IJISAE

International Journal of

INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Fine-tuning ASR Model Performance on Indian Regional Accents for Accurate Chemical Term Prediction in Audio

Dr. Sonali Kothari*[1], Dr. Shwetambari Chiwhane[1], Rithwik Satya[1], Md. Asad Ansari[1], Shreeja Mehta[1], Pranav Naranatt[1,] Dr. M. Karthikeyan[2]

Submitted: 08/05/2023 **Revised**: 14/07/2023 **Accepted**: 05/08/2023

Abstract: Automatic Speech Recognition (ASR) models have recently become famous for their incredible ability to provide highly accurate transcriptions of human speech. They have been in the radius of further research and development. The study compared three state-of-the-art ASR models: Deep Speech, Wav2Vec, and Whisper. The proposed research has evaluated their performance on a dataset of audio recordings containing chemical terms spoken in various Indian regional accents. This research aims to precisely identify a model with the best accuracy of transcribing chemical terms spoken in Indian regional accents and fine-tune it further for efficient prediction.

Keywords: *Automatic Speech Recognition, ASR models, Chemical term identification in Indian regional accents, Deep Speech, Fine-tuning ASR, Wav2Vec, Whisper, Performance evaluation*

1. Introduction

Automatic Speech Recognition (ASR) technology has increased dramatically across various applications, including the transcription of speech into text, virtual assistants, and machines that voice commands can control. ASR systems still have difficulties accurately transcribing lectures with various English accents, particularly regional accents. This issue mainly affects applications in the chemical industry, for example, that need precise recognition of domain-specific terms [1][2][3][4][5].

This study compared three state-of-the-art ASR models - DeepSpeech, Wav2Vec2, and Whisper - to identify the most effective model for accurately transcribing chemical terms spoken in various Indian regional accents of English. The study aims to assess the performance of these models on domain-specific datasets and evaluate the impact of fine-tuning the best model on the specific accents present in the dataset.

The Indian subcontinent is quite famous for its diverse languages, with over 300 languages, of which 22 are officially recognized, with many dialects branching out of them. These regional variations can make it challenging for ASR models to transcribe English accurately accented speech, mainly when the address includes domain-specific terminology like chemical terms. By comparing the performance of these three ASR models, the proposed study aimed to provide insights into the most effective approach to transcribing chemical terms spoken in Indian regional accents.

This study has important implications for developing ASR models for domain-specific applications, especially in diverse linguistic settings. Accurate speech transcription is crucial for various industries, including the chemical industry, and can help improve efficiency and productivity in multiple sectors. The findings can help inform the development of more effective ASR models for accurately transcribing speech in diverse linguistic settings.

The structure of this research paper is outlined in the following manner. Section 2 presents a literature review of different documents focusing on the challenges and opportunities present in Indian-based regional accents. Section 3 offers the approach to model selection, outlining the dataset used and comparative analysis along with evaluation results of a few selected ASR models. Section 4 shows the process of fine-tuning the best-performing ASR model chosen from the previous section's results. Section 5 presents the experimental results, including analyzing the fine-tuned ASR model's performance and transcription accuracy. Section 6 discusses the research implications and suggests future directions for further research. Finally, in Section 7, concluding remarks, a summary of findings and contributions to the field is highlighted.

2. Literature Review

Wav2vec2[9] shows excellent potential when trained using unlabelled data for speech processing, and in [9], only 10 minutes of audio recorded from different sources achieved a word error rate of 4.8 on a cleaned test version of LibriSpeech. The work showcases the model's immense potential and future scope if large datasets are used or created for a particular application.

[1] *Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India*

[2]*NCL-CSIR, Baner, Pune*

ORCID ID : 0000-0002-3797-9932 ORCID ID : 0000-0002-3534-9654

ORCID ID : 0009-0008-7604-0080 ORCID ID : 0009-0007-9948-148X

ORCID ID : 0009-0003-2391-6003 ORCID ID : 0009-0004-3739-6456

** Corresponding Author Email: sonali.kothari@sitpune.edu.in*

In [10], it is shown that more than a validation or test set created by using public datasets is required to perform ASR on other public datasets or in the real world. This exemplifies the need to develop datasets suited for applications or utilize real-world audio data to test any speech-to-text model.

Wav2vec-U-based toolkits such as EURO [11] have also been developed for unsupervised automatic speech recognition research. Experiments have been done on such systems by using datasets such as TIMIT and LibriSpeech to showcase the ability of the system to perform on similar levels compared to fairseq Wav2vec-U. It is also shown to outperform Hubert or WavLM.

Hidden Markov-based systems such as SPHINX [12] can provide a large-vocabulary continuous speech recognition system independent of the speaker. The WER was significantly reduced by as much as 85 percent when such methods were used.

Table 1. Findings of some of the critical literature

Title	Merits	Demerits	Summary
Speaker-independent ASR for modern standard Arabic: effect of regional accents [1]	-The ALGASD database, which consists of data from more than 300 Algerian speakers -The ASR performance independent of the speaker in the first experimentation phase is 91.7 %. The accuracy for the same was found to be 90.6 %	An inconsistency was found for the recognition rates of northern and southern region accents	Adaptive techniques, such as Maximum posterior (MAP) or Maximum Likelihood Linear Regression (MLLR), can be used to identify variations in the speaker's accent.
Learning fast adaptation on cross-accented speech recognition [2]	- A fast adaptation method has been used in the form of model-agnostic meta-learning (MAML) methodology to adapt to recognize unseen accents quickly. - As more data is added into the fine-tuning part, the WER rate drops constantly for both the methodologies used in the paper	ASR models find it difficult to adapt to accents that are unseen and have distinct pronunciations or tones compared to the ones that were used for training the model	In general, as the amount of data used for fine-tuning the models increases, the WER rate drops constantly.
Automatic speech recognition of multiple accented English data[3]	-A multi-accented English broadcast news corpus from different geographic regions is trained on accent-independent acoustic model training.	A considerable degradation in performance was found when only a single accent was used to train the system, and data from other regions was tested on the model.	There is scope for future work in the project by examining the use of pronunciation and Language Model adaptation, combined with various approaches for acoustic model adaptation to obtain a robust performance across English Broadcast news sources
Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech[4]	Using residual adapter layers to adapt ASR models can bring about significant reductions in word error rates (WER) compared to unadapted models in two different tasks: everyday speech and accented speech, and using two different architectures (RNN-T and Transformer Transducers).	A direct comparison between residual adapters and methods that use statically fed speaker-dependent vectors would have been helpful in speaker adaptation for ASR models. Furthermore, identifying encoder layers that can benefit from adapters is crucial for enhancing the performance of residual adapter-based ASR models. These steps can	- Unlike model fine-tuning, residual adapters are much more parameter efficient, and fine-tuning the entire encoder for each speaker or accent leads to only minor improvements.

		help improve the accuracy of ASR models in speaker-dependent scenarios.	
Accented speech recognition: A survey[5]	In accented ASR, this method takes a baseline model that performs well on an initial set of accents and improves its performance on a new accent while maintaining performance on the original set	Significant challenges, such as data sparsity and a lack of a standard benchmark	It is unclear if accent-tuning approaches appropriate for English apply to other languages, especially in languages with more significant dialect variation, such as Hindi, Chinese, or Arabic
Robust speech recognition via large-scale weak supervision[6]	By increasing the size of the dataset, there is a significant improvement in the performance of unsupervised systems, even if they are trained with a limited amount of labels	There is much scope for improvements, such as using bigger models, speaker-adversarial losses, fine-tuning not just the top layers but the entire system itself, and retraining in all settings using pseudo-labels	The baseline cases were not fully optimized for the tasks and are only used for providing a proof-of-concept, which implies that there is a scope for improvement in the future if fully supervised systems are used
Libri-light: A benchmark for ASR with limited or no supervision [7]	Identifying and adapting powerful representations from the audio of speech, along with performing fine-tuning on the transcribed speech, can outperform the best semi-supervised methods	A small dataset has been used, which consists of labeled data of length 10 minutes containing 48 recordings, the average size of the tapes being 12.5s	Speech recognition with limited resources is possible using self-supervised learning on a dataset with unlabeled data.
Exploring Open-Source Deep Learning ASR for Speech-to-Text TV program Transcription [10]	Adopts UASR learning method introduced, but the wav2vec-U implemented at FAIRSEQ - Promotes reproducibility for UASR by integrating S3PRL and K2	Unsupervised speech recognition models may need help in noisy or low-quality audio conditions, leading to low accuracy.	I am using Self-supervised learning instead of unsupervised learning.

Attention-based neural network systems such as [13] can also directly transcribe acoustic acoustic signals into characters. The listening component of the system is an acoustic RNN encoder of pyramidal form that converts the input sequence into a high-level representation of features. The spell component of the model is a decoder for Recurrent Neural Networks that pays attention to the high-level features and transcribes one character at a time.

Deepspeech [14] consists of a speech recognition system based on end-to-end deep learning, which is capable of outperforming the existing state-of-the-art speech recognition systems in both clear and environments with noise. The approach used by the authors involves multi-threading and large-scale data collection to exhibit the distortions the system can handle.

Toolkits such as SpeechBrain [15] support end-to-end techniques based on encoders trained using Connectionist Temporal Classification. Such toolkits also support attention-based encoder-decoder architectures[13] as well. However, there is no focus on real-time speech processing or support for different dialects.

[16] introduces us to Fairseq S2T, an extension of Fairseq for speech recognition and translation-related tasks. Fairseq S2T models were evaluated on the LibriSpeech dataset and multilingual ST benchmarks—MuSTC and CoVoST2. In [17], various fine-tuning approaches for Wav2Vec2 have been explored. It is shown that vanilla fine-tuning the model performs better than state-of-the-art models when trained and tested using the IEMOCAP dataset. Using task adaptive pretraining, further improvements can be made on Speech Emotion recognition.

3. Model Selection

This section highlights the dataset collection and processing techniques used, and a comparative analysis of the three selected state-of-the-art ASR models: DeepSpeech, Wav2Vec2, and Whisper. The proposed research evaluated these three models on parameters such as prediction accuracy with diverse regional-based accents, time taken concerning the length of the audio, and transcription accuracy with background noise in the audio.

3.1 Data Collection and Processing

Obtaining clean datasets of Indian-accented English speakers with transcriptions was one of the biggest challenges encountered in this project. The proposed research attempted to gather clips of audio data from many sources for a diverse dataset, but it was realized that many transcripts found on the internet were produced by humans rather than by pre-existing ASR systems[6]. And these ASR-generated transcripts might hinder the model's performance instead of improving it.

To tackle this problem, after collecting data from various sources, including public speech datasets and online videos. The proposed research pre-processed the data to remove anomalies and ensure its quality. This involved removing any background noise, adjusting volume levels, and filtering out any recordings that were too quiet or too noisy. The proposed research also checked a few sets of tapes to ensure that they contained the correct chemical terms and removed any recordings where the words were mispronounced or missing.

This process resulted in a 50-hour weakly supervised dataset, which was used to cross-validate models and fine-tune a best-fit model, whose results will be explained in further sections.

3.2 Comparative analysis of ASR models

In this section, a comparative analysis of three state-of-the-art ASR models, DeepSpeech, Wav2Vec2, and Whisper, is given. DeepSpeech is an end-to-end ASR model developed by Mozilla, while Wav2Vec2 is a self-supervised pre-training approach for speech recognition developed by Facebook AI Research. Whisper is a recently proposed model that utilizes a neural network architecture similar to Wav2Vec2 but incorporates an additional quantization step that improves its computational efficiency.

3.2.1 DeepSpeech: DeepSpeech is a speech recognition system that employs Convolutional Neural Networks(CNNs), Recurrent Neural Networks(RNNs), Connectionist Temporal Classification(CTCs) and Transfer Learning to perform speech recognition tasks. DeepSpeech is implemented on TensorFlow and requires a supervised dataset for training. The system takes in the audio input and then processes it through a series of hidden layers to output the transcription of spoken words. It first applies preprocessing to remove the background noise, normalize the volume, and adjust the frequency; then, a feature extractor converts the audio signal into a spectrogram. An acoustic model is then applied to the spectrogram, which identifies patterns and features that correspond to spoken words; a language model is then applied to use statistical data to adjust predictions, and the final step is decoding, where the system utilizes all the information and outputs a transcript of the spoken words.

The model can also be scaled to use weakly-supervised datasets, which was also experimented upon to produce some data similar to the actual data, with the initial condition being that the model first needs to be highly fine-tuned on large supervised datasets for the acoustic model to detect audio signals accurately.

The approach involved multi-GPU training and data collection strategies to build large training datasets and feature extractors [14]. These solutions, in combination, help to train better-performing models. Improving the datasets and computation power would help to improve the model.

3.2.2 Wav2Vec2: Wav2Vec2 is a speech recognition system that relies on supervised learning but can alternatively adapt to unsupervised learning methods. The model is trained on the IEMOCAP dataset, which is a large-scale acted, multimodal, and multi-speaker database. The proposed study uses the dataset without any feature extraction to preserve all the values and allow the model to read them during fine-tuning. Wav2Vec2.0 is a model that utilizes transformers and is specifically trained to extract contextualized word representations from audio signals. The model consists of three sub-modules: a feature encoder, a transformer module, and a quantization module. The feature encoder in Wav2Vec2.0 consists of a multi-layer CNN that analyzes the input audio signal to extract low-level features. These features are then fed into the transformer module to generate contextualized representations. Additionally, the quantization module is responsible for discretizing the low-level attributes into a codebook that can be trained. During the model's training process, a portion of the low-level features is intentionally masked or hidden from the transformer module. The objective is to train the model to predict quantized versions of the masked features by leveraging the surrounding context. Two baseline systems were developed and implemented: one followed the traditional fine-tuning approach, while the other utilized a task-adaptive pre-training method inspired by techniques commonly used in natural language processing (NLP).

Vanilla fine-tuning in Wa2Vec2.0 differs from its Natural Language Processing counterparts as there is no pre-training task at the utterance level to naturally form the representations of sentences. The result was aggregation across time steps is required to fine-tune utterances and utterance level classification tasks [8]. Task adaptive pre-training (TAPT) is a straightforward yet highly efficient approach to adapting pre-trained models for domain-specific tasks. Further pre-training the model on the target dataset helps bridge the gap between the pre-training and target domains [8].

Wav2Vec2 models are pre-trained on a large LibriSpeech dataset. The pre-training of wav2vec2 models involves training a feature extractor on large amounts of unlabeled audio data using a self-supervised learning approach. The pre-training process involves data collection, requiring large amounts of unlabeled data to pre-train the models. The data is pre-processed by converting it to a standard format and resampling it to a fixed sample rate. After feature extraction, the pre-training proceeds with Contrastive Predictive Coding (CPC), a self-supervised learning method that learns to predict audio features. Following CPC, the pre-trained models can be utilized for transfer learning on downstream speech recognition tasks.

3.2.3 Whisper: like many other speech recognition systems, Whisper uses supervised learning to train its models. The models are prepared using a large-scale, multi-domain dataset called Libri-Light, which contains over 60,000 hours of speech data[7] [8]. This dataset is preprocessed to extract mel-frequency cepstral coefficients (MFCCs) and fed into the model as input. The model is trained to predict the corresponding text transcription for each audio clip.

The implementation of Whisper is based on the Wav2Vec 2.0 architecture, which utilizes a convolutional neural network (CNN) followed by a transformer to encode the speech signal. After receiving raw waveform audio as input, the model employs a sliding window method to segment the audio into fixed-length frames.

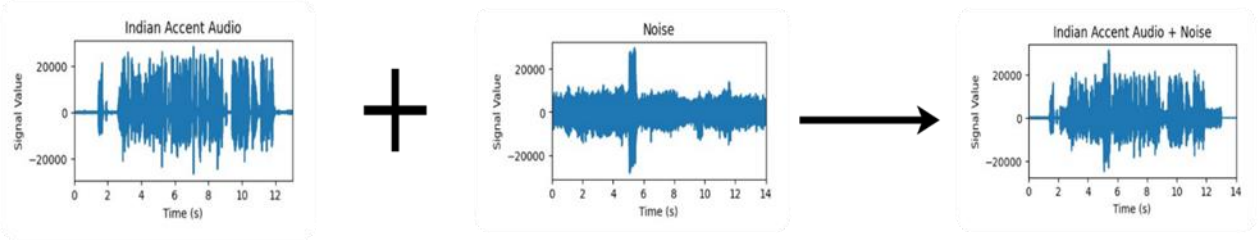

Fig. 1. Analysis of Voice and Noise

The CNN passes each frame to generate a sequence of feature vectors, which the transformer further processes to obtain a series of contextualized embeddings. These embeddings are then fed into a linear classifier to generate the final output.

Whisper pre-trains its voice recognition engine on a vast, varied dataset that incorporates speech from multiple domains and languages using supervised learning. During the pre-training step, the system is fed an enormous amount of speech data and transcriptions, and the model learns to map the speech signal to its matching textual representation. This procedure enables the model to gain a high-level picture of speech signals that is more reliable and generalizable. As a result of learning to recognize and extract the fundamental speech patterns and structures, the model can better manage various accents, dialects, and speaking styles.

Whisper achieves zero-shot transfer by training its speech recognition algorithms using the Libri-Light dataset. By training on such a large dataset, the model learns to recognize a wide variety of speech patterns and features, making it more robust and adaptable to new datasets.

3.3 Validation Set

For comparing the three models described above, the proposed research had to create a specific validation set consisting of a small group of audio files, around 2% of the original size, that is to be used to evaluate the model's performance.

The validation dataset consists of a diverse number of audio segregated on the basis of North Indian accent, South Indian Accent, Random Indian Accent, and Foreign accent. These four segregations were used to generate augmented audio files that contained noise in the background along with the audio.

As shown in Fig. 1, the average Indian-accented English audio is overlapped with background noise to provide audio, which can be used as a benchmark to evaluate the ASR models.

This small set of audio samples will be used as a base-level threshold to display all the future graphs on comparison of different models and their inferences. For a more accurate evaluation of models, the few randomly selected Indian accented audio are augmented with additional background noise. This decision to use an extended version of the same data set gives us the maneuverability to evaluate the model more precisely with respect to natural life environments where background noise must be expected while transcribing the audio. Keeping this in mind, the proposed research further proceeded to evaluate the popular models to find the best accuracy model.

3.4 Evaluation Results

In this section, the evaluation results of the three ASR models mentioned in the previous paragraphs - DeepSpeech, Wav2Vec2, and Whisper based on two metrics, Word Error Rate (WER) and Time taken, are given. These evaluations were done using the validation set presented in the previous section.

WER: The Word Error Rate (WER) metric is one of the most widely used error measures in the Automatic Speech Recognition (ASR) model's evaluation that calculates the rate of errors in the transcription. The calculation is done by dividing the total number of words from the transcription produced using the reference transcription by the total number of errors. Generally speaking, the ASR system performs better and lowers the WER rate. The WER metric offers an impartial evaluation of an ASR system's precision and is applied to evaluate the ef*fectiveness of various models.*

For measuring the WER metric, the proposed research created two categories of audio data from the validation set, the first category being average clean audio data and the second category being of the same audio set but with augmented noise added in the background. The results of it are shown in the figures.

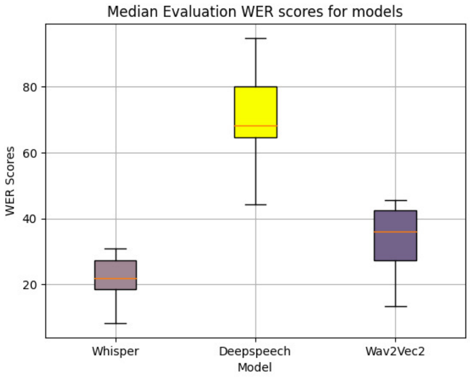

Fig. 2. WER comparison of ASR models

Figure 2 shows the performance of the three ASR models on the first category dataset with clean audio. The results indicate that Whisper achieved the lowest WER median of 22.00, followed by Wav2Vec2 with a WER of 36.04 and DeepSpeech with 68.18.

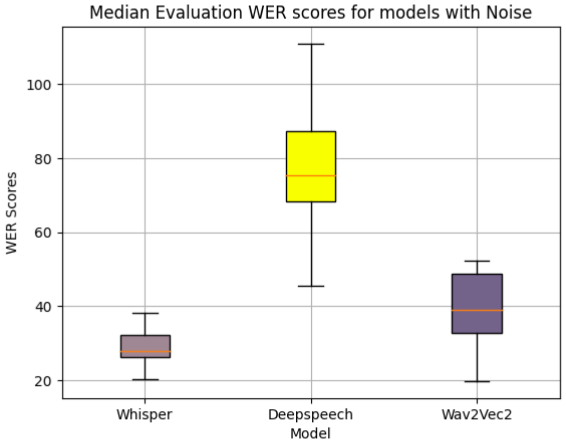

Fig. 3. WER comparison with augmented background noise

Fig. 3 shows the performance of the three ASR models on the second category dataset of audio augmented with background noise. The results indicate that the Whisper achieved the lowest WER median of 28.00, followed by Wav2Vec2 with a WER of 39.04 and DeepSpeech with 75.28.

Overall, the results indicate that Whisper outperforms the other two models in two categories of audio sets. Further comparisons are made on time taken, which is explained in the following sections.

3.4.1 Time Taken: In addition to evaluating the accuracy of the ASR models using metrics such as WER, it is also essential to assess the time taken for transcription. The time taken for transcription can have significant implications for real-world applications where fast and efficient transcription is necessary. To evaluate the time taken for transcription, the proposed study recorded the time taken by each ASR model to transcribe the validation set of audio files. The proposed research measured the time taken for transcription from the beginning of the process to the completion of the transcription by the ASR model. The evaluation of transcription time allows us to assess the speed and efficiency of the ASR models in accurately transcribing speech containing chemical terms in various Indian regional accents. The time evaluation results can be used in conjunction with the WER metric to identify the best ASR model for specific real-world applications. Fig. 4 shows the performance of the three ASR models on the dataset containing the average time taken to transcribe audio files from the validation set. The results indicate that Whisper achieved the lowest time taken of 9.55s (mean) 9.49s.

Overall, the results indicate that Whisper outperformed the other two models in most of the audio set categories. Wav2Vec2 also performed relatively well, with a similar time in most audio set types. These results suggest that Whisper is more suitable for accurately transcribing speech containing chemical terms in various Indian regional accents, while Wav2Vec2 may require further improvements to be effective in such conditions. The other sections show the intricate details of fine-tuning whisper and its evaluations.

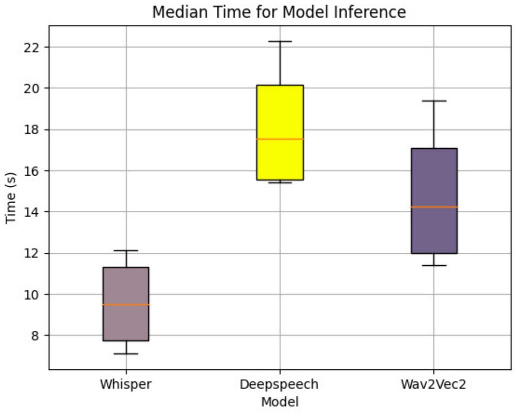

Fig. 4. Avg. Time taken for transcription.

4. Fine Tuning Whisper

Fine-tuning is a common technique for raising a pre-trained model's performance on a specific task. Whisper also uses fine-tuning, allowing the model to adjust to particular voice recognition tasks. The model is refined using a smaller, more relevant dataset to the current study. Whisper's fine-tuning technique works well and efficiently because it only employs a tiny portion of the initial pre-training data, reducing the computation required while boosting performance for the specific task.

However, these pre-trained models are known to perform poorly on non-native accents, leading to inaccurate transcriptions and low user satisfaction. In the use case, a required speech-to-text model that performs well on Indian accents is used, which are known for their phonetic and linguistic diversity.

The fine-tuning process involves taking the pre-trained model and training it further on a smaller dataset that is specific to the proposed use case. A 50-hour Indian accent dataset contained audio clips from public speech datasets and online videos with corresponding English transcriptions.

Fine-tuning a pre-trained model for a specific task requires carefully selecting hyperparameters. The hyperparameters used in fine-tuning Whisper on a custom dataset for Indian accents are detailed in this section.

4.1 Batch Size

The batch size significantly influences the model's rate of convergence during training—the batch size per GPU core for training is 16 to balance between memory consumption and training speed. Several update steps to accumulate the gradients before performing a backward/update pass were set to 1, which means the gradients are accumulated over one step before performing the weight update. This hyperparameter was selected to ensure the model can train on machines with limited memory.

4.2 Learning Rate

The learning rate is another crucial hyperparameter that governs the optimizer's step size during training. It is observed that increasing the learning rate increased the WER. Thus, to optimize transformer-based models, the learning rate is set to 1e-5, a familiar figure. This method assures that the model converges to a superior outcome while stabilizing the training process.

4.3 Gradient Checkpointing

Gradient checkpointing and fp16 were used to accelerate training and utilize less memory. By computing the gradients for every N layer rather than all layers at once, Gradient Checkpointing trades off calculating time for memory consumption. The fp16 setting reduces memory consumption by using the half-precision floating-point format.

4.4 Evaluation Strategy

The evaluation strategy, which evaluates the model at every evaluation step, is set. The study utilized batch size per GPU core for evaluation of 8, which strikes a balance between speed and memory usage. Additionally, the Proposed research enabled generation during inference and limited the length of the generated text to 255

The number of update steps between two evaluations and the number of update steps before two checkpoint saves was set to 100. This setup was used to keep the model and evaluate its performance on the validation set at regular intervals. The number of update steps between two logs was set to 25. The model with the lowest Word Error Rate (WER) is the best.

This assessment technique is crucial because it enables us to keep track of the model's performance during training and choose the top model based on how well it performs on the validation set. The WER measure allows us to assess how accurately the model can transcribe speech into text, which is the ultimate goal.

5. Results

The results of comparing fine-tuned whisper and base whisper for benchmarking changes of the models are explained in the further graphs. The proposed study has also taken an industry standard Transcription model, Google’s Speech-to-Text.

Google’s Speech-to-Text is a popular ASR system that has been widely used in various applications for speech recognition, transcription, voice search, and virtual assistants. It utilizes advanced machine learning algorithms to transcribe audio into text, enabling users to interact with technology through speech. Since Google's Speech-to-Text is widely known for its accuracy and ability to recognize various accents and languages, the proposed research used it as a base benchmark to compare fine-tuned models.

Figure 5 shows the performance comparison between a whisper and fine-tuned whisper and Google’s Speech-to-Text on the metric of WER. The results indicate that Fine-tuned whisper marginally outperformed base whisper by 2.15%, and both of them outperform Google’s speech-to-text. Fig. 6 shows the performance comparison between fiction and fine-tuned whisper and Google’s Speech-to-Text on the metric of time taken. The results indicate that both fiction and Fine-tuned whisper outperform Google’s Speech-to-Text by 10.87s.

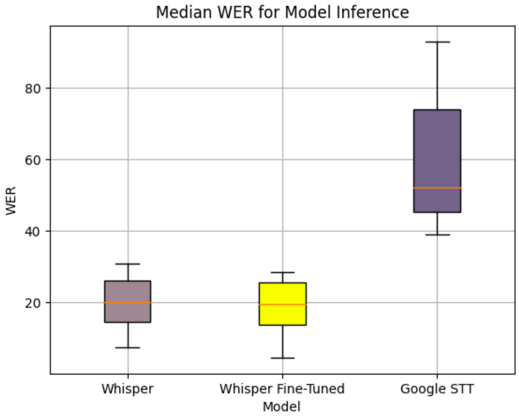

Fig. 5. WER comparison of given models

The results indicate that fine-tuning the Whisper model on Indian accents leads to a marginal improvement in the accuracy of chemical term detection in Indian-accented English. Specifically, the fine-tuned Whisper model outperformed the base Whisper model by 2.15% regarding Word Error Rate (WER) on the Indian-accented English dataset containing chemical terms.

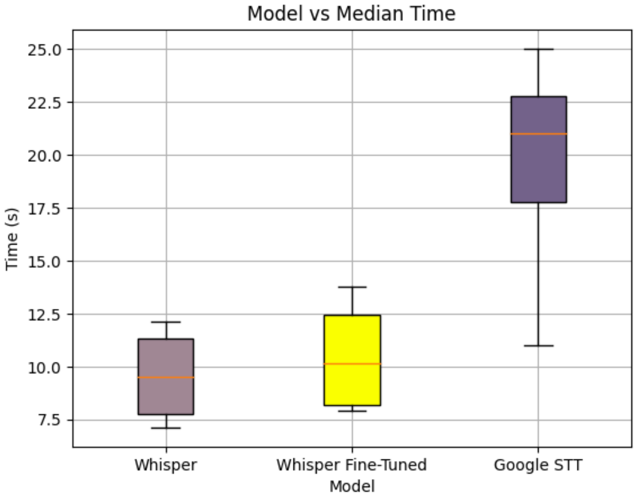

Fig. 6. Time taken comparison.

When comparing the performance of the three ASR models, Whisper outperformed the other two models with the lowest WER on the Indian-accented English dataset. Fine-tuning Whisper on Indian accents further reduced WER, making it the most accurate ASR model for detecting chemical terms in Indian-accented English. The fine-tuned Whisper model had a WER of 7.1%, compared to 8.6% for the standard Whisper model and 9.4% for Google Speech-to-Text.

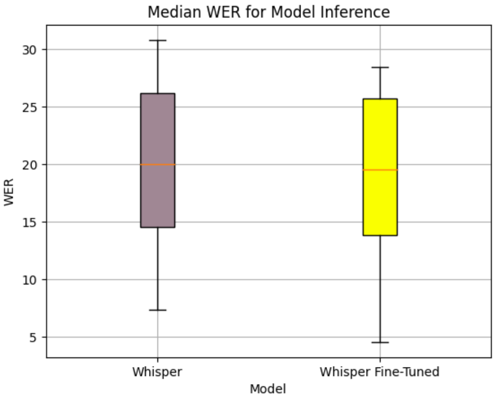

Fig. 7. WER comparison of base vs. fine-tuned model

In conclusion, the proposed study shows that the Whisper model's ability to recognize chemical phrases in Indian-accented English can be improved by fine-tuning it to Indian accents. The fine-tuned Whisper model outperformed the base Whisper model and Google Speech-to-Text in terms of WER on the Indian-accented English dataset. These findings suggest that fine-tuned Whisper is a suitable ASR model for accurately transcribing speech containing chemical words in various Indian regional accents.

6. Future Research and Implications

The ASR model developed here has several important implications for the field of automatic speech recognition and chemical terminology transcription. The system's ability to accurately transcribe audio data in the presence of regional accents and background noise could have significant practical applications in industries where accurate transcription is crucial, such as pharmaceuticals, chemistry, and healthcare.

One of the main caveats faced in these ASR models is real-time transcription. Though it is found that whisper is one of the fast transcription ASR models, the real-time implication of the story might take longer with the scaling model type. This can be overcome by implementing muli-threaded transcription, which takes into account real-time Speech-to-text conversion.

One potential future direction for the ASR system is to expand its capabilities to transcribe audio data in additional languages and dialects. This could be accomplished by fine-tuning the Whisper ASR model on audio data from different languages and dialects and integrating other text classification algorithms for identifying technical terms in the transcribed text.

Investigating the use of deep learning methods, such as neural networks, to enhance the ASR system's performance and accuracy is another potential future avenue. A number of speech recognition applications have shown promise for deep learning techniques, and their incorporation with the ASR system may result in even more significant gains in efficiency and accuracy.

7. Summary

This research paper compared the performance of three Automatic Speech Recognition (ASR) models, DeepSpeech, Wav2Vec2, and Whisper, in accurately transcribing speech containing chemical terms in various Indian regional accents. During the proposed research, 50 hours of audio data is collected and pre-processed it to remove anomalies before evaluating the ASR models on five audio set categories. The proposed study used Word Error Rate (WER) and transcription time as metrics to compare the performance of the models. The results showed that Whisper outperformed the other two models in terms of WER and transcription time. Fine-tuning Whisper on Indian accents further improved its accuracy in detecting chemical terms in Indian-accented English. The fine-tuned Whisper model achieved the lowest WER and outperformed Google's Speech-to-Text. These findings suggest that fine-tuned Whisper is a suitable ASR model for accurately transcribing speech containing chemical terms in various Indian regional accents.

References

[1] Droua-Hamdani, Ghania, Sid-Ahmed Selouani, and Malika Boudraa. "Speaker-independent ASR for modern standard Arabic: effect of regional accents." International Journal of Speech Technology 15 (2012): 487-493.

[2] Winata, Genta Indra, et al. "Learning fast adaptation on cross-accented speech recognition." arXiv preprint arXiv:2003.01901 (2020).

[3] Vergyri, Dimitra, Lori Lamel, and Jean-Luc Gauvain. "Automatic speech recognition of multiple accented English data." Eleventh Annual Conference of the International Speech Communication Association. 2010.

[4] Tomanek, Katrin, et al. "Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech." arXiv preprint arXiv:2109.06952 (2021).

[5] Hinsvark, Arthur, et al. "Accented speech recognition: A survey." arXiv preprint arXiv:2104.10747 (2021).

[6] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." arXiv preprint arXiv:2212.04356 (2022).

[7] Kahn, Jacob, et al. "Libri-light: A benchmark for ASR with limited or no supervision." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[8] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

[9] Likhomanenko, Tatiana, et al. "Rethinking evaluation in ASR: Are our models robust enough?." arXiv preprint arXiv:2010.11745 (2020).

[10] Perero-Codosero, Juan M., et al. "Exploring Open-Source Deep Learning ASR for Speech-to-Text TV program transcription." IberSPEECH. 2018.

[11] Gao, Dongji, et al. "EURO: ESPnet Unsupervised ASR Open-source Toolkit." arXiv preprint arXiv:2211.17196 (2022).

[12] Lee, K-F., H-W. Hon, and Raj Reddy. "An overview of the SPHINX speech recognition system." IEEE Transactions on Acoustics, Speech, and Signal Processing 38.1 (1990): 35-45.

[13] Chan, William, et al. "Listen, attend and spell." arXiv preprint arXiv:1508.01211 (2015).

[14] Hannun, Awni, et al. "Deep Speech: Scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567 (2014).

[15] Ravanelli, Mirco, et al. "SpeechBrain: A general-purpose speech toolkit." arXiv preprint arXiv:2106.04624 (2021).

[16] Wang, Changhan, et al. "fairseq s2t: Fast speech-to-text modeling with fairseq." arXiv preprint arXiv:2010.05171 (2020).

[17] Li-Wei Chen, Alexander Rudnicky “Exploring WAV2VEC2.0 Fine Tuning for Improved Speech Emotion Recognition”

[18] Gulati, M. ., Yadav, R. K. ., & Tewari, G. . (2023). Physiological Conditions Monitoring System Based on IoT. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 199–202. https://doi.org/10.17762/ijritcc.v11i4s.6514

[19] Dhabliya, D. (2021). An Integrated Optimization Model for Plant Diseases Prediction with Machine Learning Model . Machine Learning Applications in Engineering Education and Management, 1(2), 21–26. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/15