ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1816

Comparative Analysis of State-of-the-Art Speech Recognition Models for Low-Resource Marathi Language

Suhas Waghmare[1]; Chirag Brahme[2]; Siddhi Panchal[3]; Numaan Sayed[4]; Mohit Goud[5]
[1]GUIDE [2,3,4,5] BEAIDS/New Horizon Institute of Technology and Management/University of Mumbai/India

Abstract:- In this research, we present a comparative analysis of two state-of-the-art speech recognition models, Whisper by OpenAI and XLSR Wave2vec by Facebook, applied to the low-resource Marathi language. Leveraging the Common Voice 16 dataset, we evaluated the performance of these models using the word error rate (WER) metric. Our findings reveal that the Whisper (Small) model achieved a WER of 45%, while the XLSR Wave2vec model obtained a WER of 71%. This study sheds light on the capabilities and limitations of current speech recognition technologies for low-resource languages and provides valuable insights for further research and development in this domain.

Keywords:- *Speech Recognition, State-of-the-Art Models, Whisper, XLSR Wave2vec, Marathi Language, Low-Resource.*

I. INTRODUCTION

In recent years, speech recognition technology has made remarkable progress, primarily due to the development of sophisticated deep learning models. Among these models, Whisper by OpenAI and XLSR Wave2vec by Facebook have demonstrated impressive capabilities in transcribing speech into text across various languages and domains. Although these models have shown high performance in general, their effectiveness in low-resource language settings remains a subject of ongoing research and scrutiny. To address this research gap, this study aims to provide a detailed comparative analysis of the performance of two state-of-the-art speech recognition models, Whisper and XLSR Wave2vec, specifically applied to the Marathi language. Marathi is a low-resource language spoken by millions in India, and it presents unique challenges for speech recognition due to limited available data and linguistic variations. To assess the performance of these models in the Marathi language domain, we will use the Common Voice 16 dataset, which is a valuable resource for training and testing speech recognition systems.

We will use the word error rate (WER) as the chosen metric for performance assessment, which is a widely used measure that quantifies the accuracy of transcription by comparing predicted text with ground truth transcripts. Through this investigation, we aim to elucidate the strengths and limitations of the Whisper and XLSR Wave2vec models in the context of Marathi speech recognition. Our analysis will include a detailed evaluation of the models' performance in terms of accuracy, robustness, and efficiency. We will also examine the impact of different factors, such as data size, model architecture, and training strategies, on their performance. By shedding light on the comparative performance of these models, we seek to provide insights that can inform further advancements in speech recognition technology, particularly for low-resource languages like Marathi. Ultimately, such insights are crucial for enabling the development of more inclusive and effective speech recognition systems that cater to diverse linguistic communities worldwide.

II. DATASET

Common Voice is a publicly available dataset that contains recordings of people reading sentences in multiple languages. These recordings are made available for free to researchers and developers to train speech recognition systems. By collecting voice recordings from volunteers around the world, Common Voice aims to create more inclusive and accurate speech recognition models. Researchers and developers can use this dataset to advance the development of speech technology for various purposes, including accessibility, language learning, and voice-controlled devices. The Common Voice dataset includes various data fields, such as 'client_id', 'path', 'audio', 'sentence', 'up_votes', 'down_votes', 'age', 'gender', 'accent', 'locale', 'segment', and 'variant'. For our purposes, we only use the 'audio' and 'sentence' fields. We have a total of 7016 samples for the Marathi language, which we split into a combined training and validation dataset of 4906 samples and a separate test dataset of 2212 samples. This approach optimizes the use of the small dataset we have available.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1816

III. DATA PREPROCESING

After obtaining the dataset, we utilize the Whisper Feature Extractor on the audio samples. The feature extractor follows a series of steps for processing the audio signal. In the first step, the audio array is normalized to a length of 30 by either truncating or padding it. Then, we apply the windowing process, where the signal is divided into overlapping frames of fixed length. Each frame is typically overlapped with the preceding and succeeding frames to ensure continuity. Next, we apply the Fast Fourier Transform to each windowed frame, which converts the time-domain signal into the frequency domain. This results in a series of complex-valued frequency bins representing the magnitude and phase of different frequency components. We then calculate the power spectrum by taking the squared magnitude of each complex-valued frequency bin. This yields the power spectral density of the audio signal. After that, we multiply the obtained power spectrum by a set of triangular filter banks. These filter banks are spaced evenly in the Mel-frequency scale, which better approximates human auditory perception of sound. Then, we take the logarithm of the filter bank energies, and finally, we compute the spectrograms. These spectrograms are then passed into the Whisper model. In the case of the XLSR-Wav2Vec model, we extract the features and then convert them to logits for further computation.

IV. MODEL ARCHITECTURE

A. Whisper Model

This system is based on the transformer architecture, which consists of encoder blocks and decoder blocks with an attention mechanism that propagates information between them. The system takes an audio recording and splits it into 30-second chunks, processing them one by one. For each 30-second recording, it encodes the audio using the encoder section and saves the position of each word said. It leverages this encoded information to find what was said using the decoder. The decoder predicts tokens from this information, which are the individual words being said. It repeats this process for the next word using all the same information as well as the predicted previous word, which helps to guess the next one that would make more sense.

B. XLSR-wav2vec2 -

The XLSR-Wav2Vec2 model architecture consists of various essential components for efficient speech processing. The core of this architecture comprises convolutional layers that play a crucial role in transforming the raw waveform input into a latent representation denoted as Z. These convolutional layers extract critical features from the input audio, capturing important characteristics of the signal. After this initial processing, transformer layers are used to create contextualized representations denoted as C. These transformer layers help the model capture long-range dependencies and contextual information within the audio sequence, facilitating accurate and nuanced understanding of the input. Finally, a linear projection layer is applied to refine the representation and prepare it for downstream tasks such as speech recognition or speaker identification. Together, these components form a comprehensive architecture that enables the XLSR-Wav2Vec2 model to achieve state-of-the-art performance in cross-lingual speech recognition and related tasks.

V. RESULTS AND CONCLUSION

Two speech recognition models, the Whisper model by OpenAI and the XLSR Wave2vec model by Facebook, were tested on Marathi language data. The Whisper model achieved a WER of 45%, indicating higher accuracy than the XLSR Wave2vec model, which had a higher WER of 71%. The Whisper model was found to be more effective in low-resource settings, highlighting the potential of advanced neural network architectures in addressing challenges associated with speech recognition in underrepresented languages.

REFERENCES

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

[2]. Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self supervised learning of speech representations. arXiv preprint arXiv:2006.11477

[3]. Billa, J. (2018). Isi asr system for the low resource speech recognition challenge for indian languages. In INTERSPEECH, pages 3207–3211.

[4]. Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., and Wu, Y. (2021). W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. arXiv preprint arXiv:2108.06209.

[5]. Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision." In *International Conference on Machine Learning*, pp. 28492-28518. PMLR, 2023.

[6]. Shetty, V. M. and NJ, M. S. M. (2020). Improving the performance of transformer based low resource speech recognition for indian languages. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8279–8283. IEEE