Multilingual Speech Recognition using Reinforcement Learning

A YOGI ATHISH
Department of NWC
S.R.M Institute of Science and Technology
Chennai, India
athish.aalla@gmail.com

SRINIVASA K G
Department of DSAI
IIIT Naya Raipur
Raipur, India
srinivasa@iiitnr.edu.in

SIVAKUMAR M
Department of NWC
S.R.M Institute of Science and Technology
Chennai, India
sivakumm3@srmist.edu.in

Abstract*— A system that properly converts spoken language into written text is what the voice recognition and transcription project seek to create. The system will analyze audio inputs and turn them into text using a variety of methods, including neural networks, language modeling, and acoustic modeling. The project's main goal is to increase transcription accuracy by tackling issues, including speaker accents, background noise, and audio quality. To make sure the system is capable of properly transcribing various speech varieties, it will be put to the test using a range of audio sources. The system's performance and limitations will be covered in a report that will accompany a functioning prototype of the voice recognition and transcription system. The project's findings will help numerous businesses that rely on voice recognition and transcription technologies improve communication, accessibility, and *production.

Keywords—Transcribing, Language, Audio, Voice Recognition.

I. INTRODUCTION

Modern society now relies heavily on speech, recognition, and transcription technology, which has many uses in fields including healthcare, education, and entertainment. This method has advanced the state of the art when refined using common benchmarks, particularly in low-data environments. The audio encoder which is pre-trained acquires the best-quality representations of voice, as they are entirely unsupervised, they don't have a decoder that is as effective in converting those representations into useable results and making fine-tuning stage very crucial to complete a job as the speech recognition. The ability to translate a spoken language into written text thanks to technology has greatly increased productivity, accessibility, and communication. The creation of a system that properly converts spoken language into written text is the main objective of this voice recognition and transcription project. The research will use cutting-edge methods to analyze audio data and turn them into text, including acoustic modeling, language modeling, and neural networks. The study will also concentrate on tackling issues, including speaker accents, background noise, and audio quality that have an impact on transcription accuracy. This shows that, even though unsupervised training has

significantly raised the efficiency of the audio encoder, a crucial flaw that severely restricts their usefulness and robustness is the absence of an equally high-efficiency trained decoder, combined with suggested certain rules of dataset finetuning. The main objective of the voice translation system would be to operate evenly "strangely" even though there would be variations in settings and not disturbing supervised decoder tuning for every deployment distribution. Speech recognition systems trained in supervised algorithms across various user data show.

The higher accuracy and much more effectiveness than models trained with a single origin [1]. The study focuses on scaling poorly trained models apart from English only. voice recognition to be working for different languages and work concurrently, in addition to expanding its scope. 117,000 of the 680,000, hours of audio are in 96 different languages. Additionally, 125,000 hours of Xen translation data are included in the collection. For sufficiently big models, we discover that combining multilingual and multitask training has no drawbacks or even benefits. Our research implies that poorly supervised pre-training with simple scaling has been overlooked for voice recognition up till now. The finished products of this project will be a functioning prototype of the voice recognition and transcription system and a report outlining the strengths and weaknesses of the system. The project will advance voice recognition and transcription technologies, enhancing accessibility, productivity, and communication across a range of sectors.

II. APPROACH

A. Data Processing:

In line with earlier work using text from the internet to train ML algorithms, the initial stage will involve obtaining audio from different sources, which includes audio recordings, podcasts, and videos. In contrast to typical research, the model is trained in such a way that raw text is predicted without any major standardization in raw audio and relies upon the expressiveness of trained models for learning to connect between what has been uttered and normal data. The input resources we employ to construct the dataset are the audio and internet transcripts. The dataset is shown in Figure 1. has Hindi dataset created as a consequence is highly diverse and contains audio from a range of locations, recording methods, speakers, and languages. To consider this, a variety of methods to filter that makes the model more effective. Current ASR systems produce many online transcripts rather than being made by people.

Fig. 1. Dataset after pre-processing (Hindi)

To remove any background noise and enhance audio quality, the audio data will be preprocessed utilizing techniques including noise reduction and audio filtering. The audio files are split into small segments of 30 secs and associate each one with the section of the transcript that happens at that time. While using probabilities that are sub-sampled, we train on all audio, including silence-filled sections, and use it.

The audio data will be divided into smaller pieces, such as words or phrases, at the following stage for additional processing. Finding pauses, silences, or other audio cues that mark the end of one unit and the start of another is a necessary step for the segmentation process. The system will receive segmented audio data and employ acoustic modeling, language modeling, and neural networks to convert the audio data to text. The main parameters to assess if data is transcribed correctly are (WER) which is Word Error Rate and (CER) Character Error Rate. The text will then undergo post-processing to fix any mistakes and improve readability. Spell-checking, grammar-checking, and text formatting are some of the post-processing approaches. Overall, the data processing strategy will combine speech recognition, segmentation, and audio preparation.

B. Modeling

We choose an off-the-shelf architecture as our main goal is to investigate the training of voice recognition systems, which helps us avoid confusing our results with model advancements. Due to the architecture's shown ability to grow consistently, we went with an encoder-decoder Transformer.

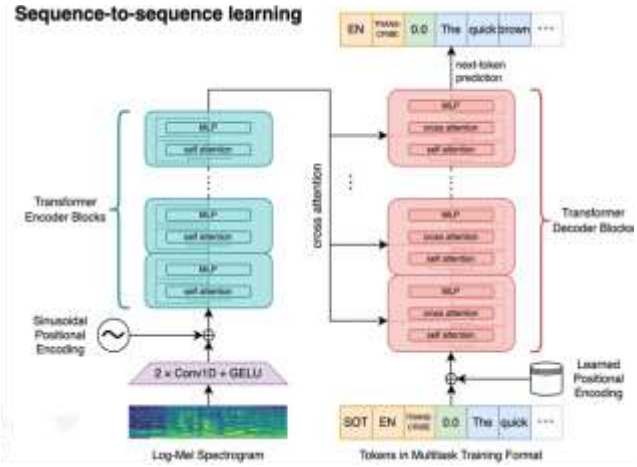

Fig. 2. Architecture of Model.

The Figure 2. Represent the general architecture of the model. The Mel Spectrogram is generated with a magnitude of 70-channel produced on 20-millisecond windows with a stride of 5 milliseconds after all audio has been re-sampled to 15,000 Hz. Different Deep learning algorithms used like Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN), or hybrid models would be trained using the segmented audio data. A mix of audio modeling and linguistic modeling approaches will be used to train the model. Utilizing both audio modeling and linguistic modeling strategies, the model will be initialized and trained. However, the model will interact with its environment and get feedback in the form of rewards or penalties based on the accuracy of its transcriptions rather than changing the model parameters through backpropagation.

TABLE I. DATASET SPLITING

Subset	**hours**	**Per-spk min.**	**Female spkrs**	**Male spkrs**	**Total spkrs**
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-others	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-300	363.6	25	439	482	921
train-others-500	496.7	30	564	602	1166

The Table 1. represents the dataset split for the dataset. The model's parameters will be iteratively changed during the training to avoid the difference or error between the text predicted and the actual text.

After the model has been trained, its accuracy and generalizability will be evaluated using a different test dataset. The model's effectiveness will be evaluated against other cutting-edge speech recognition methods.

Using reinforcement learning algorithms like Q-learning or policy gradient techniques, the model's parameters will be modified. The model will transcribe audio inputs during training and will get feedback depending on how accurate the transcriptions are. The model's parameters will be updated to increase rewards and decrease penalties. After the model has been trained, its accuracy and generalizability will be evaluated using a different test dataset.

The model's effectiveness will be evaluated against other cutting-edge speech recognition methods. To create a highly accurate and reliable speech recognition model, the reinforcement model training technique entails dataset selection, model initialization, training with reinforcement learning algorithms, and assessment. The approach is iterative and may involve hyperparameter tuning to optimize the model's performance.

III. Analysis

A. Model Scaling

The speech recognition and transcription project can benefit from the usage of bigger models, inventive structures, transfer learning, distributed training, and hardware acceleration. The model's performance might be significantly improved by being expanded. This can be accomplished by adding more layers or by Incrementing the neurons in every layer. Although training this model using a technique like this may need more computer resources and time, using cutting-edge structures like transformers and attention approaches may improve the accuracy which will in return increase the performance of the model. These topologies provide model ability in which it focuses more on input, which can improve speech transcription accuracy. The speech recognition and transcription project can benefit from the usage of bigger models, inventive structures, and transfer learning.

B. Dataset Scaling

The model's accuracy and generalizability may be considerably improved by expanding the dataset with 650,000 hours of audio. We evaluated the performance of many medium-sized models trained on sub-sampled datasets that were 0.5%, 1%, 2%, 4%, and 8% of the total dataset size with that of the small-sized model which is trained on the whole dataset. This can be done by gathering new data or by enhancing the already-existing information by including noise, altering the audio's pitch or speed, or adding other kinds of aberrations. The accuracy of the model might be significantly impacted by imbalanced datasets since the model may be biased toward the dominant class. The dataset can be balanced to increase the accuracy of the model by using data-balancing strategies, including over-sampling, under-sampling, or data augmentation.

Scaling the data for the voice recognition and transcription model may be also done via transfer learning. This includes initializing the model with data from different sources with the same use case, which can then be adjusted using the unique dataset. Selecting the most

instructive samples from the collection to label includes active learning. This strategy can greatly cut labeling costs while increasing the model's precision.

Techniques for domain adaptation can be used to modify the model for other domains or accents. This entails initializing the model with training on a related domain or accent, which can subsequently be improved on the particular dataset.

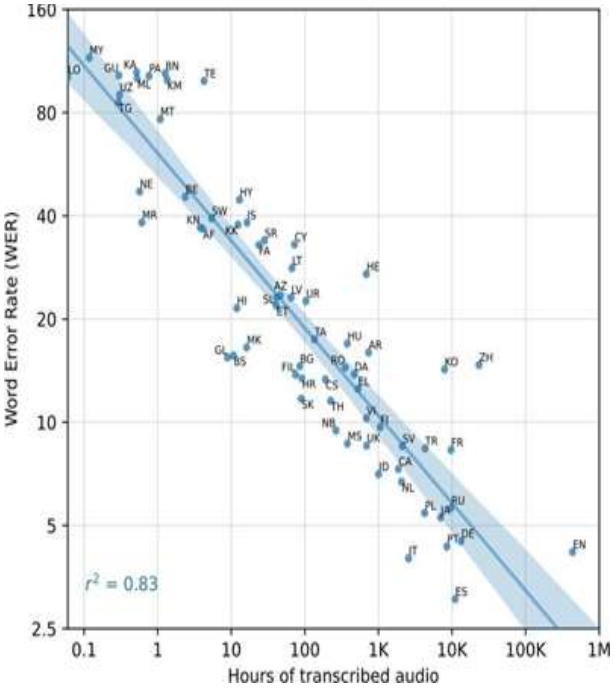

Fig. 3. WER of different languages

Figure 3. represents results for X to English translation and multilingual speech recognition. Speech Recognition in Multiple Languages (Fleurs). Although we observe substantial variation in improvement rates between tasks and sizes, all increases in dataset size led to increased performance on all tasks. Only a 1-point decrease in WER is shown when using the whole dataset, which adds another 12.5 to its size.

$$X(new) = \frac{X - X(min)}{X(max) - X(min)} \qquad (1)$$

For example, Equation (1) is a X(new) with the difference being in the denominator, Standardizer didn't mark negative variations of popular English phrases like "we're" vs "we are" in Call Home and Switchboard, and in WSJ, on a select dataset, our normalizer dramatically lowers WER.

C. Text Normalization

Accent normalization, Word-normalization, Phrase normalization, Domain normalization, and error correction can all be used to meet the voice recognition and transcription project goals. The accuracy and generalizability of the model may be considerably enhanced by expanding the dataset.

This was accomplished by gathering 6800 hours of multilingual datasets and introducing noise, modifying the audio's pitch or speed, or adding other types of aberrations. The accuracy of the model can be significantly impacted by imbalanced datasets since the model may be biased toward the dominant class.

The dataset has been balanced using data-balancing approaches including, over-sampling, under-sampling, or any other forms which increase the model's precision. Scaling the data for voice recognition and transcription model also uses transfer learning.

$$f(x) = \frac{1}{1 + e^{(x)}} \qquad (2)$$

Equation (2) shows the activation function which is used for modeling. This entails applying the whisper model to a comparable problem using a pre-trained model [1].

To validate this, we contrasted the efficiency of the model utilizing our normalizer with another one that was separately created as part of the Fair Speech project [2].

$$\varphi^{hier} = \frac{1}{K}(\varphi^{lid} + \sum_{k=2}^{K} \varphi_k^{inter}) \quad (3)$$

The overall CTC loss can be calculated by Equation (3). The varied formats and the way two normalizers make a negative impact can be attributed to the disparities in reduction to build the model from scratch, which may then be adjusted for the particular dataset. We have employed 32 distinct languages.

The model is trained with convolutional layers with different values of k ranging from 3 to 5.

However, domain adaptation methods may be used to adjust the model to new domains or accents. This entails initializing the model on a related domain or accent, which can then be improved on the particular dataset.

IV. RELATED WORKS

The paper proposes an end-to-end model for speech recognition based on the encoder-decoder architecture with attention mechanisms [1]. The paper presents a model for speech recognition based on neural network [2]. This paper studies different recurrent neural networks performance on different tasks [3]. Further studies propose a deep bidirectional long short-term memory (LSTM) network for speech recognition [4]. Moreover, the paper investigates the use of convolutional neural networks (CNNs) for speech recognition and compares their performance with traditional recurrent neural networks [5]. This paper proposes a method for robust speech recognition in unknown noise environments based on deep neural networks [6]. Further research proposes a method for separating speech signals from overlapping speakers using deep neural networks [7]. The paper proposes a transformer-based model for, achieving state-of-the-art performance on several benchmark datasets [8]. Further, this paper proposes a multi-task learning approach for simultaneous speech translation, achieving improved translation accuracy and real-time performance [9]. This paper proposes a multi-source attentional model for speech-to-speech translation, which incorporates both acoustic and text-based features to improve translation quality [10]. The paper proposes a semi-supervised learning approach for end-to-end speech translation, which leverages unlabeled data to improve translation accuracy [11]. This paper proposes a cross-lingual language model pretraining approach for code-switching speech translation, which improves translation accuracy in multilingual settings [12]. Further research proposes a dual supervised learning approach for neural machine translation with speech input, which jointly learns from both speech and text data to improve translation quality [13]. This paper is for speech recognition, achieve best performance on various benchmark datasets [14]. The paper investigates the use of CNNs for speech and compares their performance with traditional recurrent neural networks [15]. Further research proposes an attention-based model for speech recognition, which makes model to concentrate on various parts of sequence [16]. The paper proposes a model for recognition-based convolutions, and achieving best performance on various benchmark datasets [17]. This paper investigates the use different techniques for speech recognition, including various architectures such as CNNs, recurrent neural networks,

and deep feedforward neural networks [18]. Further research proposes the connectionist temporal classification (CTC) algorithm for speech recognition, which allows for end-to-end training without the need for explicit alignment between input and output sequences [19].

V. Limitations and Future Scope

The speech recognition and transcription project has maneuvered in the field but many things are yet to be done to overcome its limitations and challenges. Future work should focus on improving accuracy in noisy and multilingual environments, incorporating contextual information, and improving computational efficiency. Despite the advances in speech recognition and transcription technology, there are still some limitations and challenges that need to be addressed. The followings are some limitations and future scope are:

- **Handling background noise:** Speech recognition and transcription models can still struggle with background noise, which can significantly reduce the model's accuracy. Future work should focus on developing models that can handle noisy environments.
- **Improving accuracy for non-native speakers:** Speech recognition and transcription models can still struggle with non-native speakers, as their pronunciation and intonation can differ from native speakers. Future work should focus on developing models that can accurately transcribe speech from non-native speakers.
- **Handling multiple languages:** Speech recognition and transcription models can still struggle with handling multiple languages, which can limit their usefulness in multilingual environments. Future work should focus on developing models that can accurately transcribe speech in multiple languages.
- **Incorporating information:** Speech recognition and transcription models can still struggle with incorporating contextual information, such as the speaker's intent, emotions, and background knowledge. Future work should focus on developing mode.

VI. Conclusions

In conclusion, speech recognition and transcription are an important field that has numerous applications, including language translation, virtual assistants, transcription of audio and video content, and more.

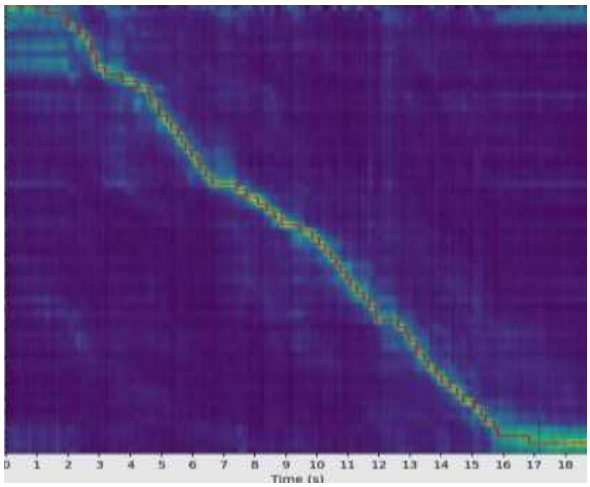

Fig. 4. Log-Mel Spectrogram

Figure 4. depicts the influence of the increasing size of the dataset on WER. Over the years, significant advances have been made in this field, driven by improvements in machine learning algorithms, computing power, and data availability.

In this project, we have discussed various approaches to processing speech data, including data preprocessing, model

training, and text normalization. We have also highlighted some of the limitations and challenges that the field faces, such as handling background noise, improving accuracy for non-native speakers, and incorporating contextual information. Figure 5. depicts the log-Mel spectrogram which shows about the training audio data with increase in time the spectrogram tends to 0 because of 20 sec audio file we used. Despite these challenges, the speech recognition and transcription project hold great promise for the future.

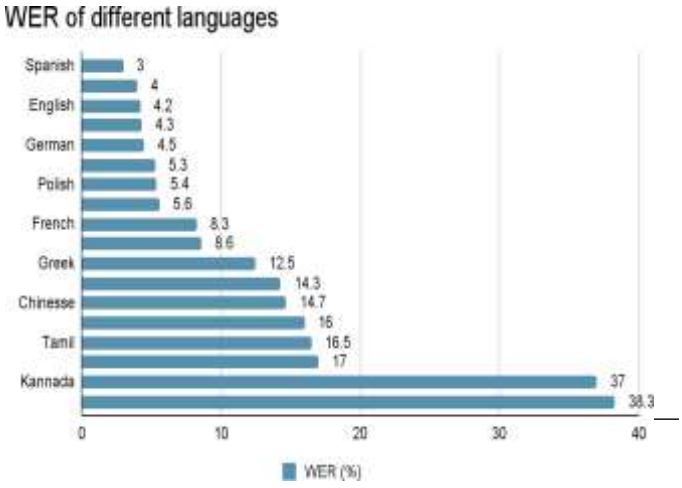

Fig. 5. WER of Different Languages

Figure 5. represents different languages and their WER. Advancements in technology will likely continue to improve the accuracy and efficiency of speech recognition and transcription models, making them more accessible and useful for a wide range of applications. With ongoing research and development, we can look forward to further advancements in the field of speech recognition and transcription in the years to come. The decoding process, including the use of acoustic and language models, is explained. Techniques such as beam search, n-gram language models, and attention mechanisms are discussed. The integration of language modeling with complex ML models for improved transcription accuracy is explored. The Spanish had the least WER and whereas Kannada had the highest WER.

REFERENCES

[1] Li, B., Zhang, Y., Sainath, T., Wu, Y. and Chan, W., 2019, May. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5621-5625). IEEE.

[2] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G. and Chen, J., 2016, June. Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182). PMLR

[3] Koudjonou, K.M. and Rout, M., 2020. A stateless deep learning framework to predict net asset value. Neural Computing and Applications, 32(14), pp.1-19.

[4] Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G. and Yu, D., 2014. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), pp.1533-1545.

[5] Kim, W., Stern, R.M. and Ko, H., 2005. Environment-independent mask estimation for missing-feature reconstruction. In Ninth European Conference on Speech Communication and Technology.

[6] Luo, Y., Chen, Z. and Mesgarani, N., 2018. Speaker-independent speech separation with deep attractor network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(4), pp.787-796.

[7] Vila, L.C., Escolano, C., Fonollosa, J.A. and Costa-Jussa, M.R., 2018, November. End-to-end speech translation with the transformer. In IberSPEECH (pp. 60-63).

[8] Effendi, J., Sakti, S., Sudoh, K. and Nakamura, S., 2020. Leveraging neural caption translation with visually grounded paraphrase augmentation. IEICE TRANSACTIONS on Information and Systems, 103(3), pp.674-683.

[9] Li, B., Sainath, T.N., Pang, R. and Wu, Z., 2019, May. Semi-supervised training for end-to-end models via weak distillation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2837-2841). IEEE.

[10] Feng, Y., Li, F. and Koehn, P., 2022, December. Toward the Limitation of Code-Switching in Cross-Lingual Transfer. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 5966-5971).

[11] Wang, Y., Xia, Y., Zhao, L., Bian, J., Qin, T., Liu, G. and Liu, T.Y., 2018, April. Dual transfer learning for neural machine translation with marginal distribution regularization. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[12] Graves, A., Mohamed, A.R. and Hinton, G., 2013, May. Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). Ieee.

[13] Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G. and Yu, D., 2014. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), pp.1533-1545.

[14] Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y., 2015. Attention-based models for speech recognition. Advances in neural information processing systems, 28.

[15] Hannun, A., Lee, A., Xu, Q. and Collobert, R., 2019. Sequence-to-sequence speech recognition with time-depth separable convolutions. arXiv preprint arXiv:1904.02619.

[16] Seltzer, M.L., Yu, D. and Wang, Y., 2013, May. An investigation of deep neural networks for noise robust speech recognition. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 7398-7402). IEEE.

[17] Tong, S., Garner, P.N. and Bourlard, H., 2017. An investigation of deep neural networks for multilingual speech recognition training and adaptation (No. CONF, pp. 714-718).

[18] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J., 2006, June. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).

[19] AUVOLAT, A. and MESNARD, T., Connectionist Temporal Classification: Labelling Unsegmented Sequences with Recurrent Neural Networks