Marathi Speech Recognition Using Machine Learning

Kaustubh Patil
Department of Computer, Pimpri Chinchwad College of Engineering and Research,Ravet,Pune
@kaustubh.patil_comp2020@pccoer.in

Darshan Nerkar
Department of Computer, Pimpri Chinchwad College of Engineering and Research,Ravet,Pune
@darshan.nerkar_comp2020@pccoer.in

Vishal Sharma
Department of Computer, Pimpri Chinchwad College of Engineering and Research,Ravet,Pune
@vishal.sharma_comp2020@pccoer.in

Sanskriti Narlawar
Department of Computer, Pimpri Chinchwad College of Engineering and Research,Ravet,Pune
@sanskriti.narlawar_comp2020@pccoer.in

***Abstract* - Speech recognition technology has seen significant advancements in recent years, enabling more efficient and accessible communication across diverse linguistic communities. In this paper, we explore the application of pretrained models for Marathi speech recognition, focusing on the "tanmaylaud/wav2vec2-large-xlsr-hindi-marathi" model provided by Hugging Face. Our study aims to evaluate the effectiveness of this pretrained model in accurately transcribing Marathi speech to text.**

We describe the methodology used to implement Marathi speech recognition, including data collection, preprocessing, and model selection. The experimental setup details the hardware and software environment, along with training procedures and evaluation metrics. Through experimentation and analysis, we assess the performance of the pretrained model and compare it with baseline approaches.

Our findings demonstrate the viability of pretrained models for Marathi speech recognition, showcasing their potential applications in diverse fields such as accessibility tools, language learning platforms, and transcription services. We discuss the implications of our research in improving communication and technology accessibility for Marathi speakers and outline future directions for advancing Marathi speech recognition technology.

keywords - Marathi speech recognition,Automatic speech recognition ,Natural language Processing.

I. INTRODUCTION

The ability to convert spoken language into text, known as automatic speech recognition (ASR), plays a crucial role in various applications, including virtual assistants, transcription services, language learning tools, and accessibility technologies. Marathi, one of the major languages spoken in India, presents unique challenges for ASR due to its linguistic characteristics and dialectal variations. In recent years, advancements in deep learning and the availability of large-scale datasets have significantly improved the performance of ASR systems.

This paper introduces a Marathi speech recognition system based on pretrained models, specifically leveraging the Wav2Vec2 architecture. Wav2Vec2, developed by Hugging Face, is a state-of-the-art framework for self-supervised learning of speech representations. The pretrained model used in this study, "tanmaylaud/wav2vec2-large-xlsr-hindi-marathi," is fine-tuned on large-scale multilingual datasets, making it suitable for handling Marathi speech.

The primary objective of this paper is to demonstrate the effectiveness of pretrained models for Marathi speech recognition tasks. We evaluate the performance of the model on various metrics, including word error rate (WER) and accuracy, to assess its suitability for real-world applications. Additionally, we discuss the challenges associated with Marathi ASR, such as dialectal variations, code-switching, and limited annotated data.

By leveraging pretrained models, we aim to provide an efficient and accessible solution for Marathi speech recognition, thereby facilitating communication, transcription, and language processing tasks in Marathi-speaking regions. This research contributes to advancing language technology for linguistic diversity and promoting inclusivity in speech-related applications.

II. LITERATURE REVIEW

Title: Designing and Developing a Marathi Speech Database for Native and Non-Native Emotional Speech in the Marathi Language

Authors: Bharati D. Borade, Ratnadeep R. Deshmukh, Santosh K. Maher, Swapnil Waghmare

Abstract: This research paper explores emotional speech recognition's applications in areas like assistive technology and human-computer interaction. Our focus is on developing Marathi vocabulary speech recognition systems using native and non-native audio datasets, considering mono/stereo channels and employing denoising. We engaged 24 speakers to mimic emotions in Marathi speech, recording happiness, sadness, and anger categories. The study details database construction, emphasizing design and development processes. This work contributes to understanding emotion-driven

technology, offering insights into linguistics and enhancing human-computer interaction across various domains.

Title: Marathi and Konkani Speech Recognition using Cross-Correlation Analysis
Authors: Alvan D'mello, Aishwary Jadhav, Jui Kale, Reena Sonkusare
Abstract: Systems that extract information from speech are common in communications and automation, where an individual's voice is analyzed and recognised by the system to follow the intended course of action. Extant research on speech recognition deals with the application of complex performance techniques of Artificial Intelligence (AI) and Machine Learning (ML) to study large non-isolated voice patterns. This paper uses simplistic performance parameters, namely, cross correlation for the identification of individual voice samples consisting of isolated Marathi and Konkani words in MATLAB. The study examines a dataset of 90 samples comprising 10 words from both languages for 10 individuals. The study presents a comparative analysis of .wav and .ogg formats of the audio samples with efficiencies of 93.5% and 88.5% for Marathi and Konkani respectively.

Title: Feature extraction using fusion MFCC for continuous Marathi speech recognition
Authors: Santosh Gaikwad, Bharti Gawali, Pravin Yannawar, Suresh Mehrotra
Abstract: This paper presents the performance of feature extraction techniques for speech recognition, for the classification of speech represented by a particular continuous sentence model. The goal of this study is to present independent as well as comparative performances of popular appearance-based feature extraction techniques, i.e., Linear Discriminative Analysis and Mel Frequency Cepstrum Coefficient. Mel Frequency Cepstrum Coefficient (MFCC) helps us in extracting feature while linear discriminant analysis (LDA) is used for reducing dimension of extracted feature. We experimented MFCC feature extraction individually and proposed a Fusion of MFCC and LDA for feature extraction.

Title: Large Vocabulary Continuous Speech Recognition System for Marathi
Authors: Snehal Chandulal Bajaj, Kamal Kant, Amol V Bole, Pranaw Kumar
Abstract: This paper presents an effort toward the development of a large vocabulary continuous speech recognition system for Marathi and its integration into practical applications, such as a speech-based typing system. The speech recognition system has been developed using the Kaldi toolkit, which facilitates the exploration of different acoustic models such as Gaussian Mixture Model with Hidden Markov Model (GMM-HMM) and Deep Neural Networks with Hidden Markov Model (DNN-HMM). In this paper, a Marathi Speech Recognition system is developed using around 95 hours of audio data and a trigram language model. The system achieves an average word error rate of

Title: Automatic recognition of class variants of Marathi consonants
Authors: Sameer Deshmukh, Chaitali Laulkar, Supriya Rajankar
Abstract: Marathi, spoken predominantly in the state of Maharashtra, India, has close to 70 million native speakers and is based on the Devanagiri script. This paper focuses on identifying individual class variants of the first five varnas of the Marathi Devanagiri script. The study aims to improve the understanding of Marathi phonetics and contribute to speech recognition systems for the language.

Title: A Critical Insight into Marathi Speech recognition and techniques
Authors: Satish Balshankar, Ratnadeeep R. Deshmukh
Abstract: This paper provides an overview of Marathi speech recognition systems and techniques. It discusses the current state of speech recognition technology for Marathi, focusing on signal processing, feature extraction, modeling, and matching techniques. The paper aims to provide valuable insights into the development and improvement of Marathi speech recognition systems.

Title: Naturalness improvement in text to speech synthesis using threshold amplitude of the syllable (Marathi language)
Authors: Pravin M Ghate, S. D. Shirbhadurkar
Abstract: This paper proposes an approach to improving the naturalness of text-to-speech synthesis in Marathi language. The method involves the concatenation of syllable-like units to produce speech output from text input. By determining a threshold point of amplitude for concatenating two words during sentence synthesis, the paper aims to enhance the naturalness and intelligibility of the synthesized speech.

Title: Syllabic Speech Synthesis for Marathi Language
Author: Soumitra Das
Abstract: This paper presents a novel methodology for synthesizing speech in the Marathi language based on syllabic division. The goal of the text-to-speech system is to convert written text into spoken form efficiently and naturally. By dividing words into syllabic clusters, the proposed approach improves speaking simplicity and naturalness. The paper discusses the process of syllabic decomposition, which provides valuable insights into pronunciation length, pause points, and emphasis. The technique is based on accurately determining the structure of words, either as consonant-vowel (CV) or phonetic (PH) structures, to ensure precise syllabic division and enhance accuracy and training efficiency.

Title: Marathi connected word speech recognition system
Authors: Priyanka P. Patil; Sanjay A. Pardeshi
Abstract: The paper presents a Marathi connected word speech recognition system utilizing the Mel-frequency Cepstral Coefficient (MFCC) feature extraction technique and

Continuous Density Hidden Markov Model (CDHMM). The system segments recorded speech signals into words using a speech segmentation algorithm based on Short Time Energy (STE) and Spectral Centroide (SC). MFCC features are extracted from isolated words, and CDHMMs are developed for each word in the speech signal. Multivariate Gaussian mixture density functions define observation probabilities generated by each state, with the Baum-Welch algorithm re-estimating model parameters. Bi-gram language models are employed to find Bi-gram pairs, and a log Viterbi beam search is used for maximum likelihood state sequence path determination. Experimental results demonstrate the overall system accuracy.

III. II. METHODOLOGY

:

Pretrained Model Selection: We begin by selecting a suitable pretrained model for Marathi speech recognition. In this study, we utilize the "tanmaylaud/wav2vec2-large-xlsr-hindi-marathi" model provided by Hugging Face. This model is specifically trained for speech recognition tasks and has been fine-tuned on large-scale multilingual datasets, including Marathi.

Google Colab Setup: To facilitate the execution of our experiments, we leverage Google Colab, a cloud-based platform that provides GPU resources for running machine learning tasks. Google Colab offers a convenient environment for training models, conducting experiments, and analyzing results without the need for extensive hardware resources.

Data Preparation: We collect or utilize existing datasets containing Marathi speech recordings. These datasets may include audio recordings of various speakers covering a range of topics and speaking styles. Proper preprocessing of the data is performed to ensure compatibility with the ASR model.

Model Integration: The selected pretrained model is integrated into the Google Colab environment. We install the necessary dependencies and libraries to ensure smooth execution of the ASR pipeline. This includes installing the Transformers library, which provides easy access to pretrained models for natural language processing tasks.

Speech Recognition Pipeline: We develop a speech recognition pipeline using the pretrained model. This pipeline takes audio recordings as input and returns the corresponding text transcription. We leverage the pipeline functionality provided by the Transformers library to streamline the ASR process.

Evaluation: The performance of the speech recognition system is evaluated using standard metrics such as word error rate (WER), accuracy, and processing speed. We assess the model's ability to accurately transcribe Marathi speech across different datasets and scenarios.

Fine-Tuning (Optional): Depending on the specific requirements and available resources, we may explore the option of fine-tuning the pretrained model on domain-specific or additional Marathi speech data. Fine-tuning can help improve the model's performance on specific tasks or adapt it to new domains..

IV. Dataset Description:

Key characteristics of the dataset include:

Language: The dataset primarily consists of speech in the Marathi language, spoken by native speakers.

Speakers: The dataset may feature recordings from multiple speakers, representing diverse demographics and backgrounds. This diversity helps ensure robustness and generalization of the speech recognition system.

Topics: The recordings cover a wide range of topics, including but not limited to everyday conversations, news, interviews, speeches, and monologues. This diversity in topics reflects real-world scenarios and improves the system's ability to recognize speech across different contexts.

Audio Quality: The quality of audio recordings may vary, ranging from high-quality studio recordings to noisy or low-quality recordings captured in different environments. Addressing variations in audio quality is crucial for developing a robust speech recognition system.

Annotations: The dataset may be accompanied by annotations or transcripts containing the corresponding text for each audio recording. These annotations serve as ground truth labels for training and evaluation purposes.

Size: The dataset size may vary depending on the sources and availability of recordings. Larger datasets typically provide more training data, leading to better model performance, while smaller datasets may suffice for specific applications or experiments.

V. Results and Analysis:

In this section, we present the results of our Marathi speech recognition system using the predefined model "tanmaylaud/wav2vec2-large-xlsr-hindi-marathi" provided by Hugging Face. We analyze the performance of the system based on various metrics and provide insights into its effectiveness and limitations.

Performance Metrics:

Word Error Rate (WER): WER measures the rate of inaccuracies between the recognized transcript and the ground truth transcript. A lower WER indicates better performance.

Experimental Setup:

We trained the speech recognition system using the predefined model and evaluated it on a held-out test set.

The test set consists of Marathi speech recordings from diverse speakers and topics, ensuring a comprehensive evaluation.
We calculated WER to quantify the system's performance.

Analysis:
Effect of Model Complexity: The large-scale pre-trained model "tanmaylaud/wav2vec2-large-xlsr-hindi-marathi" demonstrated strong performance, leveraging its ability to capture complex linguistic patterns in Marathi speech.
Data Variability: The system's performance may vary depending on the variability in the test data, including speaker accents, environmental noise, and speaking styles.
Limitations: Despite its strong performance, the system may encounter challenges in recognizing speech with heavyaccents, background noise, or rare vocabulary words.

Future Directions:
Fine-tuning: Fine-tuning the pre-trained model on domain-specific data may further improve performance, especially for specialized applications.
Data Augmentation: Augmenting the training data with variations in accents, noise levels, and speaking styles can enhance the system's robustness.
Model Architectures: Exploring different model architectures and training strategies may lead to improved performance on specific tasks or datasets.

References

1. Mahmud, N., Shohan, S. A., & Salam, K. M. A. (2021). Deep Borade, Bharati D., Ratnadeep R. Deshmukh, Santosh K. Maher, and Swapnil Waghmare. "Designing and Developing a Marathi Speech Database for Native and Non-Native Emotional Speech in the Marathi Language." In 2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 1-6. IEEE, 2023.
2. D'mello, Alvan, Aishwary Jadhav, Jui Kale, and Reena Sonkusare. "Marathi and Konkani Speech Recognition using Cross-Correlation Analysis." In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2021.
3. Gaikwad, Santosh, Bharti Gawali, Pravin Yannawar, and Suresh Mehrotra. "Feature extraction using fusion MFCC for continuous Marathi speech recognition." In 2011 Annual IEEE India Conference, pp. 1-4. IEEE, 2011.
4. Bajaj, Snehal Chandulal, Kamal Kant, Amol V. Bole, and Pranaw Kumar. "Large Vocabulary Continuous Speech Recognition System for Marathi." In 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS), pp. 1-6. IEEE, 2023.
5. Deshmukh, Sameer, Chaitali Laulkar, and Supriya Rajankar. "Automatic recognition of class variants of Marathi consonants." In 2015 International Conference on Pervasive Computing (ICPC), pp. 1-5. IEEE, 2015.
6. Balshankar, Satish, and Ratnadeeep R. Deshmukh. "A Critical Insight into Marathi Speech recognition and techniques." In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 1-5. IEEE, 2022.
7. Ghate, Pravin M., and S. D. Shirbhadurkar. "Naturalness improvement in text to speech synthesis using threshold amplitude of the syllable (Marathi language)." In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 1-5. IEEE, 2017.
8. Das, Soumitra. "Syllabic Speech Synthesis for Marathi Language." In 2023 1st International Conference on Cognitive Computing and Engineering Education (ICCCEE), pp. 1-5. IEEE, 2023.
9. Patil, Priyanka P., and Sanjay A. Pardeshi. "Marathi connected word speech recognition system." In 2014 First International Conference on Networks & Soft Computing (ICNSC2014), pp. 1-6. IEEE, 2014.
10. Brahme, Aparna, and Umesh Bhadade. "Marathi digit recognition using lip geometric shape features and dynamic time warping." In TENCON 2017 - 2017 IEEE Region 10 Conference, pp. 1-5. IEEE, 2017.