-Decker and Lamel (2005) attribute this difference at least in part to female speakers using speech that is either "more standard" or hyperarticulated. Given the highly-rehearsed nature of the speech in our corpus, it is likely that similar patterns may be driving these results.

Language Bias. Another consistent pattern that emerged is the fact that all platforms perform worse for speakers whose first language is not English. In an earlier model that compared U.S. English speakers with non-U.S. English speakers as well as those who speak a first language other than English, no significant differences emerged. As a result, we collapsed the speakers in two categories, those whose first language is English (L1), and those who had a first language other than English (L2). Results of the LMER model indicate significantly lower WER for L1 English speakers (Est=-0.393, SE=0.066, $p<.001$) when controlling for other speaker characteristics, as shown in Fig. 6.

Different platforms have substantial variation in how they perform for L1 English speakers versus others. The mean error rates and differences between L1 and L2 English speakers are illustrated in Table 2. FB and WX show the largest difference in WER, respectively, for social media and videoconferencing platforms between the two speaker groups, while YT and BJ show the smallest, indicating that of the

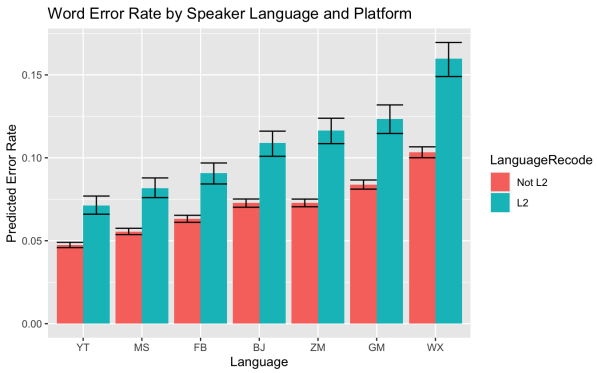


Figure 6: WER by Speaker 1st Language and Platform.

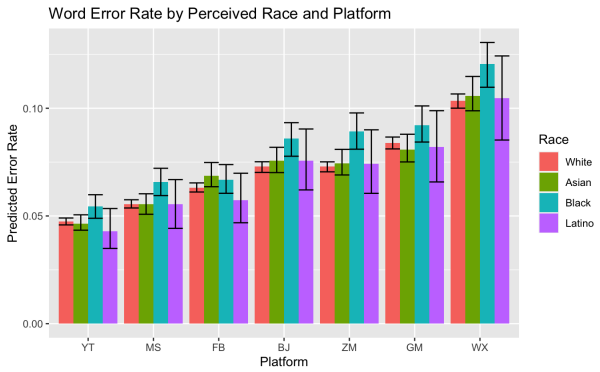


Figure 7: WER by Perceived Race and Platform.

Platform	English	Not English	Difference
BJ	0.086	0.111	0.027
FB	0.074	0.090	0.022
GM	0.097	0.128	0.030
MS	0.061	0.080	0.019
WX	0.115	0.150	0.049
YT	0.051	0.071	0.019
ZM	0.081	0.114	0.032

Table 2: Overall mean WER by Language.

platforms tested, they perform best for speakers whose first language is not English. Differences between L1 and L2 speakers are significant for every platform examined.

Several studies have found significant bias in captioning against speakers who are not speaking in their first language. Adda-Decker and Lamel (2005) observed this pattern in both French and English, and Feng et al. (2021) observed a similar pattern for Dutch. More recently, Chan et al. (2022) observed an increase in WER for non-native speakers of English engaged in a reading task. Accurately captioning the speech of L2 English speakers, even in a highly-rehearsed context, remains a challenge for all platforms.

Race Bias. For differences by speaker Perceived Race, we find that platforms perform slightly worse for non-White speakers, though these differences do not reach statistical significance. Figure 7 shows the differences in WER by speaker race, when controlling for other speaker demographic features. Of note is that Latino speakers are less well-represented in the dataset (N=14), leading to limited ability to draw conclusions about potential differences However, comparisons between White (N=651), Asian (N=104) and Black (N=77) speakers are more reliable.

Overall, the LMER model does not report race as a significant main effect (Est=0.11, SE=0.45, p=0.191). Results of the ANOVA testing from the LMER model indicate that of the seven platforms, none apparently perform significantly worse for Black speakers than for White speakers on average. Additionally, none of the platforms performed worse for Asian speakers than for White speakers. Racial bias in automatic captioning is a major issue, especially in places like the U.S. with substantial racial diversity. Previous studies, such as Koenecke et al. (2020) found robust differences in caption accuracy between Black and White American speakers. The lack of significant differences in the current study should not be interpreted as evidence that such biases do not exist; rather that in this corpus of highly-rehearsed speech by prominent individuals, racial differences in WER may not be as evident as in more casual speech situations. Indeed, TED talks represent a style where the speaker has motivations to pay extremely close attention to their speech, thus predicting the use of a highly-rehearsed style that would select against non-standard linguistic variants (Labov 1966, 1972).

Age bias. Since we provide age labels, we evaluated the presence of age bias. We initially noticed some bias for the younger and oldest groups, but our LMER model did not confirm this result as statistically significant (Est=0.164, SE=0.121, p=0.174).

Inter-attributes bias. Inspired by the methodology in Jaiswal et al. (2022), we used our LMER model to find cases of interdependence among speakers characteristics (*e.g.,* interaction between race and English as a second language). However, we did not find statistically significant cases in which such interdependences matter.

4.3 Speech Characteristics

Speech Rate Bias. When using preliminary descriptive statistics to visualize the data, we noticed an interesting pattern with respect to speech rate in Words Per Second (WPS). LMER model results indicate a significant effect of WPS on WER, though interestingly the distribution of errors is somewhat bimodal, as we observe a general trend such that error rates increase with both fast speech and with slow speech. This can be seen in Fig. 8, which shows mean WER by WPS. Note especially the concentration of errors at the slowest speech rates and increase for the fastest ones.

This result is likely attributable to properties of the captioning systems that have not been trained on extremely fast or extremely slow speech. Both fast and slow speech rates can result in a distortion of sounds that may present additional challenges for captioning systems, but future work should also examine this in greater detail.

F_0 Bias. We aggregate F_0 over each talk, considering both the mean and median F_0 values, and find that the outcomes

Figure 8: WER by Words Per Second and Platform.

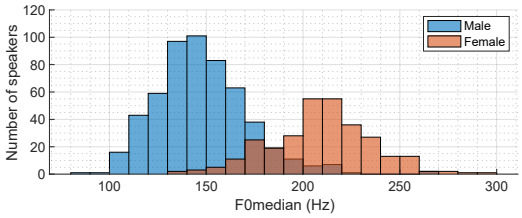


Figure 9: Median F_0 distribution of perceived gender.

are comparable when assessing their impact on WER. In Fig. 9 we can see that gender and median F_0 are, as expected, highly correlated, with low frequencies associated with male speakers and high frequencies associated with female speakers.

To better understand the role of F_0, we used our LMER model to determine whether F_0 remains a good predictor of WER after filtering out the effect of gender. We present the findings in Fig. 10, where we can see that (if we exclude the tails with fewer samples) lower F_0 results in slightly lower WER, while higher F_0 results in slightly higher WER. However, these differences are not large enough to confirm that the gender bias is exclusively due to the F_0 differences.

Speech Readability Bias. We analyze the presence of readability bias defined as changes in the WER with respect to the New Dale-Chall readability score (Sternlicht and Edgar 1995) of the transcripts. We use this score as an indicator for lexical frequency and syntactic complexity in terms of average sentence length measured by number of words, where higher values mean that the transcript contains fewer common words and longer sentences, and lower values mean a transcript with more common words and shorter sentences.

Our LMER model results did not indicate statistically significant interactions between speaker characteristics and readability. However, from Fig. 11, we can surprisingly see that readability scores less than 7 (*i.e.,* higher lexical frequency and shorter sentences) strongly correlate with higher WERs, while for scores higher than 7 (*i.e.,* lower lexical frequency and shorter sentences), the WER is lower and mostly stable among all platforms.

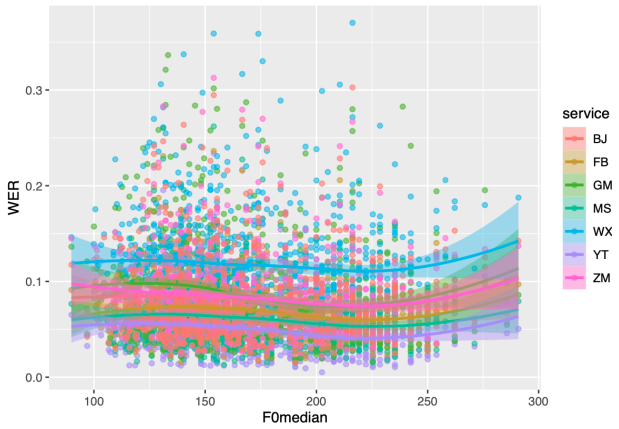


Figure 10: WER by Median F_0 and Platform.

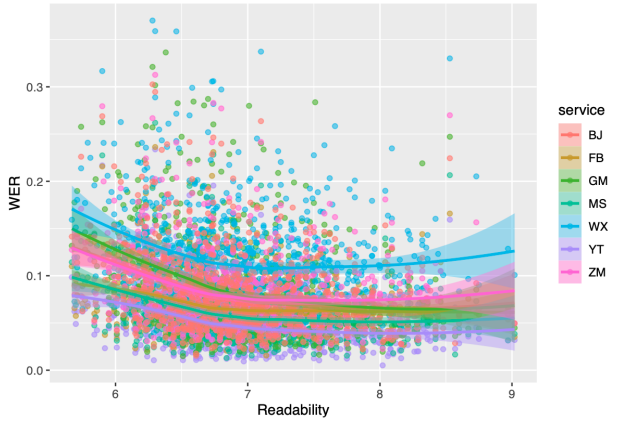


Figure 11: WER by Readability (Dale-Chall) and Platform.

5 Discussion

In this section we first posit explanations for the bias we observed in §4 and propose mitigation strategies. Then, we summarize the feedback from the companies running the platforms after disclosing our results. Finally, we discuss the limitations, broader perspective, and ethics.

5.1 Bias Explanation and Mitigation

Since we have no visibility on how the platforms internally work, the bias explanations we present here are all hypotheses supported by our observations and our understanding of how speech recognition systems work in general.

Gender Bias. The strongest bias we observe based on this analysis is between speakers who are perceived as male and those who are perceived as female. Although this was not fully explained by our F_0 bias analysis, other speech features such as voice intensity may also correlate with gender. Those different features may be harder to capture because of hardware (speakers and microphones) or software (machine learning) limitations. A possible way to mitigate this would be to use different approaches that are tailored to different genders. For example, to tailor the model to have different settings depending on the speaker's baseline

pitch. An alternative explanation for the gender bias is related to how the speakers themselves use linguistic variation. In sociolinguistics, there is ample evidence that female speakers are more likely to employ a more formal style, especially in situations where they may be negatively stereotyped (O'Barr and Atkins 2005). The current dataset represents only highly-rehearsed speech by prominent individuals, which would likely prompt a very careful style, especially for speakers from marginalized groups. This provides motivation for considering the inclusion of speech style in automatic captioning systems, since less formal speech frequently shows a higher WER (Koenecke et al. 2020).

English as First or Second Language. We observed significant bias in WER for individuals who are using English as a second language and for individuals that have unusual speech rates. One explanation is that these speakers may be under-represented in the training data that speech recognition systems use to learn how to transcribe speech. This is especially pernicious for non-native English speakers, given estimates that up to three-quarters of the world's English speakers do not speak it as their first language (Crystal et al. 2003). A way to mitigate this would be to encourage platforms to re-train the speech recognition models using a larger set of samples from under-represented groups. Another way, specific for the language bias, is that platforms could leverage their existing acoustic models to contain information about likely error types in English based on a speaker's first language.

Lexical Frequency and Sentence Length. Our readability bias analysis has shown that the talks that use more common words and shorter sentences are associated with larger WERs. Intuitively, we were expecting the opposite results since those talks would likely be easier to understand for a human according to our readability score. A possible explanation is that some speech characteristics we did not analyze, such as other phonetic properties of a more casual speech style, may be correlated with the use of common words and shorter sentences. This result should reinforce the need, as a possible mitigation strategy, to consider a larger variety of speech styles when training captioning systems.

Type of Platform. Overall, the WER is larger for videoconferencing platforms with respect to social media platforms. As mentioned above, one hypothesis about this difference is that the social media platforms may have been more likely to be trained on single-channel data more similar to our corpus, while video-conferencing platforms may be trained on multi-channel data. An alternative explanation for this may be that video-conferencing platforms produce the transcripts in real-time, while social media platforms typically take hours. This additional time may give the opportunity for social media platforms to use more sophisticated recognition algorithms with respect to videoconferencing platforms. To mitigate this, videoconferencing platforms can, after the end of a meeting, replace the real-time transcripts with transcripts generated using more accurate offline algorithms.

5.2 Feedback from the Platforms

We shared our results with the companies running the platforms we analyzed and received the following feedback. YT and MS confirmed that our findings are in line with internal testing, while WX replied that their product has not been trained for lecture-style talks such as the ones from TED and that explains the higher error rate. The remaining platforms did not give specific comments to our results.

5.3 Limitations

Our methodology has the following limitations.

Ground-truth Correctness. We are assuming the transcripts from the TED website to be correct (*i.e.,* our ground-truth), but human errors are always possible. An error in the ground-truth would result in false errors when calculating the WER. To evaluate the effect of this limitation, we manually checked random TED talk transcripts without finding errors. Another limitation of our ground-truth is that the number of samples is different, depending on which speaker characteristic we consider. Our statistical analysis takes this into account using 95% confidence intervals: cases with less samples and higher variance will result in larger confidence intervals (represented as error bars in our results).

Ground-truth Generalization. The new MonoTED corpus we use as ground-truth is based on TED talks, so it may be more polished and rehearsed than general speech, potentially limiting the generalization of our results. To address this, future work should also carefully examine the effects of speech style on the bias. However, despite this limitation, we are still able to show the presence of bias in our context.

One of the main reasons we decided to use TED-based material and keep highly-rehearsed ground truth is that we were able to produce a corpus that is large-scale (large number of speakers/talks and long durations), that is accurate (manually annotated), that has English monologues that allow us to isolate different speaker characteristics (including nonverbal ones such as race, that can be perceived by watching the videos). This is also the motivation for the success and impact of other corpora based on TED, such as TED-LIUM (Hernandez et al. 2018), which cannot be used directly in our context, but that we used as inspiration.

Perceived Speaker Characteristics. In this work, we could not find enough ground-truth for speaker characteristics such as gender, race, and English language status from unbiased sources to provide a representative sample of labeled speakers; therefore, we relied on a combination of Wikipedia crawling, manual qualitative research on speaker background, and impressionistic coding. As a result, it is likely that we have not captured the true characteristics of all 846 speakers in the corpus, especially given that gender and racial identities can be fluid. Existing work (Field et al. 2021; Ogbonnaya-Ogburu et al. 2020) has discussed how giving perceived labels is non-trivial and handled inconsistently in the research community. However, given the size of the corpus and our research questions, we believe our labeling methods were adequate for our purposes. We hope that this work can act as a first pass for understanding how race and gender may affect bias, though future work should carefully consider how it labels and measures complex and evolving aspects of speaker characteristics.

English Variety Determination. We primarily relied on online information from Wikipedia to determine each

Related work	Platform	Bias (higher error)	No/low bias
Adda-Decker (2005)	Generic ASR	Gender (M)	-
Feng et al. (2021)	DNN	Gender (M), Age (>75), Language (L2)	-
Meyer et al. (2020)	DeepSpeech	Language (L1 non-US)	Gender
Lima et al. (2019)	Voice Assistants	Gender (M), Language (various)	-
Koenecke et al. (2020)	Commercial ASR	Gender (M), Race (black), Location (various)	-
Chan et al. (2022)	Otter	Language (L2)	-
This work	Social media and videoconf.	Gender (M), Lang. (L2), WPS tails, Readability (<7)	Age, Race, Med. F_0

Table 3: Transcription error bias detected by related work.

speaker's place of birth and residence to determine their English variety (and then use perception for the ambiguous cases). A limitation of this method is that Wikipedia information may be inaccurate, and even when accurate, place of birth or residence may not correspond to the actual English variety of the speaker. To be confident that this methodology was still viable for us, we manually verified a sample corresponding to 15% of the speakers in our corpus without detecting significant mislabelings.

Non-speech. We try to avoid the inclusion of talks that contain music, multiple speakers, or external videos. To do this we rely on markers in the transcripts. However, sometimes such markers are missing and in that case, we still include the talk. To reduce the effect of this limitation we did an outlier analysis with respect to the WER and manually checked each outlier to remove talks violating our assumptions.

Non-determinism. We observed that most platforms produce different transcripts when they process the same video. We call this non-determinism. We reduce the effect of non-determinism by using a large sample of data and by considering a 95% confidence interval for statistical significance. Another way to address this problem, which we leave as future work, is to repeat experiments with the same video multiple times, and average the WER.

5.4 Broader Perspective

Our work has the following broader impacts: (*i*) we are the first to analyze transcription error biases in popular social media and videoconferencing platforms, analyzing new characteristics such as speech rate, and pointing out results that are different from previous work (Koenecke et al. 2020); (*ii*) we are the first to release a corpus that, contrary to existing corpora, is optimized for analyzing such platforms, due to the large scale, and the simultaneous presence of a single voice and manually curated ground truth. This will enable future research to extend our study and/or to compare with it. Further, our corpus, methodology, and results may be used by the companies running such platforms to reduce their biases and overall error rate for the benefit of consumers.

5.5 Ethics

Obtaining Ground-truth. All the data we used as ground truth either comes from public sources (TED website and Wikipedia) or is created by us. We accessed the data respecting the terms and conditions of every source, and in cases in which we had to automatically scrape the content from a website, we throttled the scraping operation to minimize any disturbance to the website.

Human Effort and IRB Approval. All human effort needed to label perceived speaker characteristics was conducted by one of the authors and two non-authors working on the same team as one of the authors. Due to the presence of people (and their real names) in the corpus of TED monologues, we consulted our IRB to determine whether we needed approval to conduct our study using this publicly available dataset. The IRB indicated that this study was exempt from IRB review. Note that since TED talks are released under a Creative Common license, we did not need to obtain speaker permissions for this research.

Mislabeled Speaker Characteristics. Due to the difficulty of obtaining the real self-identified speaker characteristics, we relied on perceived labels. A drawback of this choice is that we might use and publish perceived labels that are different from self-identified ones, potentially offending a speaker that does not want to be associated with the assigned perceived label. To mitigate this ethical concern, our data is annotated with a disclaimer that our perceived labels should not be considered as the real speaker labels. Also, should a speaker be offended by a perceived label, we will offer to update our published data by either updating the label with the desired one, or by removing the label entirely.

Platforms Usage and Transcript Sharing. We used the platforms in compliance with their terms and conditions, and paid for all the needed subscriptions and licenses. Also, the terms and conditions of all the platforms we analyzed do not prohibit or require us to request permission to publish the transcripts generated by them.

Implications for the Platforms. Our results provide some insight on platform performance with respect to transcribing the TED talks in our corpus; however, some of the platforms (WX) claim they have not been optimized for that. Therefore, our results should not be interpreted as a conclusion that one platform is better than another in general or absolute terms. However, the bias we measured can still be used by each platform to identify sources of transcription errors and improve their product.

6 Related Work

Transcription Bias. Several existing works have analyzed transcription bias for various platforms and speaker characteristics, as shown in Table 3. Our results confirm most of their findings for gender and English language variation; however, we did not find significant bias for age and race.

Fairness. We assume a system to be fair when its bias is

Corpus	Source material	Duration	Labels	Limitations
CORAAL	108 interviews	30m each	Gender, age, social class, location	Two speakers, single race (African American), limited samples
Voices of California	109 interviews	60m each	Gender, age, race, location	Two speakers, limited samples
Broadcast News	Radio/TV	600h total	Gender	Multiple speakers, limited labels
Common Voice	86,942 clips	5s each	Gender, age, English accent	Clips very short and unnatural, limited labels
Artie Bias	1,712 clips	5s each	Gender, age, English accent	Clips very short and unnatural, limited labels
LibriSpeech	Audiobooks	960h	Gender	Clips very long and unnatural, limited labels
TED-LIUM 3	2,351 talks	12m each	-	Multiple speakers, music/videos, incomplete transcripts, no labels
MonoTED (ours)	846 talks	14m each	Gender, age, race, language status	Labels assigned from public sources or perceived characteristics

Table 4: Corpora comparison with MonoTED (our corpus of labeled TED monologues).

zero; however, other related work about fairness in different recognition contexts (*e.g.,* Hajavi and Etemad (2023); Fenu et al. (2021); Meng et al. (2022)) define fairness in other ways (*e.g.,* systems having the same probability of error among different demographic groups).

Corpora. In addition to TED-LIUM (described in §3.1), other English corpora used for transcription bias include CORAAL (Kendall and Farrington 2018), Voices of California (Stanford 2012), Broadcast News (Lamel et al. 2004) and LibriSpeech (Panayotov et al. 2015). Other corpora exist with transcript ground-truth, typically to train voice-recognition systems, *e.g.,* Common Voice (Ardila et al. 2020) and Artie Bias (Meyer et al. 2020), but they are not suitable to study transcription bias due to the short sentences and the likelihood that they have been already used to train those systems. Table 4 reports a detailed comparison of these existing corpora and MonoTED.

Other Works. Some other works focus on studying biases in other recognition contexts. The following is a non-exhaustive list: *face recognition* (*e.g.,* Jaiswal et al. (2022); Kyriakou et al. (2019); Barlas et al. (2019); Raji et al. (2020); Buolamwini and Gebru (2018)), where face recognition bias is evaluated for several personal characteristics; *smart speakers wake-word recognition* (*e.g.,* Chen et al. (2021); Dubois et al. (2020)), where the authors show that certain categories of people are more likely to misactivate popular smart speakers; *automated speaker recognition* (Hutiri and Ding 2022; Fenu et al. 2021; Meng et al. 2022; Hajavi and Etemad 2023), where the authors analyze the fairness in recognizing speakers.

7 Conclusion

We analyzed, for seven popular social media and video-conferencing platforms, the transcription error bias among groups of speakers with different characteristics using a new corpus (MonoTED) of 846 speakers, 1.8M spoken words, and 194 hours of speech.

As far as we know, this is the first time all these characteristics are being analyzed at the same time and in an opaque context such as the popular platforms we consider. From this study we have learned that **transcription error bias does exist**, with significant occurrences with respect to perceived gender, speaker first language, speech rate, and speech readability, and no significant occurrences with respect to perceived race, age, and F_0 (when removing the interaction with gender). This result is surprising since previous work (in the different context of general-purpose automated voice recognition systems) reported statistically significant bias also for race (Koenecke et al. 2020).

One item of note is that the corpus we used for this analysis consists of highly-rehearsed speeches delivered by prominent individuals. Given that these biases still emerge when examining such tightly-controlled, highly-rehearsed data, it is likely that the biases we observe are amplified when systems transcribe messier, real-world data. Based on our experimental observations, we hypothesize that most biases are likely due to the fact that these platforms have been trained with samples that did not account for systematic variation between speakers with different characteristics. As a result, the platforms under-perform for speakers who were either not well-represented in training data, or who were using a linguistic style that differed from that of the training data. Possible ways to reduce such biases are: re-training these system with more representative training samples, training them to account for differences in speech style, and adjusting acoustic models to be better-suited for speakers of different linguistic backgrounds.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. This research was partially supported by the NSF (SaTC-1955227) and Consumer Reports Innovation Lab.

References

Adda-Decker, M.; and Lamel, L. 2005. Do speech recognizers prefer female speakers? In *Interspeech*.

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2020. Common voice: A massively-multilingual speech corpus. In *LREC*.

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2019. Social b (eye) as: Human and machine descriptions of people images. In *ICWSM*, volume 13, 583–591.

Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. of Statistical Software*, 67(1): 1–48.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 77–91. PMLR.

Chan, P. M.; Choe, J.; Li, A.; Chen, Y.; Gao, X.; and Holliday, N. 2022. Training and typological bias in ASR performance for world Englishes. In *Interspeech*.

Chen, Y.; Bai, Y.; Mitev, R.; Wang, K.; Sadeghi, A.-R.; and Xu, W. 2021. FakeWake: Understanding and Mitigating Fake Wake-up Words of Voice Assistants. In *CCS*, 1861–1883.

Crystal, D.; et al. 2003. *English as a global language*. Cambridge university press.

Dubois, D. J.; Kolcun, R.; Mandalari, A. M.; Paracha, M. T.; Choffnes, D.; and Haddadi, H. 2020. When Speakers Are All Ears: Characterizing Misactivations of IoT Smart Speakers. In *PETS*.

Feng, S.; Kudina, O.; Halpern, B. M.; and Scharenborg, O. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Fenu, G.; Marras, M.; Medda, G.; Meloni, G.; et al. 2021. Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition. In *Interspeech*, 1892–1896. International Speech Communication Association.

Field, A.; Blodgett, S. L.; Waseem, Z.; and Tsvetkov, Y. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *AACL/IJCNLP*, 1905–1925.

Gerhard, D.; et al. 2003. *Pitch extraction and fundamental frequency: History and current techniques*. Univ. of Regina, SK, CA.

Hajavi, A.; and Etemad, A. 2023. A Study on Bias and Fairness in Deep Speaker Recognition. In *ICASSP*, 1–5. IEEE.

Hernandez, F.; Nguyen, V.; Ghannay, S.; Tomashenko, N.; and Esteve, Y. 2018. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *SPECOM*, 198–208. Springer.

Hutiri, W. T.; and Ding, A. Y. 2022. Bias in automated speaker recognition. In *FAccT*, 230–247.

Jaiswal, S.; Duggirala, K.; Dash, A.; and Mukherjee, A. 2022. Two-Face: Adversarial Audit of Commercial Face Recognition Systems. In *ICWSM*, volume 16, 381–392.

Kendall, T.; and Farrington, C. 2018. The corpus of regional African American language. https://oraal.uoregon.edu/coraal. Accessed: 2024-04-11.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *Proc. of the National Academy of Sciences*, 117(14): 7684–7689.

Kyriakou, K.; Barlas, P.; Kleanthous, S.; and Otterbacher, J. 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *ICWSM*, volume 13, 313–322.

Labov, W. 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, W. 1972. *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Lamel, L.; Gauvain, J.-L.; Adda, G.; Adda-Decker, M.; Canseco, L.; Chen, L.; Galibert, O.; Messaoudi, A.; and Schwenk, H. 2004. Speech transcription in multiple languages. In *ICASSP*, volume 3, iii–757. IEEE.

Lima, L.; Furtado, V.; Furtado, E.; and Almeida, V. 2019. Empirical analysis of bias in voice-based personal assistants. In *WWW 2019 Companion*, 533–538.

MathWorks, Inc. 2018. Estimate fundamental frequency of audio signal. https://www.mathworks.com/help/audio/ref/pitch.html. Accessed: 2024-04-11.

Meng, Y.; Chou, Y.-H.; Liu, A. T.; and Lee, H.-y. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models. In *ICASSP*, 3258–3262. IEEE.

Meyer, J.; Rauchenstein, L.; Eisenberg, J. D.; and Howell, N. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *LREC*, 6462–6468.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. *Interspeech*, 2616–2620.

O'Barr, W. M.; and Atkins, B. K. 2005. "Women's Language" or "Powerless". *Language, Communication and Education*, 202.

Ogbonnaya-Ogburu, I. F.; Smith, A. D.; To, A.; and Toyama, K. 2020. Critical race theory for HCI. In *CHI*, 1–16.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*, 5206–5210. IEEE.

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *AIES*, 145–151.

Reid, L. D. 2010. The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *J. of Diversity in higher Education*, 3(3): 137.

Sims, J. P.; Pirtle, W. L.; and Johnson-Arnold, I. 2020. Doing hair, doing race: The influence of hairstyle on racial perception across the US. *Ethnic and Racial Studies*, 43(12): 2099–2119.

Stanford. 2012. Voices of California. http://web.stanford.edu/dept/linguistics/VoCal/. Accessed: 2024-04-11.

Sternlicht, C. J.; and Edgar, D. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Tatman, R.; and Kasten, C. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Interspeech*, 934–938.

Vaessen, N. 2018. JiWER: Similarity measures for automatic speech recognition evaluation. https://github.com/jitsi/jiwer. Accessed: 2024-04-11.

Wang, X.; Kanda, N.; Gaur, Y.; Chen, Z.; Meng, Z.; and Yoshioka, T. 2021. Exploring end-to-end multi-channel ASR with bias information for meeting transcription. In *SLT'21*, 833–840. IEEE.

Woodard, J.; and Nelson, J. 1982. An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, NADC*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see §5.5 for details.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract and the introduction present a summary of contributions and findings that we discussed more in detail in the remainder of the paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see §3 for details.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see §3.1 for details.
   (e) Did you describe the limitations of your work? Yes, see §5.3 for details.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see §5.5 for details.
   (g) Did you discuss any potential misuse of your work? Yes, we explained in §5.5 the ethical implications of using perceived demographic labels and our approach for mitigating them.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see §5.5 for details. Also, to promote reproducibility and support future research, we released MonoTED and the rest of our dataset at https://github.com/NEU-SNS/MonoTED.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, see §5.5 for details.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, we provide citations or URL footnotes for all external tools and data sources we used.
   (b) Did you mention the license of the assets? Yes, see §3.1 for details.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we released our new MonoTED corpus and the code for reproducing our results (see §1).
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see §5.5 for details.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see §5.5 for details.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR[14]? Yes, the introduction section contains a link to our dataset, which we released under FAIR principles.
   (g) If you are curating or releasing new datasets, did you create a Datasheet[15] for the Dataset? Yes, our dataset has a Datasheet attached.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

[14]FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

[15]Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. Communications of the ACM, 64(12): 86–92.

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, in §5.5 we discuss the absence of risk and our IRB exemption.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA