A Deep Dive into the Disparity of Word Error Rates across Thousands of NPTEL MOOC Videos

Anand Kumar Rai, Siddharth D Jaiswal, Animesh Mukherjee
Indian Institute of Technology Kharagpur, India
anand@kgpian.iitkgp.ac.in, siddsjaiswal@kgpian.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Automatic speech recognition (ASR) systems are designed to transcribe spoken language into written text and find utility in a variety of applications including voice assistants and transcription services. However, it has been observed that state-of-the-art ASR systems which deliver impressive benchmark results, struggle with speakers of certain regions or demographics due to variation in their speech properties. In this work, we describe the curation of a massive speech dataset of 8740 hours consisting of $\sim$ 9.8K technical lectures in the English language along with their transcripts delivered by instructors representing various parts of Indian demography. The dataset is sourced from the very popular NPTEL MOOC platform. We use the curated dataset to measure the existing disparity in YouTube Automatic Captions and OpenAI Whisper model performance across the diverse demographic traits of speakers in India. While there exists disparity due to gender, native region, age and speech rate of speakers, disparity based on caste is non-existent. We also observe statistically significant disparity across the disciplines of the lectures. These results indicate the need of more inclusive and robust ASR systems and more representational datasets for disparity evaluation in them.

Introduction

Automatic speech recognition (ASR) systems have become increasingly prevalent in recent years with applications including voice assistants, transcription services and language translation. Auto-generated transcripts serve an integral part in providing equitable access of online video content to a wide variety of individuals and groups while voice based assistants enable users to avail a lot of online services with voice-based commands. In the past two decades, designing efficient ASRs have been an active area of research resulting in substantial advancement in the accuracy of these tools (Hannun 2021).

Learning Through Video Lectures

The advent of COVID-19 pandemic has hastened the pace of adoption of online education and, Massive Open Online Courses (MOOC) platforms like NPTEL (Krishnan 2009), Coursera [1] etc. play a pivotal role in it. The transcripts of these videos are typically used to generate captions (Kent et al. 2018) for the videos. Such captions reduce the disconnect between the viewer and speakers with distinct accents and speaking styles. Many speakers rely on ASR tools like YouTube [2], Zoom [3], Otter.AI [4], etc. for generating automated transcriptions of their videos to reduce their own manual effort. Thus the error from these ASRs can directly impact the video captions and hence the overall understanding of the viewer.

Disparity in Caption Generation

These platforms are expected to work without disparity for all speech rates, accents, tonality, independent of gender or age to ensure the generation of correct transcripts, and captions, so that the listener does not misinterpret the speaker. However, concerns have been raised about the potential for such systems to exhibit bias (Feng et al. 2021; Koenecke et al. 2020; Tatman 2017) toward certain demographics.

There has been ongoing research evaluating disparities in ASRs toward different racial (Koenecke et al. 2020), gender (Tatman 2017) and other groups (Feng et al. 2021). Most of these studies have focused on Western languages, accents, and social demographic groups. There have also been efforts to create public datasets like Artie Bias (Meyer et al. 2020), CORAAL (Kendall and Farrington 2018), AAVE (Rickford 1999), TED-LIUM (Rousseau, Deléglise, and Esteve 2012), Librispeech (Panayotov et al. 2015) etc., with dialectical and vernacular variations to evaluate the performance of existing ASR platforms, but these too have primarily focused on Western speakers. Moreover, there has been a lack of datasets exploring technical or pedagogical content.

ASR Disparity Impact in Indian Context

The problem of disparity in ASR systems is of particular concern in the Indian context, given the country's highly diverse linguistic landscape across its vast geography. With the rapid advancement in digital economy in the country, it has become the second largest country in terms of internet

[1] https://www.coursera.org

[2] https://blog.research.google/2009/12/automatic-captioning-in-youtube.html

[3] https://blog.zoom.us/zoom-auto-generated-captions

[4] https://otter.ai

users[5] with almost 1 billion users. Moreover, a significant number of students within India consume educational content from YouTube or other e-learning platforms (Buddayya, LG et al. 2019), but no studies have evaluated the transcription accuracy of ASRs for educational content delivered by Indian speakers. Due to the scale at which content is disseminated and consumed in India, it is important for researchers to study the disparities in widely deployed and state-of-the-art ASRs to ensure proper learning pedagogy is maintained.

Our Contributions

In this study, we evaluate the correctness of captions for two popular ASR platforms – YouTube Automatic Captions[6] and OpenAI's Whisper (Radford et al. 2022) on a new large-scale dataset developed by us. This is called the Technical Indian English (TIE) dataset, and has more than 9800 educational videos from the NPTEL MOOC learning platform, comprising over 8700 hours of content and 62 million spoken words. These lecture videos are delivered by 332 speakers belonging to different gender, caste, age and regions within India. Ours is the first such annotated dataset on Indian educational videos and our in-depth analysis of the performance of the two ASRs on this dataset shows that disparities exist not only between the two platforms but also between categories of speakers such as gender and experience level on each platform. While studies have been conducted to evaluate YouTube's ASR capabilities, ours is the first study to evaluate OpenAI's Whisper for its accuracy on technical English content, spoken by non-native English speakers.

We now list the research questions that we address in this study.

RQ1. Since no prior dataset exists for studying ASR disparities across educational videos for non-native speakers the first question we were faced with was to develop such a dataset from scratch. In response we built the TIE dataset.

RQ2. How do the two ASRs perform on the TIE dataset? We perform statistical tests to identify which of the two platforms is better at transcribing technical educational content spoken by Indian speakers.

RQ3. Do the two ASR platforms exhibit disparity in performance among the different categories within each attribute like gender, caste, age, etc. Here as well we perform statistical tests to identify the disparities in performance not only for a given platform but also across platforms.

Finally, we list our contributions. In this paper, we propose the TIE dataset of 9860 audio files spanning a massive duration of 8740 hours worth of lectures sourced from the NPTEL MOOC platform and annotated for attributes corresponding to the lectures and the 332 instructors who have delivered these lectures. We believe that this dataset can serve as an excellent benchmark for studying the performance of various AI softwares including ASRs. We audit the YouTube Automatic Captions and OpenAI Whisper ASRs for existing disparities towards gender, caste, teaching experience, native region of the speakers, speech rate and technical discipline of the lectures.

Our analysis reveals a significant disparity against non-native English speakers in the Indian population, particularly for male speakers, speakers from southern India and those with a slow speech rate. In addition, we found that the experience of speakers and the topic of speech had an impact on the accuracy of ASRs. We also observed that while Whisper had better overall accuracy than YouTube, disparities were higher in Whisper. We conclude by discussing possible causes of these disparities and proposing strategies to address them.

Related Work

Automatic Speech Recognition

ASR systems, designed to recognize and translate spoken language have been actively developed since the 1950s. Today, this technology has become highly ubiquitous (Smith 2020) and has various use cases ranging from generating captions for pre-recorded videos [7] and live videos [8] to personal voice assistants like Siri [9], Alexa [10] and Cortana [11].

Audit of ASRs

Recently, there has been a growing interest in auditing (Ngueajio and Washington 2022) these ASR platforms for potential disparities against various gender (Adda-Decker and Lamel 2005; Tatman and Kasten 2017; Tatman 2017; Garnerin, Rossato, and Besacier 2019) and racial groups (Koenecke et al. 2020; Tatman and Kasten 2017). (Koenecke et al. 2020) have highlighted the existing racial bias against Black speakers prevalent in state-of-the-art commercial ASRs by benchmarking them against CORAAL (Kendall and Farrington 2018) and AAVE (Rickford 1999) datasets having high representation of speakers from African American community. (Adda-Decker and Lamel 2005; Garnerin, Rossato, and Besacier 2019) have pointed out the gender bias in ASR systems which favors female speakers when they benchmarked the ASRs performance for English and French news broadcast dataset. (Vipperla, Renals, and Frankel 2010) have audited the impact of speaker age on the ASR performance. Most of the existing studies fall under the purview of *black box audits* (Sandvig et al. 2014) due to the lack of access to model architecture and training data for ASRs supplied by commercial vendors (Koenecke et al. 2020; Tatman and Kasten 2017).

Non-Native English Datasets and Models

The authors in (Meyer et al. 2020) proposed a dataset which had speaker age, accent and gender annotated along

[5] https://www.statista.com/statistics/271411/number-of-internet-users-in-selected-countries/

[6] https:/blog.research.google/2009/12/automatic-captioning-in-youtube.html

[7] https://blog.research.google/2009/12/automatic-captioning-in-youtube.html

[8] https://blog.zoom.us/zoom-auto-generated-captions

[9] https://www.apple.com/in/siri/

[10] https://developer.amazon.com/en-US/alexa

[11] https://www.microsoft.com/en-us/cortana

with the their associated audio and benchmarked the DeepSpeech model (Hannun et al. 2014). They reported more accurate transcriptions of US English compared to Indian English. Some of the other curated datasets to facilitate benchmarking of ASRs on non-native speakers are L2-Arctic (Zhao et al. 2018), EdAcc (Sanabria et al. 2023) and AccentDB (Ahamad, Anand, and Bhargava 2020).

In addition (Hinsvark et al. 2021) presented a comprehensive survey on bias in ASR systems due to variety in the speakers' accents. (Sullivan, Shibano, and Abdul-Mageed 2022), (Shibano et al. 2021) and (Vu et al. 2014) presented different approaches like transfer learning and language model decoding that allow ASRs to perform better on non-native speaker dataset.

The present work. In order to bridge the lack of data we present a large-scale dataset of technical lecture videos comprising 8740 hours of speech by 332 Indian speakers on more than 20 diverse lecture topics. Each video file has been annotated with demographic attributes like teaching experience, gender, caste and native region of respective speaker plus audio metadata like speech rate, discipline and topic of the lectures. Unlike (Meyer et al. 2020), our dataset is highly diverse comprising speakers from all the four regions of India. In addition, our dataset is richer than AccentDB (Ahamad, Anand, and Bhargava 2020), another Indian ASR dataset in terms of the number and diversity of native speakers represented and the total duration of speech data available. In particular, while AccentDB includes only speakers with native languages such as Bangla, Malayalam, Odiya and Telugu and has a duration of only 9 hours, our dataset covers a wider range of Indian languages and comprises a total duration of 8740 hours. Next, existing literature (Tatman and Kasten 2017; Tatman 2017) has focused on US speakers' whose native language is English. Ours is one of the first *large-scale* study focusing on non-native English speakers from the Global South. Finally, ours is the first study to perform a comprehensive audit of OpenAI Whisper on technical speech by non-native English speakers.

Dataset and Platforms

In this section, we give a detailed overview of the dataset that we curate for this audit study, along with a description of the platforms that we evaluate this dataset on.

TIE Dataset

In this study, we curate a new large-scale dataset[12] of 9,860 lecture videos from the NPTEL (National Programme on Technology Enhanced Learning) (Krishnan 2009) platform, which is a government-funded joint initiative of the Indian Institutes of Technology (IITs)[13] and the Indian Institute of Science (IISc)[14] and provides high-quality educational content in the form of video lectures, online courses and other resources to students and educators throughout India. The NPTEL platform has more than 2500 courses across 29 engineering and non-engineering disciplines with over 78,000 videos distributed on both the NPTEL website[15] as well as YouTube. The official YouTube channel of NPTEL[16] has more than 20,000 videos from 22 disciplines with more than 2 million subscribers and around 400 million cumulative views on all videos. The platform is similar to other MOOC platforms like MIT OCW [17], Coursera [18], edX [19] etc.

For this study, we sample 9860 videos from NPTEL's YouTube channel for which both ground truth transcripts and YouTube captions are available in English. These lecture videos have been delivered by 332 speakers of Indian origin belonging to both genders – male & female, reserved and unreserved caste groups (Merriam-Webster 2021), multiple experience ranges and residing in diverse geographical regions. The speakers are faculty members at the premier educational institutes of India (e.g., IITs and IISc). Thus our dataset is both acoustically and linguistically rich and diverse. Each lecture video is ≈ 53 minutes long, giving us a massive 8740 hours of lecture videos with more than 62 million spoken words in total. We note that even though each individual speaker speaks in ≈ 31 videos, there are variations in the video content, speaker's tone, accent, and style. Hence each video lecture can be considered to be a unique input video resulting in 9860 unique data points.

All the video lectures collected as part of this dataset are non-interactive and pre-recorded technical monologues delivered by the lecturers. Such monologues are generally less noisy than lectures delivered in interactive settings. This also helps us in avoiding speech preprocessing and enhancement tasks which could potentially add bias to the dataset. More details regarding the dataset metadata are available in Table 1.

Dataset Preparation

We prepare the TIE dataset by sampling a list of courses from the NPTEL (Krishnan 2009) website in two phases. In the first phase, we collect metadata for all courses – name, discipline, institute, instructor and the course weblink. This resulted in a corpus of 2567 courses. Each course has ≈ 31 lecture videos, for which we collected the titles and weblinks. This resulted in a total dataset size of 78,222 lecture videos. In the second phase, we filter out only those videos which were hosted on YouTube and had ground truth transcripts available. This filtered dataset has 9860 lecture videos along with all metadata information mentioned previously. We collect all video files and extract the audio tracks in MP3 format. The size of the final corpus of 9860 audio files is approximately 700 GB.

Dataset Annotation

We annotate each video file in our dataset with multiple features for the speaker viz. gender, caste, experience, lecture

[12] Available at https://github.com/raianand1991/TIE/
[13] https://en.wikipedia.org/wiki/Indian_Institutes_of_Technology
[14] https://en.wikipedia.org/wiki/Indian_Institute_of_Science
[15] https://nptel.ac.in/courses
[16] https://www.youtube.com/@iit
[17] https://ocw.mit.edu
[18] https://www.coursera.org
[19] https://www.edx.org

Attribute	Category	% Speakers	% Lectures	# Hours	#Words in ground truth transcripts
Gender	Female	5.4	5.8	498	3.7 M
	Male	94.6	94.2	8242	58.5 M
Caste	Res.	27.4	27.0	2285	16.2 M
	Unres.	72.6	73.0	6455	46.0 M
Experience	$\leq$ 1980	14.5	14.2	1287	9.0 M
	1981-90	22.9	22.1	1895	13.1 M
	1991-00	32.8	33.9	3033	21.7 M
	$\geq$ 2001	29.8	29.8	2526	18.5 M
Native Region	East	35.2	37.9	3405	23.1 M
	West	8.4	6.6	584	4.3 M
	North	21.7	19.2	1663	12.3 M
	South	34.7	26.3	3088	22.5 M
Speech Rate	Slow	72.9	39.1	3442	19.3 M
	Average	71.4	21.4	1880	13.3 M
	Fast	75.3	39.5	3418	29.6 M
Discipline	Engg.	70.5	70.8	6269	44.8 M
	Non-engg.	30.7	29.2	2472	17.4 M
TOTAL				**8740**	**62.2 M**

Table 1: Metadata statistics for the Technical Indian English (TIE) dataset. The attributes and the categories therein, are presented in the rows. % Speakers and % Lectures represent the share of speakers and lectures from 332 speakers and 9860 lectures respectively. # Hours and # Words represent the time and number of words (in Millions) for all lecture videos corresponding to a given attribute's category. There are approximately 30 lectures per speaker, and each lecture runs for ~ 53 minutes and has $\approx$ 6300 words.

discipline, affiliation and native region. Here, we define experience as the year since which the speaker has been affiliated to their institute and the native region as the geographical region to which the speaker belongs – north, east, south or west India. We do the above by scraping the NPTEL course website to collect information regarding the course, discipline, institution, instructor and the Youtube URL for the videos.

Demographic information annotation Next, to identify the demographic information of the speaker, we use a pre-trained BERT-based model proposed by (Medidoddi et al. 2022), which takes as input the individual's full name and returns the gender and caste for the same. This model has a claimed accuracy of 96.06% for gender classification and 74.7% accuracy for caste category classification. Note that our definition of gender represents the perceived gender, and not necessarily the self-identified gender of the speaker. Similarly, the caste of individual speakers represents the perceived caste category which is generally determined by the surname of the individual. We have used only two categories for caste – unreserved (historically privileged groups) and reserved (SC/ST/OBC – historically discriminated groups) for ease of analysis.In order to check the robustness of the accuracy of the caste classification model reported in (Medidoddi et al. 2022) we manually verified its performance for a random sample of 50 speakers from our TIE dataset. We obtained an accuracy of 80% in identifying speaker caste category based on their names for our dataset.

For annotating the experience of speakers, we use the year when they started their teaching career. This information is collected from the speakers' curriculum vitae (CV) or profile page. We created four categories for experience viz. ≤ 1980, 1981-90, 1991-2000, and ≥ 2001 indicating the time range when the speaker started their professional career.

To annotate for the native region, we use the information available in the speaker's CV. If this is unavailable, the surname of the speaker is used to infer the native state as Indian names are based on naming conventions that are region-specific[20]. The inter-annotator agreement (Cohen's $\kappa = 0.71$) of co-authors is used for finalizing the annotation for the native state attribute.

Non-demographic information annotation Finally, we also annotate the dataset with non-demographic information like speech rate and the speaker's affiliation and topic discipline (engineering or non-engineering). Speech rate is a measure of the word utterance rate of the speaker and is calculated in terms of words per minute (wpm). We calculate this value by dividing the total number of words in the ground truth transcript by the lecture duration. We use this speech rate to categorize the dataset into three classes viz. slow, average and fast representing those with $\leq$ 33, 33-66, and $\geq$ 66 percentile speech rates respectively.

Dataset statistics Table 1 presents the statistics for the TIE dataset that we curate as part of this study. We can immediately see that there is a significant skew toward male speakers ($\approx$ 95%) and those belonging to the unreserved caste category ($\approx$ 73%). This is representative of the distribution observed in higher educational institutes in India. (Prathap 2017; Chanana 2006) have reported on the marginalized representation of females in teaching and lead-

[20]https://en.wikipedia.org/wiki/Indian_name

ership profiles in Indian Higher Education institutions and in fact a recent survey[21] from the year 2019-20 amongst Institutes of National Importance in India shows that females occupy only 20% of the teaching positions. Similarly, it has been observed that only 21% of the lecturers in higher education institutes belong to the reserved caste category. In terms of teaching experience, the dataset has a higher representation ($\approx$ 63%) of speakers who started teaching after 1991. The distribution of speakers from the native regions is fairly balanced except for west India which is underrepresented with a share of only 8.4% speakers. We annotate speech rate and the discipline for each lecture, instead of each speaker. Hence, there is an overlap of speakers across these subcategories resulting in the total sum exceeding 100%. In the speech rate category, it can be seen that most lectures are delivered by speakers speaking either slowly or fast (each having a 39% share). As NPTEL hosts lectures from higher educational institutes specializing in engineering, the share of engineering lectures has a majority (70.8%).

Platforms Evaluated

In this study, we audit two ASR platforms – YouTube Automatic Captions (Google 2009) and OpenAI's Whisper (Radford et al. 2022). While Youtube is a full-fledged video-sharing platform with its own proprietary ASR solution to generate captions for uploaded videos, Whisper is an open-source ASR system developed by OpenAI that can generate transcripts in multiple languages.

We now give a brief description of the two platforms under consideration.

- YOUTUBE AUTOMATIC CAPTIONS: YouTube launched its proprietary ASR software – YouTube Automatic Captions in 2009 to generate automatic captions for videos uploaded on YouTube. The captions are textual representations of the audio and get displayed on the screen while the video is being played. The feature was developed to make it easier for people with hearing impairments to access the video content on the platform. As this is a commercial black-box model, the architecture and training data are not available in the public domain.
- WHISPER: This is an open-source ASR system launched by OpenAI in 2022. The architecture is based on an end-to-end encoder-decoder transformer and trained on 680k hours of multilingual and multitask supervised data collected from the web. The authors claim that their model has improved robustness to accents and technical language. The model is available in nine variants of multiple sizes viz. tiny, base, small, medium, and large. The model architecture has been open-sourced and is available for zero-shot transcription of speech audio in multiple languages.

Reasons for choosing these platforms: We use the above two ASR platforms for disparity evaluation due to the following reasons.

- YouTube ASR is directly used for generating captions for the YouTube videos, which is by far the largest commercial video-sharing platform. Thus the transcripts used for caption generation has a significant influence on the accessibility and reach of the videos on the platform. The NPTEL videos used in our TIE dataset have more than 200M views cumulatively along with 831K likes on YouTube, indicating wide content consumption.
- We experimentally observe that Whisper outperforms other open-source ASRs like DeepSpeech (Hannun et al. 2014) and Wav2Vec 2.0 (Baevski et al. 2020) in terms of overall WER on the TIE dataset when evaluated in a zero-shot setting. For this evaluation, we use pre-trained DeepSpeech[22], Wav2Vec 2.0[23] and Whisper[24] models trained on English datasets. The median WER on TIE dataset for these models are 96.4%, 28.6% and 11.8% respectively. The performance of Whisper being substantially better than the other two competing SOTA models, we use it for all our subsequent experiments.

Methodology & Evaluation Metrics

We now describe the experimental methodology and the metrics used to evaluate the observations. We first discuss the process of collecting the ASR-generated transcripts, followed by a description of the metric – word error rate (WER) and the statistical tests performed to determine the disparity among the various categories for each attribute as well as between the two ASRs under consideration.

Methodology

We generate the transcripts for each video in our TIE dataset using the transcription services of both the platforms– YouTube and Whisper. For YouTube, we use the YOUTUBETRANSCRIPT API (Depoix 2022) to generate the textual transcripts and post-process these by removing frame-wise timestamps. To generate transcripts for Whisper, we first extract the audio in MP3 format and provide these as input to the pre-trained base model of Whisper (base.en), trained on an English corpus. Each MP3 file is of approx 50 min duration and inference on Whisper base model took $\approx$ 2 mins per file when executed on a server with 2 Tesla P100 PCIe 16 GB GPUs. The overall inference time in generating transcripts from Whisper for the entire dataset of 9860 files was $\approx$ 250 hours.

Evaluation Metric

We choose WER as the evaluation metric as it is the most commonly used metric for evaluating ASRs, having been used by many studies in literature (Tatman and Kasten 2017; Tatman 2017; Koenecke et al. 2020; Garnerin, Rossato, and Besacier 2019) as well as for benchmarking public datasets. The WER is measured in two steps – (i) a preprocessing step in which we remove all the punctuation marks and standardize the case of all words and (ii) the WER calculation step

[21] https://aishe.gov.in/

[22] https://deepspeech.readthedocs.io/en/r0.9/USING.html

[23] https://huggingface.co/facebook/wav2vec2-large-robust-ft-libri-960h

[24] https://huggingface.co/openai/whisper-base.en

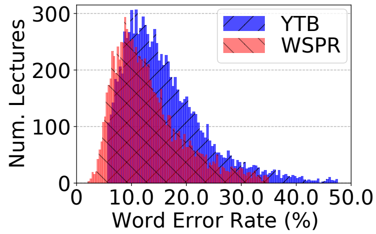


Figure 1: WER distribution for YouTube and Whisper ASRs on the TIE dataset. Both distributions are right-skewed.

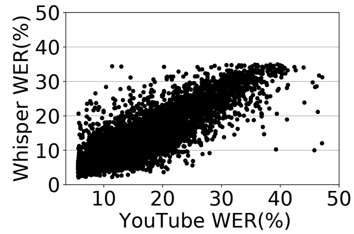


Figure 2: Correlation plot of the WER for YouTube and Whisper ASRs on the TIE dataset. The correlation co-efficient is 0.81.

for which we use the JiWER (Vaessen 2022) Python library to compare the ground truth and ASR generated transcripts.

In addition to being the most popular measure, WER also presents a better idea of what went wrong in the automatic transcription. It does so by taking into account all types of errors, including insertion, deletion, and substitution errors as WER is defined as

$$WER = \frac{S + D + I}{N}$$

where S, D and I are the number of substitutions, deletions and insertions respectively needed to transform the reference text into the hypothesis text and N is the number of words in the reference text. The share of these three components in WER can explain the underlying cause of the error.

In order to further strengthen our observations we also compute two more recently introduced measures BERTScore (Tobin et al. 2022) and SemDist (Kim et al. 2021) that focus on semantic closeness of ground-truth sentences and ASR generated sentences.

Disparity Determination

To measure the disparity between categories within an attribute like gender, experience, etc. we perform statistical significance tests. If the differences between the WERs for the two categories like male and female are statistically significant, we can conclude that there exists a disparity within the ASR between the two categories. From Figure 1, we observe that the WER distributions for both the softwares are right-skewed. We, therefore, use non-parametric statistical test – Kruskal-Wallis (α = 0.001) to determine whether there exist statistically significant differences between the WER medians of the various categories for every attribute. We choose a very strict p-value ($p < 0.001$) to test the hypothesis related to disparities among all attributes to avoid any bias in our testing approach.

In addition, we also compare the median of WER components viz. insertion, deletion and substitution to ascertain the cause of disparity within each category.

Results

We now present the results from our experimental evaluation of the YouTube and Whisper ASRs' transcriptions for the TIE dataset. We first evaluate the overall performance of the two platforms, followed by an in-depth study of the disparities between the different categories for each demographic attribute and speech characteristic mentioned in Table 1.

Overall Disparity

From Figure 1, we see that for the YouTube platform, only 75.6% videos have a WER lower than 20% whereas this ratio is 84.0% for Whisper's ASR transcriptions. This indicates that on average, Whisper reports lower errors in transcription compared to YouTube.

We also evaluate both YouTube and Whisper ASRs using SemDist and BERTScore measures. The median SemDist score (the lower the better) for YouTube was 0.038 and for Whisper was 0.01. Similarly, the BERTScore (the higher the better) for YouTube was 0.843 and for Whisper was 0.891. Thus in both cases the performance of the Whisper ASR is better than the performance of the YouTube ASR which is what we also observe using the WER metric. Thus, the trends being exactly the same, to reduce verbosity we only report WER and its subparts S, D, I in the rest of the analysis in the paper.

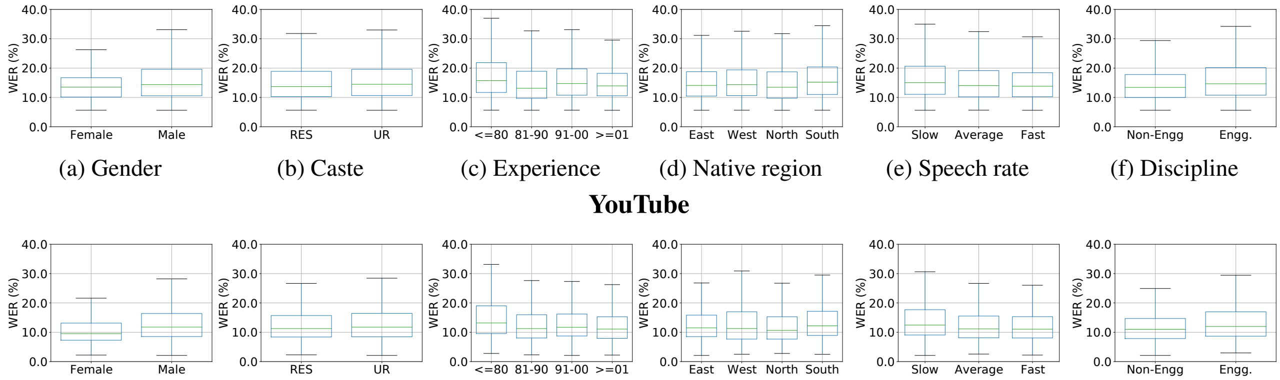
In Figure 2 we analyze the overall correlation between WER exhibited by YouTube and Whisper ASR corresponding to each lecture. We observe a strong positive correlation with a co-efficient of 0.81 indicating that the errors in both ASRs follow the same pattern. The lectures for which the transcription error is high in YouTube ASR-generated transcripts have a high correlation with the lectures for which the Open AI Whisper performs poorly and vice versa. However, there are few exceptions to this trend.

In Table 2 we report the attributes of the top *five* videos with the highest WER for Whisper ASR. We observe that the majority of lectures with the highest WER are from male speakers in the unreserved caste category teaching engineering subjects.

Disparity Within Attributes

We now look at the YouTube and Whisper ASR performances for transcript generation across various demographic and non-demographic attributes.

Gender We compare the WER between male and female speakers in the TIE dataset. Even though the dataset has more than 95% male speakers, we notice that female speakers report lower median WER from Table 3 (difference of 0.9% for YouTube and 2.3% for Whisper) and the Kruskal-Wallis test's null hypothesis of no disparity in WER toward gender attribute can be rejected for both platforms as the p-value corresponding to test statistic is less than 0.001 (refer



(a) Gender (b) Caste (c) Experience (d) Native region (e) Speech rate (f) Discipline

YouTube

(g) Gender (h) Caste (i) Experience (j) Native region (k) Speech rate (l) Discipline

Whisper

Figure 3: WER Disparity across various attributes for the YouTube and Whisper ASRs. The figures from (a) to (f) depict the distribution of WER through box plots for the YouTube ASR, while the figures from (g) to (l) show the distribution of WER through box plots for the Whisper ASR.

Attribute	**Attribute values**				
Gender	M	M	M	M	M
Caste	UR	UR	UR	UR	UR
Exp.	91-00	81-90	81-90	>2001	>2001
NR	East	South	North	East	East
SR	Slow	Fast	Slow	Avg.	Fast
Disc	Engg.	Engg.	Engg.	Engg.	Non-Eng.
YT %	39.5	38.3	35.7	36.1	30.9
WP %	34.9	34.9	34.9	34.8	34.7

Table 2: Description of top 5 lectures with worst WER for Whisper ASR. Values for the attributes of speakers corresponding to these five lectures are mentioned. It can be observed that all the five lectures with the highest WER were delivered by male speakers of unreserved caste category and the lectures were related to engineering disciplines.WER corresponding to these lectures for YouTube ASR is also mentioned for reference. M: male, UR: unreserved.

Table 3). Next, looking at the boxplot distributions in Figures 3a and 3g, we notice that males not only have a higher WER (34.0% and 23.1% resp.), but also a larger interquartile range (IQR) as compared to female speakers, irrespective of the platform. On comparing the two platforms, we see that Whisper reports lower median WERs (Table 3) as well as IQR distributions and max WERs (Figs. 3a and 3g) for both genders, thereby performing better than YouTube for a given gender.

Takeaways: From the results for gender attribute, it can be inferred that –

- WER for female voices is significantly lower (WER median difference – 0.9% on YouTube and 2.3% on Whisper) than that of males, thus indicating a disparity between the transcripts for the two genders on both ASR platforms.
- Whisper performs better than YouTube for both genders, reporting a lower median WER and IQR, but it has a higher disparity between the male and female speakers as compared to YouTube.
- The existing gender disparity in WER corresponding to both YouTube and Whisper is primarily due to substitution error disparity in the ground-truth and ASR-generated transcripts.

Caste On comparing the performance of speakers belonging to reserved castes against unreserved castes, we note for YouTube (see Table 3), a significant median difference (H = 15.2, p< 0.001) of 0.8%, with reserved caste speakers having low WER (median WER = 13.8%) compared to unreserved caste speakers (median WER = 14.6%). For Whisper on the other hand, we note that the difference is not statistically significant (H = 16.4, p> 0.001) and there is no disparity between the reserved and unreserved caste speakers. From Figures 3b and 3h, we see that the distributions are fairly similar but unreserved caste speakers report a slightly higher maximum WER at 33.9% and 29.4% respectively.

Takeaways: On comparing the caste groups, it can be inferred that –

- The Kruskal-Wallis tests shows that YouTube ASR has disparity between the reserved and unreserved caste category speakers, while Whisper does not.
- Both platforms perform relatively better for reserved caste category speakers, reporting lower WER.
- For both the ASRs, the WER dispersion in both groups is similar with difference in IQR for both groups being 0.8%, indicating that the ASRs are agnostic to the caste of speakers.

Experience Here we compare the ASR performance for speakers belonging to different experience groups. From Table 3, we notice that the worst performance is reported for the speakers with the highest experience (WER – 16% for YouTube and 13.3% for Whisper). The best performance for

Category	Attribute	YouTube			Whisper		
		WER (%)	KW	I, D, S (%)	WER (%)	KW	I, D, S (%)
Gender	Female	13.6	H = 23.3, p <0.001	$I: 4.0, D: 1.6, \mathbf{S}: \mathbf{6.4}$	9.6	H = 87.1, p <0.001	$I: 2.9, D: 1.4, \mathbf{S}: \mathbf{4.5}$
	Male	**14.5**		$I: 4.3, D: 2.0, \mathbf{S}: \mathbf{7.2}$	**11.9**		$I: 3.5, D: 2.0, \mathbf{S}: \mathbf{5.5}$
Caste	RES	13.8	H = 15.2, p <0.001	$\mathbf{I}: \mathbf{4.1}, D: 2.0, S: 7.0$	11.4	H = 16.4, p >0.001	$\mathbf{I}: \mathbf{3.4}, D: 1.9, S: 5.4$
	UR	**14.6**		$\mathbf{I}: \mathbf{4.4}, D: 2.0, S: 7.2$	**11.9**		$\mathbf{I}: \mathbf{3.5}, D: 1.9, S: 5.5$
Exp	$\leq$ 1980	**15.9**	H = 125.8, p <0.001	$I: 4.9, D: 2.2, \mathbf{S}: \mathbf{8.0}$	**13.3**	H = 135.8, p <0.001	$I: 3.9, D: 2.1, \mathbf{S}: \mathbf{6.3}$
	1981-90	13.3		$I: 3.8, D: 1.9, \mathbf{S}: \mathbf{6.8}$	11.5		$I: 3.3, D: 1.8, S: 5.4$
	1991-00	14.9		$I: 4.5, D: 1.9, S: 7.2$	11.9		$I: 3.6, D: 2.0, S: 5.5$
	$\geq$ 2001	14.0		$I: 4.2, D: 1.9, S: 6.9$	11.2		$I: 3.2, D: 1.9, \mathbf{S}: \mathbf{5.1}$
NR	North	13.6	H = 80.9, p <0.001	$I: 3.8, D: 1.9, S: 6.9$	10.8	H = 80.3, p <0.001	$I: 3.1, D: 1.7, S: 5.2$
	South	**15.4**		$\mathbf{I}: \mathbf{4.6}, D: 2.1, S: 7.5$	**12.4**		$I: 3.5, D: 2.2, S: 5.7$
	East	14.2		$I: 4.5, D: 1.8, S: 6.8$	11.6		$\mathbf{I}: \mathbf{3.6}, D: 1.8, S: 5.3$
	West	14.4		$\mathbf{I}: \mathbf{3.6}, D: 2.3, S: 7.4$	11.4		$\mathbf{I}: \mathbf{3.0}, D: 1.8, S: 5.4$
SR	Slow	**15.3**	H = 104, p <0.001	$\mathbf{I}: \mathbf{4.9}, D: 1.9, S: 7.4$	**9.3**	H = 160.6, p <0.001	$\mathbf{I}: \mathbf{4.1}, D: 2.0, S: 5.7$
	Avg.	14.2		$I: 4.5, D: 1.9, S: 6.9$	7.8		$I: 3.5, D: 1.8, S: 5.3$
	Fast	13.9		$\mathbf{I}: \mathbf{3.7}, D: 2.1, S: 7.0$	7.4		$\mathbf{I}: \mathbf{2.9}, D: 2.0, S: 5.4$
Disc.	Non-Eng.	13.5	H = 110.8, p <0.001	$\mathbf{I}: \mathbf{3.7}, D: 1.9, S: 6.9$	11.0	H = 128.9, p <0.001	$\mathbf{I}: \mathbf{3.0}, D: 1.9, S: 5.1$
	Engg.	**14.8**		$\mathbf{I}: \mathbf{4.5}, D: 2.0, S: 7.3$	**12.2**		$\mathbf{I}: \mathbf{3.7}, D: 2.0, S: 5.6$

Table 3: WER corresponding to various attributes across categories and Kruskal-Wallis test results for transcripts generated by YouTube and Whisper ASRs. The highest word error rate in each category corresponding to each ASR is highlighted in bold. Except for caste category in Whisper ASR, significant difference in all sub-groups is observed as per the Kruskal-Wallis test results. I, D, S (%) column indicates median of insertion, deletion and substitution component of WER corresponding to each subgroup. Substitution error share in WER is highest for both YouTube and Whisper ASR. The disparity in subgroups in both ASRs is due to different WER components. The WER component with highest disparity corresponding to each category is highlighted in bold. E.g., in gender category for YouTube ASR the difference between median insertion, deletion and substitution error for male and female is 0.3, 0.4 and 0.8 respectively. This indicates that the substitution component is primarily responsible for the observed WER disparity between male and female speakers and the same is highlighted. Exp: Experience, NR:Native Region, SR: Speech Rate, Disc: Discipline

YouTube is reported for speakers who started teaching in 1981-90 (WER = 13.3%) and for Whisper, it was reported by the youngest speakers (WER = 11.2%). Moreover, these differences are statistically significant for both the platforms. Interestingly, the worst WER for Whisper is the same as the best WER for YouTube thus showing the overall superiority of Whisper.

Upon examining the video lectures of highly experienced teachers to uncover the reasons for high WER in their lectures, we discover that these teachers have a tendency to use filler words such as 'uh', 'um', 'ok' and 'alright' and repeat their last said words to bridge gaps between content words, which are part of the manual transcript. In addition, these experienced teachers often use blackboard teaching methods which can introduce noise into the speech signal due to the sound of writing on the blackboard and their movement while speaking.

Takeaways: We now state the takeaways –

- YouTube ASR performs best for speakers who started teaching between 1981-1990 and worst for those who started teaching before 1980. Table 3 highlights that speakers who began teaching before 1980 have the highest substitution error component for both YouTube and Whisper with values of 8.0% and 6.3% respectively.
- Whisper ASR performs best for speakers who started teaching after 2001 and worst for those who started teaching before 1980, primarily due to substitution error.
- YouTube ASR favors the young speakers as the IQR reduces from 10.3% to 7.8% as we move from speakers with the most to the least experience (Figure 3c). The IQR dispersion for Whisper ASR reduces from 9.8% to 7.4% for the same progression (Figure 3i).

Native region We now study the change in the ASR performance based on the speaker's native region which may happen due to the regional variation in the accents. From Table 3, we do not see any disparity between speakers hailing from west and east India in both ASRs, but both the ASRs perform best for speakers from north India, with median WER for YouTube being 13.6% and median WER for Whisper being 10.8%. Figure 3d shows the least IQR and maximum WER for speakers from east India, with the max values for south Indian speakers.

Takeaways: The takeaways for the ASRs performance based on the speakers' native region are –

- Whisper outperforms YouTube for speakers from all four regions.
- Both ASRs perform best for speakers from north India and worst for speakers from south India.
- The primary cause of native regional disparity in WER is insertion error, which is highest between South and West Indian speakers in case of YouTube and East and West in case of Whisper.
- The lowest IQR for 7.6% and 8.5% is reported for speak-

ASR	Lowest WER		Highest WER	
	Engg.	Non-Eng.	Engg.	Non-Eng.
YT	Textile	Agri	Comp. Sci.	Chemistry
	Mining	Basic	Electronics	Maths
WP	Nano-Tech	Basic	Comp. Sci.	Chemistry
	Textile	Agri	Electronics	Atmo Sc.

Table 4: Two disciplines having the highest and lowest WER for each of the disciplines and each of the ASRs.

ers from east India by Whisper and YouTube ASRs respectively.

Speech rate We now study the results for both ASRs for non-demographic attributes. In the TIE dataset, the share of slow, average and fast speech rate category lectures are as follows – 39.1%, 21.4% and 39.5% respectively. In Table 3, we see that interestingly, the slowest speakers have the worst WER (15.3% for YouTube and 9.3% for Whisper) and the fastest speakers have the best WER (13.9% for YouTube and 7.4% for Whsiper). Figures 3e and 3k show a similar change in IQR and it decreases with the increase in speaking speed. **Takeaways**: We have the following takeaways for the two ASR performance for speakers with different speech rates –

- Both ASRs are best in transcribing the audio files for speakers who speak the fastest.
- The performance for both ASRs become more consistent as we move from slow to fast speech rate (Figs. 3e and 3k).
- The highest disparity in WER between lectures with slow speech rate and lectures with fast speech rate can be attributed to the insertion error component of WER.

Discipline We divide the lecture videos into two disciplines – non-engineering and engineering to understand the differences in ASR performances between these two broad disciplines. From Table 3, we observe that transcripts from engineering disciplines have a higher WER (difference of 1.3% for YouTube and 1.2% for Whisper) than those from non-engineering domain. This difference is statistically significant in both ASRs and thus a disparity exists between the two categories. From Figures 3f and 3l, we see that YouTube's WER for engineering lecture videos have a higher IQR and maximum WER, thereby indicating an overall worse performance than Whisper. Table 4 illustrates the two disciplines with the best and worst WER for each ASR. **Takeaways**: We now list the takeaways for comparison between the two disciplines –

- Both ASRs perform better in lectures belonging to non-engineering category compared to the engineering category.
- In both YouTube ASR-generated transcripts and Whisper-ASR generated transcripts, the disparity in transcription accuracy between engineering and non-engineering discipline groups can be attributed to insertion errors.
- From Figs. 3f and 3l, we see that engineering videos have a much higher maximum WER and a larger IQR for both ASRs, with YouTube performing worse than Whisper.

Discussion

In this section, we summarize our findings from our experiments on the two ASR platforms – YouTube Automatic Captions and OpenAI Whisper. We evaluate these platforms on our large-scale dataset– TIE, from the NPTEL MOOC website for word error rate and associated disparities across various demographic attributes of the speakers like gender, caste, experience, native region and non-demographic attributes like speech rate and discipline of the lecture content. We take a two-pronged approach to our analysis – voice related attributes and content related attributes.

Voice Related

We first look at the demographic attributes of the speakers like gender, caste, experience, native region (indicator of accent) and the speech rate that can be correlated to differences in voices of the speakers.

Gender The results for disparity toward gender for YouTube ASR in particular (a difference of 0.9% in median WER between males and females), are surprising as a previous gender specific black-box audit of YouTube Automatic Captions (Tatman and Kasten 2017) did not report any significant differences. In fact, (Tatman 2017) found that the YouTube ASR performs better for White male speakers on an English speakers dataset. Thus, we can see that the accent (non-native Indians vs native White speakers) may be impacting the transcription accuracy and deploying the same model without accounting for regional accents can have an adverse effect on the audience's experience.

The difference in pitch, intensity, tonality along with the enunciation of speech characterize the difference in male and female voices. (Adda-Decker and Lamel 2005) have pointed out better pronunciation and articulation of word utterances in female speech helps the ASR model in extracting correct content information from the speech signal. This may explain the lower WER for female speakers on both YouTube and Whisper ASR. Whisper ASR improves the accuracy for female voices, resulting in a 4% lower WER for female voices and 2.6% lower WER for male voices when compared to YouTube ASR. However, it has been observed that Whisper ASR has a greater variation in performance based on the gender of the speaker.

We also observe that the highest disparity between the two gender groups for substitution error on both ASRs. This confirms that the models are picking up the correct phoneme from female utterances more often than that of male utterances.

These differences could be attributed to the model architecture as well as the training data.

Caste Overall, the differences based on caste are marginal which is also expected. This reinforces the fact that one's caste does not play any role in their tonal attributes or the quality of delivery.

Experience (Vipperla, Renals, and Frankel 2010) have highlighted that the organs involved in speech production mechanism of individuals like lungs, vocal cords and the vocal cavities get affected with age which in turn affects the articulation of words.The speakers with the highest experience are also expected to be the oldest and vice versa. The differences in median WER of the least and most experienced speakers in both ASRs depicts the inability of the ASR model to account for these phonetic variations.Similarly, the highest disparity in substitution error of both ASRs points toward the same direction.

While Whisper reports the lowest WER for speakers with least experience, which is intuitive, interestingly, YouTube ASR reports the lowest value for speakers who joined their institutes between 1981-90. Since the YouTube ASR is a black box model, it is difficult to explain this anomaly.

Native region Our results indicate a disparity between the word error rates for speakers belonging to different parts of India. (Pickering and Wiltshire 2000) have studied the accents of Indian English discourse and have highlighted the difference in frequencies in accented/unaccented syllables of native Tamil (southern region), Bengali (eastern region) and Hindi/Urdu (northern region) speakers. The higher variation in accent of English spoken by the northern and southern speakers is reflected in the median WER difference for YouTube and Whisper– 1.8% and 1.6% respectively.

It should be noted that the high insertion error rate in both of the ASRs for South and East Indian speakers could be due to their accent having certain regional influences. The ASR system may be adding (sub)words that were not originally part of the spoken utterances.

Speech rate The speech rate of native English speakers is understandably higher than that of non-native speakers (Guion et al. 2000). This could be the most plausible explanation for better transcription for lectures having fast speech rate by both ASRs. Whisper ASR has been found to improve the accuracy of speech recognition for lectures delivered at slow, average and fast speech rates by 6%, 5.4% and 6.5% respectively, when compared to YouTube ASR. Nevertheless, it has a greater variation in performance depending on the speech rate in comparison to YouTube.

(Wang and Narayanan 2007) has highlighted the role of speech rate in sentence boundary, disfluency and syllable detection accuracy in speech. Longer syllable duration in lectures which are categorized in slow category might be affecting the correct syllable detection leading to the insertion of extra words in ASRs. The same has been observed in WER error components of both ASRs, where insertion error disparity is highest due change of speech rate. The insertion error rate is lowest for lectures delivered in fast speech rate while those delivered at a slower pace have a higher insertion error rate.

To summarize, we observe statistically significant disparities in both ASRs due to voice related attributes of speakers except for caste where the differences are marginal.From these findings, it can be argued that the disparities observed in both Whisper and YouTube on the TIE dataset are not only due to imbalances in linguistic variations in the training set, but also due to the limitations in the model architecture that does not account for variations in speech signal features. As OpenAI Whisper is an open source model, it can be fine-tuned for the speech variations based on the use case; however the same will not be applicable for YouTube.

Content Related

Next, we look at the content related attribute – discipline to which the lecture videos belong.

Discipline Our findings on the disparity between the WER for engineering and non-engineering lectures points toward the lack of generalized performance on domain specific data sets.It is also observed that there exist disparities within each discipline as well and these may be an outcome of the difference in the number and type of subjects covered in each group. For example, in engineering discipline, average WER varies from 20% for YouTube and 18% for Whisper in Computer Science and Engineering to 12% for YouTube and 10% for Whisper in Nanotechnology. In non-engineering discipline, the average WER varies from 19% for YouTube and 15% for Whisper in Chemistry and Biochemistry to 10% each in YouTube and Whisper in Agriculture.

Our findings suggest that Whisper outperforms YouTube in terms of reducing the WER for various voice and content related attributes. However, Whisper's zero shot transcription feature is observed to exhibit more disparities across certain attributes and groups, compared to YouTube. Despite this, open-source models like Whisper are still a better alternative in comparison to proprietary models in terms of performance, flexibility and addressing disparities.

It is important to note that the model architecture should be designed to account for variations in speech signal features in order to further reduce disparities in the results. This could be achieved by having modern adversarial setups (Kumar et al. 2020) and transfer learning (Vu et al. 2014; Shibano et al. 2021) that have been found to be effective in various problems. The TIE dataset and other datasets (Sanabria et al. 2023; Ahamad, Anand, and Bhargava 2020; Zhao et al. 2018) that include speech samples from non-native speakers of English can be valuable resources for improving the accuracy of open-source ASR models like Whisper. By fine-tuning these models on a diverse range of speech samples including those from speakers with accents, slower speech rates and other potential idiosyncratic attributes, it is possible to reduce the performance gap between different subgroups of speakers. Integrating audio preprocessing techniques such as noise filtering, speaker movement compensation and other similar methods into the ASR pipeline improves the signal quality and can also contribute to enhancing the overall performance of the ASR system. In addition, by incorporating an in-domain vocabulary set, the ASR model can be fine-tuned to recognize and transcribe speech in the target domain more accurately as shown by (Saadany, Orăsan, and Breslin 2022).

Conclusion

In this paper, we explored the issue of disparity in ASR systems towards Indian demography. Here, we have proposed the TIE dataset comprising a set of technical lectures delivered by Indian speakers representing different dimensions of India's socially and linguistically diverse population. We benchmark the performance of YouTube and Whisper for measuring disparity due to speaker characteristics using the word error rate evaluation metric using this dataset. Additionally, we analyzed the insertion, substitution, and deletion components of WER to gain insights into the causes of errors in the two ASR models.

Our findings highlight that there is a strong correlation between WER patterns for both ASRs being audited. Significant disparities exist across gender, speech rate, age group and native region of speakers. We also find that the disparity due to caste is marginal. Both the ASRs seem to have high error rates for lecture videos in the engineering disciplines compared to the non-engineering disciplines.

We also discussed strategies like use of more diverse and representative training data for ASR model training and evaluation. It is important to build customized models that are specifically tailored to the speech variations of underrepresented groups and lexical variations of technical domains to mitigate the existing disparities. Adversarial and transfer learning methods could be very helpful in this context.

Potential Limitations

Our study may have some limitations, like the correctness of manual transcripts can be impacted by human error during transcribing. Similarly, self-annotated labels like gender and caste are not publicly available and hence challenging to determine. Automatic identification of these attributes from names might have some potential errors. The formal speech style in the videos and use of technical jargon may have an impact on the ASR performance. Some of the subjects involve live mathematical derivations and problem solving while others are delivered using slides. These factors may also affect the ASR performance favorably or adversely depending on the lecture topic. The non-deterministic outputs of ASR model components may result is different transcripts of same video. To address these, we experiment over a large dataset and consider a confidence interval of 99.999% for statistical significance.

Broader Perspective, Ethics and Competing Interests

ASR systems have been shown to have bias toward specific sub-population of society due to non-standard accent and dialects. For a linguistically diverse and vast population like India, bias of ASR systems may have serious consequences in providing equitable access to ASR based technologies. In this work, we evaluate disparities in two ASRs – YouTube Automatic Captions and OpenAI Whisper – on a self curated dataset of technical lectures, which is highly relevant from a pedagogical point of view in the Indian context. As online education has been an enabler in democratising education in developing economies like India, errors in ASR generated transcripts can seriously impact both lecture delivery of instructors and learning outcome of students (Willems, Farley, and Campbell 2019).

This audit study and evaluation dataset will enable the ASR system developers in investigating and addressing the biases in these systems which disproportionately affect certain demographic groups. However, if our dataset is used for training any ASR system, it may exhibit data bias due to skewed representation of some of the demographic attributes. Thus we advise practitioners to use caution.

While annotating sensitive demographic attributes of the speakers like gender and caste in our dataset, we specifically label them as perceived gender and perceived caste since self-identified gender and caste information were not available in the public domain. We take due care in anonymising the instructor name and associated institute name corresponding to each lecture to avoid PII based evaluation of WER in future works.

References

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

Ahamad, A.; Anand, A.; and Bhargava, P. 2020. AccentDB: A database of non-native English accents to assist neural speech recognition. *arXiv preprint arXiv:2005.07973*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33: 12449–12460.

Buddayya, R.; LG, N.; et al. 2019. Benefits of videos in YouTube for the undergraduate students in engineering and technology in India. *Webology*, 16(2): 57–71.

Chanana, K. 2006. Gender and disciplinary choices: Women in higher education in India. *Knowledge, power and dissent*, 267.

Depoix, J. 2022. youtube-transcript-api — pypi.org. https://pypi.org/project/youtube-transcript-api/. [Accessed 10-Dec-2022].

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

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

Garnerin, M.; Rossato, S.; and Besacier, L. 2019. Gender representation in French broadcast corpora and its impact on ASR performance. In *AI4TV*, 3–9.

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.

Google. 2009. YouTube Automatic Captions. https://ai.googleblog.com/2009/12/automatic-captioning-in-youtube.html?showComment=1263378133488. Accessed: 2023-01-01.

Guion, S. G.; Flege, J. E.; Liu, S. H.; and Yeni-Komshian, G. H. 2000. Age of learning effects on the duration of sentences produced in a second language. *Applied Psycholinguistics*, 21(2): 205–228.

Hannun, A. 2021. The history of speech recognition to the year 2030. *arXiv preprint arXiv:2108.00084*.

Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hinsvark, A.; Delworth, N.; Del Rio, M.; McNamara, Q.; Dong, J.; Westerman, R.; Huang, M.; Palakapilly, J.; Drexler, J.; Pirkin, I.; Bhandari, N.; and Jette, M. 2021. Accented Speech Recognition: A Survey.

Kendall, T.; and Farrington, C. 2018. The corpus of regional African American language. https://oraal.uoregon.edu/coraal. Accessed: 2020-02-28.

Kent, M.; Ellis, K.; Latter, N.; and Peaty, G. 2018. The case for captioned lectures in Australian higher education. *TechTrends*, 62(2): 158–165.

Kim, S.; Arora, A.; Le, D.; Yeh, C.-F.; Fuegen, C.; Kalinli, O.; and Seltzer, M. L. 2021. Semantic Distance: A New Metric For Asr Performance Analysis Towards Spoken Language Understanding. *arXiv preprint arXiv:2104.02138*.

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. *PNAS*.

Krishnan, M. S. 2009. NPTEL: A programme for free online and open engineering and science education. In *T4E*, 1–5. IEEE.

Kumar, R. S. S.; Nyström, M.; Lambert, J.; Marshall, A.; Goertzel, M.; Comissoneru, A.; Swann, M.; and Xia, S. 2020. Adversarial Machine Learning-Industry Perspectives. In *2020 IEEE SPW*, 69–75.

Medidoddi, V.; Bantupalli, J.; Chakraborty, S.; and Mukherjee, A. 2022. Decoding Demographic un-fairness from Indian Names. In *Social Informatics*, 472–489.

Merriam-Webster. 2021. Caste. https://www.merriam-webster.com/dictionary/caste. Accessed: 2021-09-01.

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.

Ngueajio, M. K.; and Washington, G. 2022. Hey ASR System! Why Aren't You More Inclusive? In *HCII*, 421–440. Springer.

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.

Pickering, L.; and Wiltshire, C. 2000. Pitch accent in Indian-English teaching discourse. *World Englishes*, 173–183.

Prathap, G. 2017. Excellence and diversity mapping of research in IISc, IITs, NUS and NTU. *Current Science*, 1012–1015.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Rickford, J. R. 1999. *Phonological and Grammatical Features of African American Vernacular (AAVE)*. Blackwell Publishers.

Rousseau, A.; Deléglise, P.; and Esteve, Y. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. In *LREC*, 125–129.

Saadany, H.; Orăsan, C.; and Breslin, C. 2022. Better Transcription of UK Supreme Court Hearings. *arXiv preprint arXiv:2211.17094*.

Sanabria, R.; Bogoychev, N.; Markl, N.; Carmantini, A.; Klejch, O.; and Bell, P. 2023. The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR. *arXiv preprint arXiv:2303.18110*.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*.

Shibano, T.; Zhang, X.; Li, M. T.; Cho, H.; Sullivan, P.; and Abdul-Mageed, M. 2021. Speech technology for everyone: Automatic speech recognition for non-native english with transfer learning. *arXiv preprint arXiv:2110.00678*.

Smith, S. 2020. Number of Voice Assistant Devices in Use Worldwide to Reach 8bn by 2024.

Sullivan, P.; Shibano, T.; and Abdul-Mageed, M. 2022. Improving automatic speech recognition for non-native english with transfer learning and language model decoding. *arXiv preprint arXiv:2202.05209*.

Tatman, R. 2017. Gender and dialect bias in YouTube's automatic captions. In *EthNLP*, 53–59.

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.

Tobin, J.; Li, Q.; Venugopalan, S.; Seaver, K.; Cave, R.; and Tomanek, K. 2022. Assessing ASR Model Quality on Disordered Speech using BERTScore. *arXiv preprint arXiv:2209.10591*.

Vaessen, N. 2022. JiWER: Similarity measures for automatic speech recognition evaluation. [Accessed 10-Dec-2022].

Vipperla, R.; Renals, S.; and Frankel, J. 2010. Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 1–10.

Vu, N. T.; Wang, Y.; Klose, M.; Mihaylova, Z.; and Schultz, T. 2014. Improving ASR performance on non-native speech using multilingual and crosslingual information. In *Interspeech*.

Wang, D.; and Narayanan, S. S. 2007. Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8): 2190–2201.

Willems, J.; Farley, H.; and Campbell, C. 2019. The increasing significance of digital equity in higher education: An introduction to the Digital Equity Special Issue. *Australasian Journal of Educational Technology*, 35(6): 1–8.

Zhao, G.; Sonsaat, S.; Silpachai, A.; Lucic, I.; Chukharev-Hudilainen, E.; Levis, J.; and Gutierrez-Osuna, R. 2018. L2-ARCTIC: A non-native English speech corpus. In *Interspeech*, 2783–2787.

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.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Disparity Determination in Methodology & Evaluation Metrics
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Potential Limitations subsection
   (e) Did you describe the limitations of your work? Yes, see Potential Limitations subsection
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Broader Perspective, Ethics and Competing Interests
   (g) Did you discuss any potential misuse of your work? Yes, see Broader Perspective, Ethics and Competing Interests
   (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 Broader Perspective, Ethics and Competing Interests
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes.
   (b) Have you provided justifications for all theoretical results? Yes.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? No.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes.
   (f) Have you related your theoretical results to the existing literature in social science? Yes.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes.
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)? Yes, the dataset is available at GitHub[25]
   (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)? No.
   (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)? Yes.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Potential Limitations.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? Yes, the license information is available at GitHub[25]
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, the ASR generated transcripts and the annotated dataset are available [25]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Broader Perspective, Ethics and Competing Interests.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, the datasheet is hosted on GitHub[25]
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (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? NA
   (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

[25] https://github.com/raianand1991/TIE.