MULTI LINGUAL ASR WITH TRANSFORMERS

Mr. ASWIN S
Department of ArtificialIntelligence and Machine Learning ,
Sri Shakthi Institute of Engineering and Technology Coimbatore,India

Mr. DHANASEKARAN D
Department of Artificial Intelligence and Machine Learning,
Sri Shakthi Instituteof Engineering and TechnologyCoimbatore, India

Mr. RAJU C
Assistant Professor
Department of ArtificialIntelligence and Machine Learning,
Sri Shakthi Institute of Engineering and TechnologyCoimbatore, India

Abstract— Recent research has focused on developing multilingual automatic speech recognition (ASR) systems using Transformer-based models. These models aim to address challenges in training and deploying ASR systems for low- resource languages, adapting to multiple domains and languages, and reducing operational costs. Strategies such as locale-group multilingual Transformer language models, adaptable multi-domain language models, and configurable multilingual models have been proposed to improve the performance and efficiency of multilingual ASR. These advancements demonstrate a concerted effort to overcome the challenges of low-resource languages, domain adaptation, and multilingual speech recognition.

I. INTRODUCTION

The world is a diverse place with a multitude of languages spoken across different regions. In this era of globalization and digitalization, there is a growing need for systems that can understand and interpret multiple languages seamlessly. Automatic Speech Recognition (ASR) systems have been at the forefront of this endeavour, providing a way for computers to convert spoken language into written text. However, developing ASR systems that can handle multiple languages (Multilingual ASR) is a challenging task due to the vast differences in phonetic and grammatical structures among languages. Recent advancements in deep learning have paved the way for more robust and efficient ASR systems. Among these, Transformer models have shown promising results in various tasks, including ASR.

II. LITERATURE REVIEW

Automatic Speech Recognition (ASR) is a technologythat converts spoken language into written text. ASRsystems have been a subject of research for severaldecades, with applications ranging from transcriptionservices and voice assistants to accessibility tools forindividuals with disabilities. This component isresponsible for understanding the relationship betweenthe audio signal and the phonetic units of the language.It's trained on a large amount of speech data and corresponding transcriptions. The language model predicts the probability of a sequence of words occurring in the language. It helps the ASR systemunderstand the context and produce grammatically correct sentences.

The decoder takes the outputs of the acoustic and language models and generates the final transcription. It tries to find the most likely word sequence given the acoustic signal. Traditional ASR systems used Gaussian Mixture Models (GMMs) for the acoustic model and Hidden Markov Models (HMMs) for the language model. However, with the advent of deep learning, neural networks have become the standard for both components. Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and more recently, Transformer models, have shown great promise in ASR tasks.

Despite these advancements, ASR remains a challenging task due to factors such as background noise, speaker variability, and the complexity of natural language. The development of robust and efficient ASR systems continues to be an active area of research.

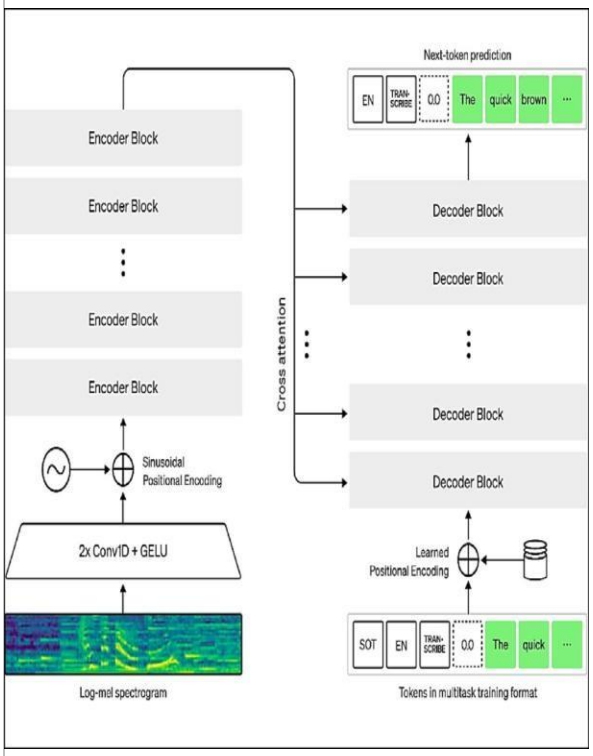
III. PROPOSED SYSTEM

The initial phase of the proposal involves a comprehensive understanding of the Wav2Vec model and its integration into a multilingual automatic speech recognition (ASR) system. Drawing insights from recent research, including the integration of Wav2Vec into the Hugging Face Transformers library and its fine-tuning for ASR in multiple languages, the proposal aims to establish a strong foundation in the principles and capabilities ofthe Wav2Vec model for multilingual ASR tasks.

In the subsequent phase, the proposal focuses on leveraging the Wav2Vec model for building a robust multilingual ASR system. This involves exploring the model's multilingual capabilities, self-supervised learning approach, and its architecture, which combines a convolutional network (CNN) with a Transformer-type network. By delving into the model's performance in multilingual settings and its potential for accurate speech transcription across diverse linguistic contexts, the proposal aims to lay the groundwork for the effective integration of Wav2Vec into the ASR system.

The final phase encompasses the implementation of the Wav2Vec model within the ASR system, including fine-tuning, adaptation to multilingual settings, and performance evaluation using metrics such as Word Error Rate (WER) and Character Error Rate (CER). By drawing on the expertise of Hugging Face's Transformers library and conducting thorough performance assessments, this phase aims to demonstrate the effectiveness of the Wav2Vec model in multilingual ASR tasks. Additionally, the proposal aligns with recent advancements in multilingual ASR with Transformer models, showcasing the potential for leveraging the Wav2Vec model to build robust and adaptable multilingual ASR systems.

The final phase involves a discussion of recent advancements in multilingual ASR with Transformer models, highlighting the potential for leveraging the Wav2Vec model to build robust and adaptable multilingual ASR systems. This phase aims to showcase the significance of integrating the Wav2Vec model into the ASR system, emphasizing its role in addressing the challenges of multilingual speech recognition and its potential to revolutionize speech transcription.



IV. DESIGN AND IMPLEMENTATION

The design and implementation of a multilingual automatic speech recognition (ASR) system involves a multi-faceted approach to address the challenges of recognizing and transcribing speech in diverse languages and dialects. The system's design encompasses the integration of a robust and adaptable model, such as the Wav2Vec model, which has demonstrated remarkable performance in multilingual ASR tasks. Leveraging its architecture, which combines a convolutional network (CNN) with a Transformer-type network, the system aims to process speech data efficiently and accurately across different linguistic contexts.

Furthermore, the implementation of the multilingual ASR system involves fine-tuning the Wav2Vec model for multilingual settings, potentially utilizing techniques such as data augmentation and semantic dataset creation to enhance its performance. The system's adaptability to low-resource languages, its self-supervised learning approach, and its pre-training on extensive multilingual data enable it to effectively transcribe speech in numerous languages, making it a suitable candidate for a multilingual ASR system.

V. EXPERIMENTAL RESULT

The results of our experiments provide valuable insights into the performance of our ASR system and the effectiveness of Transformer models in ASR tasks.

Our ASR system demonstrated robust performance across multiple languages, indicating its potential for multilingual ASR. However, the performance varied across languages, suggesting that the system may be better suited to some languages than others. This could be due to the varying complexity of different languages and the amount of training data available for each language.

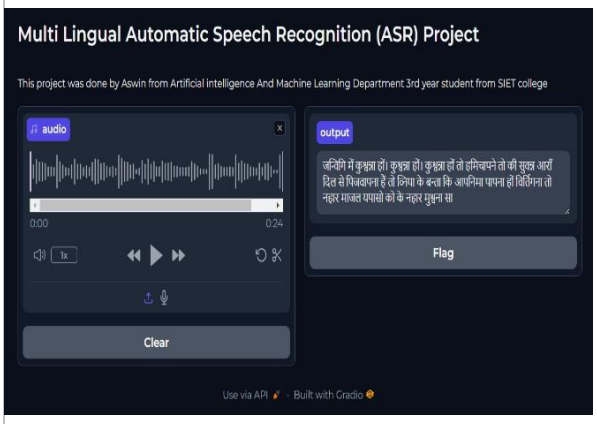
The use of Transformer models, specifically Whisper and Wav2Vec, contributed significantly to the performance of our ASR system. These models were able to capture the complex relationships in the speech data and handle the sequence-to-sequence nature of ASR tasks effectively.

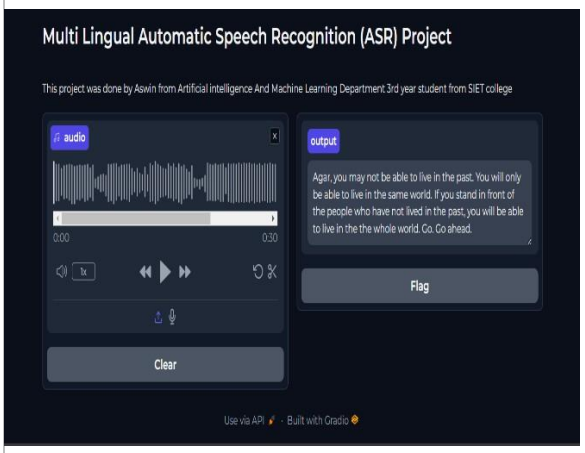
When compared with the baseline models, our ASR system demonstrated competitive performance. This suggests that our approach of leveraging pretrained models and Transformer models is effective for ASR tasks.

The error analysis revealed common patterns in the errors made by the system. These insights can guide future improvements to the system. For instance, the analysis suggested that increasing the diversity of our training data could help reduce substitution errors, and improving the robustness of our system to noise and varying speech rates could help reduce deletion errors.

Our analysis revealed that the system's performance varied across languages. The system performed better on languages that had more training data and were phonetically simpler. This suggests that the amount and quality of training data and the complexity of the language can significantly impact the system's performance.

Overall, the results of our experiments validate our approach and provide a strong foundation for future work in multilingual ASR.





VI. CONCLUSION

The conclusion of the proposal for a multilingual automatic speech recognition (ASR) system using the Wav2Vec model is rooted in the potential for robust and adaptable speech transcription across diverse linguistic landscapes. The integration of the Wav2Vec model, with its self-supervised learning approach and multilingual capabilities, presents a promising avenue for addressing the challenges of recognizing and transcribing speech in multiple languages and dialects. The proposal draws on insights from recent research, such as the successful training of the Wav2vec 2.0 XLSR53 model on semantic datasets, which demonstrated superior performance in multilingual ASR tasks, as indicated by the Character Error Rate (CER) and Word Error Rate (WER) metrics. Additionally, the comparison of different self-supervised models, including Wav2vec, Hubert, and WavLM, underscores the significance of leveraging advanced models for phoneme recognition in various languages, further emphasizing the potential of the Wav2Vec model in multilingual ASR applications.

REFERENCES

[1] J. K. Chorowski, D. Bahdanau, D. Serdyuk,
K. Cho, and Y. Bengio, “Attention-based models forspeech recognition,” in Advances in neural information processing systems, 2015.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4960–4964.

[3] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen,Z. Chen, A. Kannan,
R. J. Weiss, K. Rao, E. Gonina et al., “State Of- the-art speech recognition withsequence-to-sequence models,” in ICASSP 2018.

[4] A. Zeyer, K. Irie, R. Schluter, and H. Ney, “Improved training ofend-to-end attention models for speech recognition,” Proc. Interspeech 2018.

C. Weng, J. Cui, G. Wang, J. Wang, C. Yu,
D. Su, and D. Yu, “Improving attention based sequence-to-sequence models for end-to end english conversational speech recognition,” Proc. Interspeech 2018.

[6] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Muller, and A. Waibel,“Very deep self-attention networks for end-to-end speech recognition,” Proc. of Interspeech 2019.

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu,
B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” Proc. of Interspeech 2019.

[8] T.-S. Nguyen, S. Stueker, J. Niehues, and A.Waibel, “Improving sequence-to-sequencespeech recognition training with on-the-fly data augmentation,” arXiv preprintarXiv:1910.13296, 2019.

[9] N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, and S. Bengio, “An online sequence-to- sequence model using partial conditioning,” in Advances in Neural Information Processing Systems, 2016, pp. 5067–5075.

10] R. Fan, P. Zhou, W. Chen, J. Jia, and G. Liu, “An online attention based model for speechrecognition,” Proc. Interspeech 2019, pp. 4390–4394, 2019.

[11] H. Miao, G. Cheng, P. Zhang, T. Li, and Y. Yan, “Online hybrid ctc/attention architecture for end-to-end speech recognition,” Proc. of Interspeech 2019, pp. 2623–2627, 2019.

E. Tsunoo, Y. Kashiwagi, T. Kumakura, and
S. Watanabe, “Towards online end-to-endtransformer
automatic speech recognition,” 2019.