“Allot?” Is “A Lot!” Towards Developing More Generalized Speech Recognition System for Accessible Communication

Grisha Bandodkar[1], Shyam Agarwal[1], Athul Krishna Sughosh[1], Sahilbir Singh[1], Taeyeong Choi [2]

[1]Department of Computer Science, University of California, Davis, CA, USA
[2]Department of Information Technology, Kennesaw State University, Marietta, GA, USA
{gbandodkar, shyagarwal, asughosh, scsingh}@ucdavis.edu, tchoi3@kennesaw.edu *

Abstract

The proliferation of Automatic Speech Recognition (ASR) systems has revolutionized translation and transcription. However, challenges persist in ensuring inclusive communication for non-native English speakers. This study quantifies the gap between accented and native English speech using Wav2Vec 2.0, a state-of-the-art transformer model. Notably, we found that accented speech exhibits significantly higher word error rates of 30-50%, in contrast to native speakers' 2-8% (Baevski et al. 2020). Our exploration extends to leveraging accessible online datasets to highlight the potential of enhancing speech recognition by fine-tuning the Wav2Vec 2.0 model. Through experimentation and analysis, we highlight the challenges with training models on accented speech. By refining models and addressing data quality issues, our work presents a pipeline for future investigations aimed at developing an integrated system capable of effectively engaging with a broader range of individuals with diverse backgrounds. Accurate recognition of accented speech is a pivotal step toward democratizing AI-driven communication products.

Introduction

While Automatic Speech Recognition (ASR) systems have made tremendous progress in achieving high performance for native speakers of English, a considerable gap remains in their performance on recognizing accented English.

Our investigation analyzes multiple datasets including audio samples crawled from YouTube(NPTEL 2022), the Svarah dataset (Javed et al. 2023), and the TTS Indic dataset(Madras 2023). We later go onto the nuances of each dataset in terms of experimentation with the Wav2Vec 2.0 (Baevski et al. 2020) model and analysis. In our experiment, we dissect the complexities of accented speech and quantify the disparities that exist between accented and native English speech.

In the course of our inquiry, we further fine-tuned the Wav2Vec 2.0 model on the accented speech, increasing the adaptability of the ASR system to recognize this particular linguistic difference. Throughout our observations, we discovered instances where divergent pronunciations and the subtle nuances of syllable stress can confound the ASR systems, highlighting the challenges that underpin the representation of accented speech.

In wrapping up our exploration, we showcase potential advancements in speech recognition for non-native speakers. Our insights are backed by quantified results, offering a glimpse into a future where technology is more inclusive. Beyond these achievements, our work paves the way for future research, setting the stage for an integrated system that can connect communication between individuals from diverse linguistic backgrounds.

Related Work

Not a lot of work has been done in specifically improving Indic-accented speech recognition, but there have been quite a lot of studies that recognize the racial bias within speech recognition systems, as well as acknowledge the limited availability of accented datasets. Harwell discusses in their work how people with accents face difficulty in getting accurate responses from home assistants like Google Nest and Amazon Alexa. After testing thousands of voice commands from diverse participants, they share notable disparities in accent understanding. Similarly, Tatman found that based on gender and dialect, there were significant differences in word error rates of YouTube's automatic captions. They showed how women and Scottish speakers had higher word error rates and they discussed the ethical concerns surrounding the same - ”Robust differences in accuracy of automatic speech recognition based on a speaker's social identity is an ethical issue.” Many experiments showed lower accuracies associated with non-white speakers on other platforms as well like BingSpeech (Tatman and Kasten 2017).

As far as ASR systems are concerned, deep neural networks have contributed the most to the successes in this field recently. Elloumi et al. use convolutional neural networks in a multitasking setting and compare their results from learning the features to previous methods that used predefined feature traits. Fohr, Mella, and Illina approached the problem from a more linguistic perspective and suggested enriching semantic tags with specific error tags using multiple SLU architectures and a coder-decoder network. Most recently, verification of utterances is being explored (Lleida and Rose 2000) which is based on the idea that when humans speak to each other, the mode of communication is multimodal

[*] Equal contribution by first four authors.

Parameter	Setting
Metric For Best Model	WER
Weight Decay	0.0354792
Eval Steps	50
Warm Up Steps	150
Learning Rate	3e-5
Batch Size	1
Gradient Accumulation Steps	32
Number of Epochs	30

Table 1: Fine-tuning hyperparameters

which also involves visual information like lip movements which is especially useful in noisy environments. Our work most closely aligns with some works that have fine-tuned Wav2Vec 2.0 on Swedish audio samples and presenting a detailed analysis of their results in comparison with Google Speech-to-Text (Dabiri 2023).

There have been a lot of other studies on ASR systems as well but most of them rely on a huge amount of data, which is very hard to obtain for non-native speaker settings. Project Vaani is an initiative supported by Google which aims at creating a corpora of over 150,000 hours of speech which is transcribed in local scripts. They are particularly focused on ensuring that the dataset respects age, gender, linguistic, educational, and urban-rural diversity. Due to lack of access to huge amounts of data, our approach is rather simple since each of our datasets is less than 4 hours worth of sound samples. However, we do discuss some more future directions in the discussions later that would be possible with access to moderate-sized datasets.

Motivation

ASR systems have significantly shaped human-to-human and human-to-computer connectivity over the past few decades. Increased recognition accuracy has not only helped tear down language barriers in conversations but has also given way to live transcriptions of audio. A simple “Hey Google” allows users to access the internet, send messages to their contacts, or turn off their bedroom lights. ASR systems have, overall, enhanced communication accessibility, particularly for individuals hard of hearing or with difficulties moving around.

Though India is the second largest English-speaking country, Indic-accented data performs poorly against popular benchmarks like LibriSpeech, Switchboard, and Speech Accent Archive. This disparity is demonstrated in a lower WER when using state-of-the-art ASR systems on Indic-accented speech.

Spoken language itself is fallible and open to subjectivity regardless of variations in accents, dialects, speech impediments, background noise, and the very tone of the audio. All these factors and minor misinterpretations of speech contribute to ASR inaccuracies.

Understanding the aforementioned limitations is crucial when considering the applications of ASR and its use for speakers of foreign accents. A seemingly small transcription error can have significant consequences in communicating the speaker’s intent. Thus, the challenge persists in the equitable accessibility of these systems to all individuals, particularly those who, by virtue of their linguistic and cultural backgrounds, have alterations in their communication, i.e., accented speech.

Parameter	Setting
Mask Time Probability	0.75
Mask Time Length	10
Loss Function	CTC
Gradient Checkpointing	True
FP16	True

Table 2: Fine-tuning hyperparameters for Wav2Vec 2.0

Stage	Male WER	Female WER
Pre-trained	0.331	0.303
Fine-tuned	0.213	0.222

Table 3: WER comparison across gender within Svarah dataset

Data Exploration

Dataset Selection

The first dataset contains audio samples crawled from YouTube videos from the NPTEL channel (NPTEL 2022), containing English videos with a South-Asian accent within the General & Technical Education domain (Kumar, D, and G 2021). These videos all had subtitles generated with Google ASR that were later manually updated. Within the dataset, around 19,500 videos were crawled with an average video duration of 40 minutes. Each audio chunk length is around 3-10 seconds, but for the sake of computation, we used audios that are 7 seconds in length or less. We chose a random subset of samples with a total length of 7819.56 seconds.

The second dataset is the Svarah dataset (Javed et al. 2023). This dataset is more inclusive, covering conversational speech across several domains such as history, culture, tourism, government, sports, daily activities, and more. Data from the Svarah dataset is composed of 9.6 hours of transcribed English from 117 speakers across 17 districts. This varied vocabulary and usage scenarios contribute to a more comprehensive assessment of ASR systems in real-world applications. We chose a random subset of samples with a total length of 12810.37 seconds.

Our third dataset is the TTS Indic Dataset by IIT Madras (Madras 2023). This project was funded by the Deity, M CIT, Government of India, and aims to improve the quality of Indic TTS synthesis systems rather than ASR, so the voices within this dataset are very clean and less conversational. In this dataset, there are 6,765 audio files with both male and female speakers narrating a story. We chose a random subset of samples with a total length of 5,536.26 seconds.

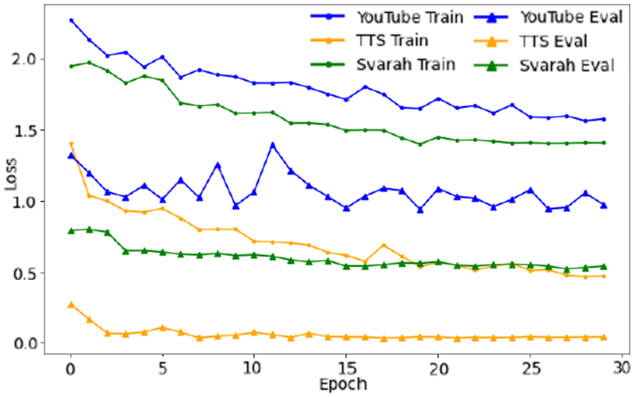


Figure 1: Fine-tuned loss graphs for all three datasets

ASR Model Selection

Wav2Vec 2.0 Wav2Vec 2.0 has emerged as a leading framework for self-supervised speech representation learning, achieving state-of-the-art results on automatic speech recognition benchmarks. Developed by Facebook AI Research, Wav2Vec 2.0 utilizes a convolutional neural network to encode raw speech audio into latent representations during unsupervised pretraining. This adapts the model to predict text transcriptions directly from raw speech audio as input, using connectionist temporal classification (CTC) loss during training.

Facebook AI team initially fine-tuned the Wav2Vec 2.0 model on 960 hours of labeled speech data from LibriSpeech (Baevski et al. 2020) which contains English speech samples from audiobooks. This large dataset teaches the model to produce effective transcriptions for English speech, which can then be applied to downstream tasks like speech recognition for Indic accents. The goal is to leverage the representations learned on LibriSpeech to improve performance on Indic accented speech compared to training on Indic data alone.

Table 2 outlines key hyperparameters used during fine-tuning for all three datasets: Mask Time Probability, Mask Time Length, and Loss Function. We selected default hyperparameters mainly because our experimentation focus is not on the parameters of fine-tuning.

Dataset Curation Criteria

Our dataset curation process included specific criteria. To maintain a manageable and consistent dataset size, all audio clips of each dataset were limited to durations of 4 seconds, except for the YouTube dataset with a maximum audio duration of 7 seconds as there were not many audio samples under 4 seconds. Furthermore, audio segments that contained any numbers in digit form were excluded altogether to ensure a homogenous dataset in terms of content and facilitate a more focused analysis.

A deliberate choice was made to give less precedence to the YouTube dataset. This is because some audio features posed as negative reinforcements since some audio clips had incorrect transcriptions that deviated entirely from the speech in the audio. Another nuance resided in the TTS dataset since each audio was very clean, and each speaker spoke slowly with emphasis on enunciation. Thus we place priority on the findings of the Svarah datasets since they contained more conversational speech and were a perfect middle ground between the other two.

Word Error Rate (WER) Findings

The Word Error Rate (WER) proved to be a central, grounded metric for evaluating transcription accuracy throughout our research. WER is defined as:

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

S: number of substitution errors.

D: number of deletion errors.

I: number of insertion errors.

N: total number of words in the reference string

On the YouTube dataset, Wav2Vec 2.0 achieved a WER of 51.64%. On the Svarah dataset, Wav2Vec 2.0 achieved a WER of 34.31%. On the TTS Indic dataset, Wav2Vec 2.0 achieved a WER of 4.60%.

With a standard in place and substantial space for improvement in Indic-accented speech recognition, the goal became to detail the underlying problem and discover methods for refinement.

Examples of Mispronounced Words and Sentences

To underscore the depth of inaccuracy within the three datasets, we provide further detail of words and sentences our ASR systems failed to transcribe appropriately. To the best of our knowledge, this is the first attempt at categorizing common errors made in transcribing accented speech:

Word Pronunciation Variation

- The word “schedule” was often pronounced as “sched-yool” instead of the American standard “sked-yool”.
- “Herb” was transcribed as “erb” without the “h”, in line with British English pronunciations.

Syllable Stress Nuance

- The phrase “television program” places stress in different syllables among Indic and American accented English with the former using “TELE-vision PRO-gram” and the latter “tele-VISION program”.
- Similarly, the word “tomorrow” was broken into “to mor-row”.

Phonetic/Cultural Variation

- The name “Raj” was transcribed as “Rodge”.

Divergent Pronunciation

- The sentence “Turn off the lights” was transcribed as “turn awff the leets.”
- “In turn” was transcribed as “intern.”

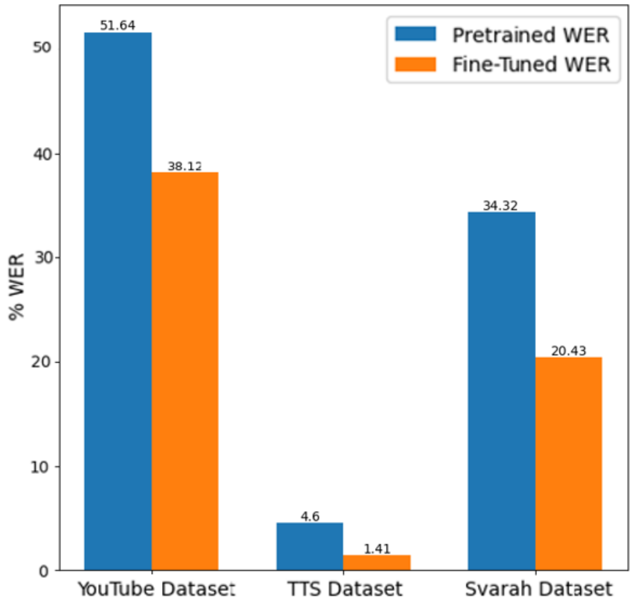


Figure 2: WER Comparison for All Datasets

Moving forward, our aim is redirected to exploring solutions to improve accuracy. However, it is important to note that our fine-tuning results in the model learning the inaccuracies of the transcribed speech. This is a problem we acknowledge, and to address this problem we need a larger database of accented language to accurately train ASR models.

In the subsequent sections of the paper, we delve into the details of our findings, offer further insights, and provide the results from our introduced methodologies.

Experimental Fine Tuning

Fine-Tuning

Fine-tuning, a popular machine learning technique, adapts pre-trained models to new tasks without the need for extensive data while also maintaining high performance. When fine-tuning models like Wav2Vec 2.0, it is important to prevent forgetting the original knowledge gained from its pretraining. This is achieved by utilizing a small learning rate during the fine-tuning stage, to not drastically modify the model. Leveraging this concept, we aim to fine-tune Wav2Vec 2.0, initially trained on native English speech, to enhance Word Error Rates (WER) on accented Indic speech. Our approach is inspired by prior efforts to improve WER for the Swedish accent (Dabiri 2023) as mentioned before. Rather than training an entire model from scratch, we will adapt the preexisting representations learned by Wav2Vec 2.0 to better recognize Indic accent in a specialized model.

For the fine-tuning process, our training parameters are outlined in Table 1. Hyperparameters were chosen with specific goals in mind, namely: computing efficiency (small batch size), model generalization (mask times), and minimizing knowledge forgetting (low learning rate). We particularly chose a batch size of 1 because we were constrained by our computational resources, otherwise, it is important to note that a default batch size of 32 would have been used. The rest are standard recommended parameter values.

Augmenting factors	Male WER	Female WER
A, B, C, D	0.523	0.568
A, D	0.353	0.386

Table 4: Data augmentation test results

Fine-Tuning all Three Datasets

We perform our experiments on all three datasets. The results can be found in Fig. 1, where we show the training as well as testing losses. It is important to note that the three differed in the numbers and sizes of audio samples and in the quality of their samples. Thus, the pre-trained and fine-tuned WER vary a lot (as seen in Fig. 2). We provide a more comprehensive analysis of our results in the upcoming sections.

Analyzing Svarah Dataset

Since we established that the other two datasets were not a perfect representation of the Indic-accented speech, we wanted to look at the Svarah dataset under a closer lens. There is a study (Clopper, Conrey, and Pisoni 2005) that suggests due to phonological differences, women are often thought to produce clearer speech with greater modulation of pitch than men. As a result of this, ASR systems seemingly tend to transcribe female audio with higher accuracy, so we ran an experiment to analyze the WER across genders.

Figure 3 shows the performance metrics for baseline fine tuning across gender.

Experimental Fine Tuning with Data Augmentation

Accented speech exhibits significant variability in pronunciation, tone, and rhythm when compared to native speech. This inherent variability poses challenges for state-of-the-art ASR systems, resulting in reduced performance. In light of this issue, data augmentation emerges as a valuable technique for the following reasons (Huh, Ray, and Karnei 2023):

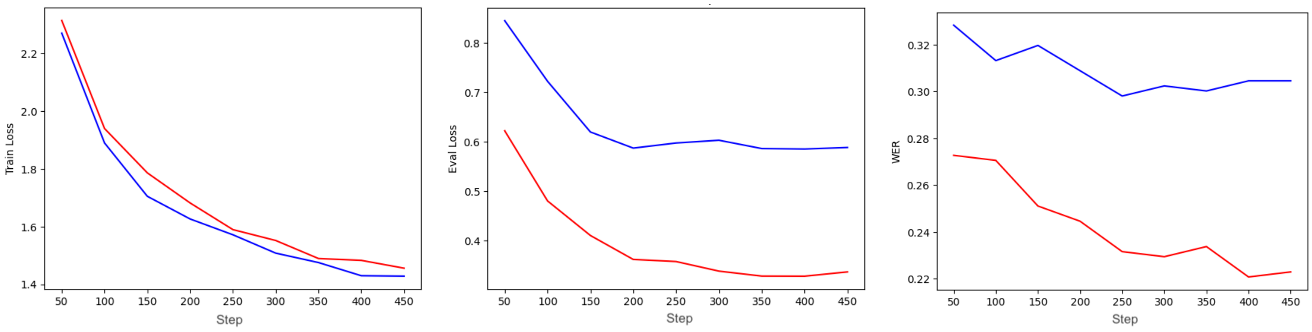
I) Generalization: By furnishing the model with augmented data examples, it improves its ability to generalize, enabling it to comprehend spoken words and phrases that it has not encountered explicitly before.

II) Robustness: Augmenting data also helps make the model more robust by exposing it to data instances that differ in just one particular attribute, such as speaking rate or pitch, for example.

We perform four different augmentations and observe the word error rates for different permutations and combinations of them, the major ones of which are mentioned in Table 4. The various augmentations are:

A) Adding Gaussian noise to the audio signal.

B) Altering audio duration by compressing or stretching it.



(a) (b) (c)

Figure 3: Performance metrics for baseline fine tuning across gender (blue = male; red = female): (a) Train loss, (b) Evaluation loss, and (c) Train WER.

C) Shifting audio in time (moved waveform along time axis).
D) Changing the frequency of the audio.

The details about training loss, evaluation loss, and training word error rates for both the permutations of augmenting factors are present in Fig. 4 and 5.

It is worth noting that after excluding the audio alterations and time shift data augmentation strategies, the Word Error Rate demonstrated a closer resemblance to the original baseline fine-tuning WER. This is largely because when the audio was altered and shifted, pieces of the audio clip were shuffled around, causing the misalignment between the altered speech and the ground-truth transcription. Consequently, this misalignment contributed to an increased WER value when transcribing the rearranged speech segments.

A persistent issue is the quality of the data that we started with, but there are two additional factors:
1) Presence of a small dataset: Since it is very difficult to get access to a large number of instances of transcribed accented speech, data augmentation on this small dataset might end up not resolving the overfitting of the model. This is because the total number of samples after augmentation is still relatively small (we did not have enough augmentation strategies to increase the number of unique samples passed to the model by a huge amount).
2) Strength of augmentation: It is very non-intuitive and difficult to come up with the right set of parameters while augmenting the dataset such that the quality of the original data is not compromised.

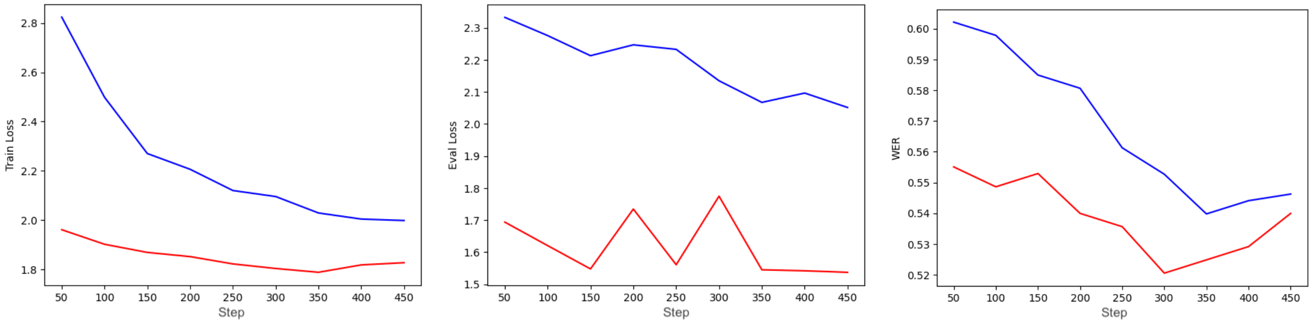
Results

1. *Performance of ASR Models across Datasets:* Figure 2 shows that when all three datasets were tested against the pre-trained ASR Model, only the TTS Indic dataset achieved a WER that was comparable to Wav2Vec 2.0's original WER metrics. We reason this is due to the fact that the dataset consisted of synthetic voices that spoke clearly and were not conversational - hence not an accurate representation of the accent. The other two datasets achieved relatively higher WER rates which indicates the poor performance of the Wav2Vec 2.0 model on real data.

2. *Effectiveness of Fine-Tuning across Datasets:* According to Fig. 2, all three datasets showed significant improvement in WER after fine-tuning the model. This shows that this method is empirically beneficial for adapting ASR systems to be more accessible across all distinct accents. However, this process is very inefficient, as the model had to be trained for one specific accent and required an extensive dataset containing conversational data. Ultimately, this method becomes too costly to implement, considering that India is the second-largest English-speaking country, and we had difficulty obtaining an acceptable dataset.

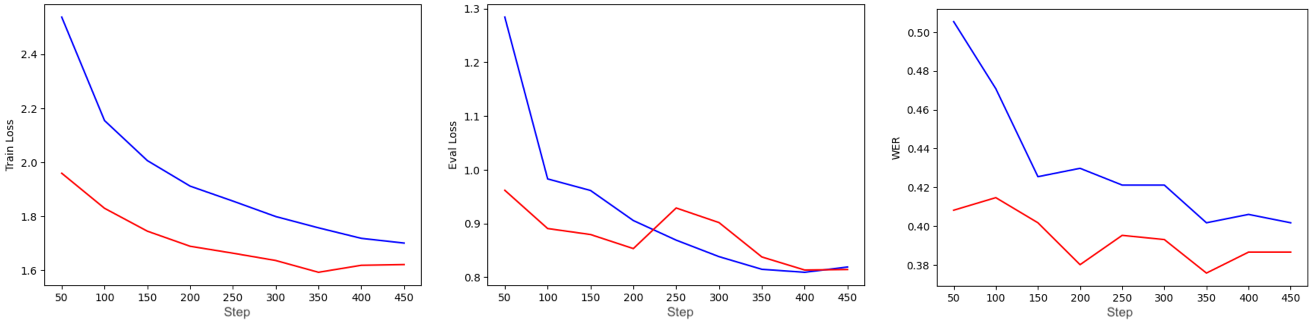
3. *Performance and Effectiveness of ASR Models across Gender (Svarah Exclusive):* Table 3 presents the pretrained WER alongside the fine-tuned WER across gender, specifically for the Svarah dataset. It is interesting to note that there was significantly more improvement in the male WER than the female WER, but this may be specific to the dataset that we used. There is a smaller scope of improvement from fine-tuning accented speech because vocal attributes of women from both native and non-native origins are relatively closer as compared to men. The details about training loss, evaluation loss, and training word error rates for baseline fine-tuning are present in Fig. 3. As for effectiveness, fine-tuning the model improved the WER for both genders, which corroborates our initial statement that fine-tuning the dataset overall is empirically effective.

4. *Data Augmentation for Accent Variation:* To enhance the model's recognition capabilities, we applied various data augmentation techniques, specifically the previously mentioned A, B, C, and D factors. The results of the two combinations are presented in Table 4. We chose these two combinations as we saw tangible performance. In the future, we plan to delve deeper into why our data augmentation results were considerably worse than before since theory contradicts this. We will explore the different combinations, especially with Gaussian noise and masking. This is detailed in a different study which is a direction we want to pursue (Klumpp et al. 2023).



(a) (b) (c)

Figure 4: Performance metrics for augmented fine tuning with factors A, B, C, and D across gender (blue = male; red = female): (a) Train loss, (b) Evaluation loss, and (c) Train WER.



(a) (b) (c)

Figure 5: Performance metrics for augmented fine tuning with factors A and D across gender (blue = male; red = female): (a) Train loss, (b) Evaluation loss, and (c) Train WER.

Though data augmentation is a valuable technique for many other tasks, it did not improve the accuracy of accented speech in our specific context. On the other hand, fine-tuning improved WERs for both female and male speech, highlighting a route to explore further.

Discussions and Limitations

Dataset Curation

It is important to note that it is difficult to find a relatively robust dataset. All of the three we found had their issues. The YouTube crawled dataset often consisted of incorrect transcriptions as well as noisy audio. The TTS Indic dataset was generated out of synthetic speech, so it was not representative of a natural Indic accented speech. The Svarah Dataset, which we concluded as the best out of the three, still had issues with audio quality and limited size, due to which we were not able to utilize the benefits of data augmentation.

In the future, we would like to build a more informative dataset and utilize different learning methods.

Next Steps and Future Directions

While our research addresses the challenges of Indic ASR, there are several promising avenues for further exploration and improvement in this domain that we intend to pursue.

1. *Active Learning:* Active learning is a machine learning paradigm that focuses on the selection of the most informative data points for model training. In the context of speech recognition, active learning can help identify audio samples that are particularly challenging or where the model is likely to make errors. By selecting, labeling, and fine-tuning these challenging samples, we can enhance the model's performance on accented speech. To implement active learning effectively, we can develop strategies to identify problematic audio segments, such as those with high WER during initial model predictions. These samples can then be reviewed for model improvement. There has been work done to apply active learning to ASR models, but it has not yet been used for accented language (Huang et al. 2016)

2. *Style Transfer for Accent Variation:* Style transfer techniques (Gatys, Ecker, and Bethge 2015) can be applied to audio data to simulate variations in accents, pronunciation patterns, or speaking styles/rates. By transforming audio samples to mimic different accents or speaking styles commonly found in Indic-accented speech, we can augment the training dataset in a novel way. This ap-

proach allows the model to adapt to a broader range of accented speech patterns. We can explore various style transfer methods, including neural style transfer, to generate augmented data and fine-tune the model on this data to improve its ability to recognize accented speech. This method is starting to be implemented in general across audio, but not yet to a particular dialect (Verma and Smith III 2018)

3. *Custom Dataset Generation:* To further refine the model's performance on specific accented speech patterns and problematic words, we can create a custom dataset. This dataset can be generated based on identifying the words or phrases that are most commonly transcribed incorrectly by the ASR model. We can curate audio samples containing these challenging words, phrases, or accent-specific nuances. These samples can be recorded by speakers with foreign accents or generated using a text-to-speech model. By fine-tuning the model on this custom dataset, we can target specific weaknesses in accented speech recognition and improve accuracy.
4. *Exploring Gender-Based Datasets:* To better understand the impact of gender on ASR, we can explore and compare male and female datasets separately. Gender-related variations in speech, including pitch, tone, and speaking style, can significantly influence recognition accuracy. By analyzing and fine-tuning the model on gender-specific datasets, we can tailor our ASR system to be more inclusive and accurate for both male and female speakers with diverse accents.

Technological Limitations

We had a few computational challenges that guided the way we conducted our experiments. The most significant of these was lack of access to a high-quality GPU due to which we had to restrict the batch size to 1 in order to avoid any memory issues arising from the free Google Colab version. Additionally, we opted for shorter audio lengths to enhance computational time efficiency and optimize runtime memory usage.

Ethics and Inclusivity

In examining the ethical dimensions and inclusivity in our research, it is vital to address the significance of data collection to promote equitable/inclusive advancements in ASR technology. The findings of our study, which specifically focuses on Indic-accented English, highlight the shortcomings of widely used pre-trained models and the need for diverse and representative datasets to ensure the development of inclusive ASR systems. The disparities revealed in error rates underscore the ethical obligation to bridge communication gaps for non-native speakers.

Our work stresses the necessity of addressing linguistic and cultural diversity in dataset curation during the training and development of ASR models. The utilization of video-crawled data is ethically justified as it provides a more realistic representation of real-world data without revealing any personally identifiable information about the speakers. Datasets generated using TTS models, like TTS Indic Dataset, are neither natural nor authentic. This approach ensures that the models are exposed to varied voices and audio qualities, contributing to their robustness and generalization avoiding any possible privacy concerns.

In short, the use of publicly available, random voices aligns with ethical considerations and allows a more realistic depiction of linguistic diversity for all our datasets.

Conclusion

This research identifies and quantifies significant challenges in ASR systems when handling Indic-accented English. Accented speech records notably higher error rates of 51.64% using Wav2Vec 2.0 on the YouTube dataset, 34.31% on the Svarah dataset, and 4.6% on the TTS Indic Dataset, compared to 8.2% on noisy speech and 5.2% on clean speech on the standard LibriSpeech benchmark (which is non-accented). Our study identifies possible categories causing erroneous transcription, like pronunciation variations, syllable stress nuances, cultural differences, and divergent pronunciation. The fine-tuning of the Wav2Vec 2.0 model results in a substantial improvement, with all three datasets showing positive net improvement in WER.

Dataset	Pre-Trained	Fine-Tuned	ΔWER
YouTube	51.64%	38.12%	13.52%
Svarah	34.31%	20.42%	13.89%
TTS Indic	4.59%	1.41%	3.18%

Table 5: WER improvement across datasets

Through data augmentation, we were able to identify the limitations caused by the poor quality of available data. Our paper also proposes future directions, including active learning and style transfer methods, custom dataset creation, and exploring gender-based datasets for improved inclusivity. Overall, our research provides a roadmap to develop more accurate and inclusive ASR systems, aiming to make AI-driven communication more accessible for individuals with diverse linguistic backgrounds. We hope that this paper will motivate further research in this field.

References

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477*.

Clopper, C. G.; Conrey, B.; and Pisoni, D. B. 2005. Effects of Talker Gender on Dialect Categorization.

Dabiri, A. 2023. *Improving accuracy of speech recognition for low resource accents: Testing the performance of fine-tuned Wav2vec2 models on accented Swedish.* Ph.D. thesis, KTH Royal Institute of Technology.

Elloumi, Z.; Lecouteux, B.; Galibert, O.; and Besacier, L. 2020. Prédiction de performance des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs. *TAL. Volume 59 - n◦ 2/2018*, 49–76. Published in French.

Fohr, D.; Mella, O.; and Illina, I. 2017. New Paradigm in Speech Recognition: Deep Neural Networks. In *IEEE International Conference on Information Systems and Economic Intelligence*. Marrakech, Morocco. HAL Id: hal-01484447.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A Neural Algorithm of Artistic Style. *arXiv:1508.06576 [cs.CV]*. (or arXiv:1508.06576v2 [cs.CV] for this version).

Harwell, D. 2018. Alexa Does Not Understand Your Accent. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

Huang, J.; Child, R.; Rao, V.; Liu, H.; Satheesh, S.; and Coates, A. 2016. Active Learning for Speech Recognition: the Power of Gradients. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain. ArXiv:1612.03226v1 [cs.CL] 10 Dec 2016.

Huh, M.; Ray, R.; and Karnei, C. 2023. A Comparison of Speech Data Augmentation Methods Using S3PRL Toolkit. arXiv:2303.00510.

Javed, T.; Joshi, S.; Nagarajan, V.; Sundaresan, S.; Nawale, J.; Raman, A.; Bhogale, K.; Kumar, P.; and Khapra, M. M. 2023. Svarah: Evaluating English ASR Systems on Indian Accents. arXiv:2305.15760.

Klumpp, P.; Chitkara, P.; Sarı, L.; Serai, P.; Wu, J.; Veliche, I.-E.; Huang, R.; and He, Q. 2023. Synthetic Cross-accent Data Augmentation for Automatic Speech Recognition. *arXiv:2303.00802 [cs.CL]*. ArXiv:2303.00802v1 [cs.CL] for this version.

Kumar, L. A.; D, K.; and G, R. 2021. Automatic Speech Recognition for Indian Accent Lectures contents using End-to-End Speech Recognition model. AI4Bharat: NPTEL2020 - Indian English Speech Dataset.

Lleida, E.; and Rose, R. 2000. Utterance verification in continuous speech recognition: Decoding and training procedures. *Speech and Audio Processing, IEEE Transactions on*, 8: 126 – 139.

Madras, I. 2023. IIT Madras. https://www.iitm.ac.in/. Accessed: December 18, 2023.

NPTEL. 2022. NPTEL - Online Learning Initiatives by IITs and IISc. Distributed under Creative Commons Attribution-ShareAlike (CC BY - NC - SA).

Tatman, R. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, 53–59. Valencia, Spain: Association for Computational Linguistics.

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

Verma, P.; and Smith III, J. O. 2018. Neural Style Transfer for Audio Spectrograms. *arXiv:1801.01589 [cs.SD]*. (or arXiv:1801.01589v1 [cs.SD] for this version).