International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Sanskrit Voice to Text

Bijal Bharadva[1], Vinit Jha[2], Jaanvi Nambiar[3] , Nidhi Sanghavi[4]

[1]Student at Atharva College of engineering, Mumbai
[2]Student at Atharva College of Engineering, Mumbai
[3]Student at Atharva college of Engineering, Mumbai
[4]Assistant Professor at Atharva college of Engineering, Mumbai

---

Abstract - The Sanskrit language has been an important part of Indian culture and heritage for thousands of years. It is one of the oldest languages in the world and has been used for religious and philosophical purposes throughout history. However, with the advent of modern technology, the use and study of Sanskrit have become more challenging. One of the significant difficulties in studying the language is the lack of efficient tools to analyze and interpret the vast amount of Sanskrit texts available. To address this issue, we developed a Sanskrit speech recognition application that can accurately recognize spoken Sanskrit words and convert them into text format. The project utilizes the latest advances in speech recognition technology and machine learning algorithms to achieve high accuracy and precision in recognizing spoken Sanskrit. The application's design ensures that it can recognize various dialects and accents of Sanskrit language, making it useful for a diverse range of users. The project's primary aim is to facilitate the preservation and promotion of the ancient Sanskrit language and culture.

Key Words: Sanskrit ASR, Voice to Text, HMM, CNN, ASR, Feature Extraction

1. INTRODUCTION

Automatic Speech Recognition (ASR) is a subfield of computer wisdom and computational linguistics that develops technologies to allow the recognition and restatement of spoken words into text by computers with the main benefit of searchability. In recent times, the reign of the traditional statistical models to process speech recognition has been replaced by deep learning technologies. However, an ASR with deep learning requires hundreds of hours of data to work fairly well. A corpus of this size only exists for a small majority of languages; hence, speech recognition researchers substantially concentrate on these high-resource languages.

In recent times, there has been a further focus on developing speech recognition for low-resource languages. However, this arrival of technology isn't yet applicable to numerous low-resource Indian languages. One similar language is Sanskrit. The Sanskrit language has been an integral part of Indian culture and heritage for thousands of years. It's one of the oldest languages in the world and has been used for religious and philosophical purposes throughout history. The Vedas are the soul of Indian culture and are the storage of all types of jñÁnas i.e. wisdom, which are veritably useful in maintaining a healthy and pressure-free life in society. It's also extensively known as the source or root of sciences, including Physics, Chemistry, Botany, Zoology, Mathematics, Agriculture, Environmental Science, Architectural Science(vÁstu- vidyÁ), figure, Divination, Metallurgy, Medicine, Meteorology, etc. These subjects are bandied in colorful texts and narratives of the vast literature of Sanskrit independently. Still, in terms of technology, there are numerous challenges associated with the study of Sanskrit.

The application we aim to develop can help overcome a few of the challenges.

First, it can aid in the preservation and revival of the Sanskrit language. In recent years, there has been a renewed interest in Sanskrit as a language of knowledge and wisdom, and this application can play a vital role in promoting its use. Second, the application has potential applications in various fields, including education, research, and literature analysis. For instance, it can assist Sanskrit students in learning the language by providing an efficient tool for pronunciation and comprehension. Researchers can also use the application to analyze and interpret ancient Sanskrit texts more accurately and efficiently. Third, the development of the application reflects the potential of technology to promote and preserve cultural heritage. The use of cutting-edge technology to revive an ancient language like Sanskrit is a testament to the power of innovation and creativity.

2. REVIEW OF LITERATURE

1. “Sanskrit Speech Recognition using Hidden Markov Model Toolkit”, [IJERT[1]

This research aims to build a speech recognition system for the Sanskrit language. Hidden Markov Model (HMM) Toolkit (HTK) is used to develop the system. The system is trained to recognize 50 Sanskrit utterances with the data for training being collected from 10 speakers.

2. “Automatic Speech Recognition in Sanskrit: A New Speech Corpus and Modelling Insights”, IITB[2]

This research paper talks about the Automatic Speech Recognition (ASR) and describes the impact of unit selection in Sanskrit ASR. It also investigates the role of different acoustic models and language model units in ASR systems.

3. Research Proposal Paper on Sanskrit Voice Engine: Convert Text-to-Audio in Sanskrit/Hindi

This paper presents the Methodology, Application area of the Sanskrit Speech Recognition System. Their system is capable of teaching “Sanskrit Language” with the help of “Hindi Language”..

4. “CTC-Based End-To-End ASR for the Low Resource Sanskrit Language with Spectrogram Augmentation”, IISC [4]

In this research paper, the trained CTC acoustic model is evaluated in terms of character error rate (CER) on greedy decoding. Weighted finite-state transducer (WFST) decoding is used to obtain the word-level transcription from the character-

level probability distributions obtained at the output of the CTC network.

5. "Voice conversion using deep bidirectional long short-term memory based recurrent neural network"

In order to increase the naturalness and continuity of speech output in speech conversion, this article presents a conversion method based on the use of DBLSTM-RNN to model not only the relationship between the base language and the target language, but also the long-range context-dependencies in the acoustic trajectory.

6. "Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks"

This research paper gives an insight into various Speech recognition algorithms such as Hidden Markov model (HMM), Dynamic time warping (DTW), Artificial Neural Network (ANN).

7. "Automatic Speech Recognition for Sanskrit"

This paper talks about the work on building a speaker-independent, large vocabulary continuous speech recognition system for Sanskrit using HMM Toolkit (HTK). It uses a Sanskrit speech corpus with a vocabulary size of 8370 and word-level transcription of 1360 sentences.

8. "ISI ASR System for the Low Resource Speech Recognition Challenge for Indian Languages"

The paper talks about the ISI Automatic speech recognition system used to generate its submission across Tamil, Telugu and Gujarati speech recognition tasks for the low resource recognition challenge for the Indian languages.

9. "Deep Neural Networks for Acoustic Modeling in Speech Recognition"

This review paper represents the shared views of various research groups who have all had recent successes in using Deep neural Network (DNN) for acoustic modeling. The paper describes steps for successfully implementing the two-stage training procedure used in DNNs.

10. "EESEN: End-To-End Speech Recognition using Deep RNN models and WFST-Based Decoding"

The paper talks about the End-to-End Speech Recognition (EESEN) framework which excessively simplifies the process to build state-of-the-art Automatic Speech Recognition (ASR) System. This paper compares EESEN framework with DNN system, where EESEN framework achieves comparable word error rate (WERs) and also speeding up decoding significantly.

11. "Audio Augmentation for Speech Recognition"

This paper investigates speech augmentation methods that processes the raw signal at audio-level. It represents the results at 4 different Large Vocabulary Conversational Speech Recognition task with the training data ranging from 100 hours to 1000 hours to examine the effectiveness in various data scenarios.

2.1 DATA / CORPUS

The dataset used contains 78 hours of speech consisting of about 46,000 sentences, for ASR in Sanskrit. Keeping the rich and long cultural heritage the language carries, the dataset is diverse both chronologically and in terms of domain coverage. Further, the dataset contains utterances from 27 different speakers, out of which 20 are male speakers and 7 are female speakers, representing 6 different native languages including Hindi, Tamil, Kannada, Malayalam, etc.

2.2 FEATURE EXTRACTION

There are many features that can be extracted from an audio file for model building. Some commonly used features are:

1. Mel-frequency cepstral coefficients (MFCCs): MFCCs are widely used for speech and audio analysis. They represent the spectral envelope of a sound and are useful for recognizing phonemes and other speech sounds.
2. Mel-spectrograms: Mel-spectrograms are similar to MFCCs, but represent the spectral content of an audio signal as a 2D matrix. They are often used as input features for deep learning models.
3. Spectral features: Spectral features include spectral centroid, spectral bandwidth, and spectral rolloff, which represent different aspects of the spectral content of an audio signal.
4. Zero-crossing rate: The zero-crossing rate is the rate at which the audio signal changes sign. It is often used as a measure of the temporal structure of an audio signal.
5. Energy features: Energy features include root-mean-square (RMS) energy, which measures the overall energy of an audio signal, and short-time energy, which measures the energy of an audio signal over short time windows.
6. Pitch features: Pitch features include fundamental frequency (F0), which represents the perceived pitch of an audio signal, and autocorrelation, which is often used to estimate F0.
7. Filter bank energies: Filter bank energies are similar to MFCCs, but are based on a bank of overlapping bandpass filters that are used to model the human auditory system.
8. Spectral features: Spectral features such as spectral centroid, spectral bandwidth, and spectral rolloff can also be used to capture the spectral content of speech signals.
9. Time-domain features: Time-domain features such as zero-crossing rate, short-term energy, and pitch can also be useful for speech recognition.
10. Delta and delta-delta coefficients: Delta and delta-delta coefficients are used to capture the dynamic changes in the spectral features over time.

The choice of features will depend on the specific application and the type of model being used. It is also common to normalize or scale the features to improve the performance of the model.

2.3 EXISTING SYSTEM AND LIMITATIONS

I. HMM:

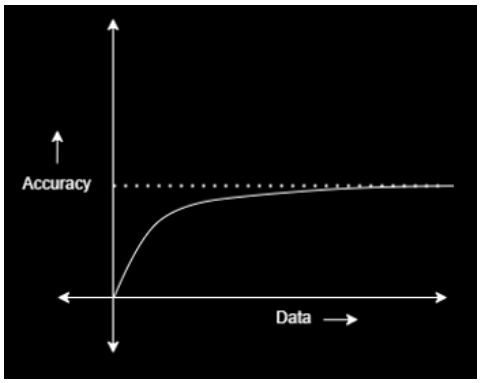

Fig-1 : HMM Model

HMM (Hidden Markov Model) is a widely used statistical model for speech recognition, including Sanskrit speech recognition. The basic idea behind HMM-based speech recognition is to model the speech signal as a sequence of hidden states, with each state corresponding to a phoneme or other speech unit.

Here are the basic steps for building an HMM-based speech recognition system for Sanskrit:

Data collection and preprocessing: Collect a large amount of speech data in Sanskrit and preprocess it to extract features such as MFCCs, Mel-spectrograms, or filter bank energies.

Acoustic modeling: Train a set of HMMs to model the acoustic properties of speech units such as phonemes or sub-phonetic units. This involves estimating the transition probabilities between states and the emission probabilities of the observed features.

Language modeling: Build a language model that captures the probability of word sequences in Sanskrit. This can be done using an n-gram model or other statistical language models.

Decoding: Given a new speech signal, apply the HMMs and the language model to find the most likely sequence of words that generated the observed speech signal.

Evaluation and tuning: Evaluate the performance of the system on a test set of speech data and tune the model parameters as needed to improve performance.

HMM-based speech recognition has been used successfully for many languages, including Sanskrit. However, it has some limitations, such as difficulty in modeling long-range dependencies and the need for large amounts of training data. More recent approaches such as deep learning-based models have shown promising results in improving speech recognition accuracy.

II. ML Based acoustic modeling:

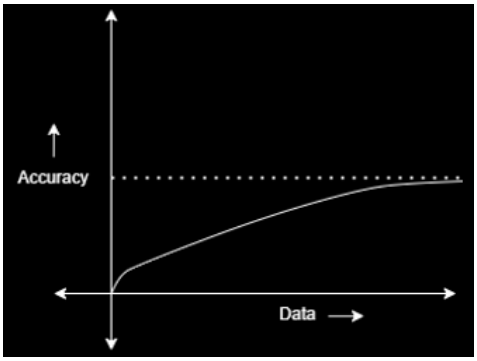

Fig-2 : ML Based acoustic model

Machine learning (ML) algorithms can be used for acoustic modeling in speech recognition systems. In fact, many modern speech recognition systems, including those for Sanskrit speech recognition, use machine learning algorithms to model the acoustic properties of speech.

There are several types of machine learning algorithms that can be used for acoustic modeling, including:

1. Gaussian Mixture Models (GMMs): GMMs are a popular choice for acoustic modeling in speech recognition systems. They are generative models that can capture the statistical distribution of speech features such as MFCCs.
2. Deep Neural Networks (DNNs): DNNs are a type of artificial neural network that can be used for acoustic modeling in speech recognition systems. They can learn complex nonlinear relationships between speech features and acoustic properties.
3. Convolutional Neural Networks (CNNs): CNNs are a type of neural network that can be used for acoustic modeling in speech recognition systems. They are particularly good at capturing local patterns in spectrogram-like representations of speech signals.
4. Recurrent Neural Networks (RNNs): RNNs are a type of neural network that can be used for modeling temporal dependencies in speech signals. They have been shown to be effective for acoustic modeling in speech recognition systems.

The choice of machine learning algorithm will depend on various factors such as the size of the training data, the complexity of the speech recognition task, and the available computing resources.

Although machine learning (ML) algorithms are widely used for acoustic modeling in speech recognition systems, they also have some disadvantages. Here are some of the drawbacks:

1. Data dependency: ML-based acoustic modeling algorithms require large amounts of high-quality training data to accurately model the acoustic properties of speech. Without sufficient training data, the models may not generalize well to new speakers, accents, or environmental conditions.
2. Complexity: ML-based acoustic models can be quite complex and computationally expensive to train and deploy, especially deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). This can limit their use in real-time or low-power applications.

3. Overfitting: ML-based acoustic models can be prone to overfitting, where the model memorizes the training data instead of learning to generalize to new data. This can result in poor performance on new data and a lack of robustness.
4. Vulnerability to noise: ML-based acoustic models can be sensitive to background noise and other sources of acoustic interference. This can lead to reduced accuracy in noisy environments or with low-quality recordings.

Overall, ML-based acoustic modeling is a powerful technique for speech recognition, but it requires careful attention to data quality, model complexity, and interpretability to achieve good performance.

III. End-to-end model:

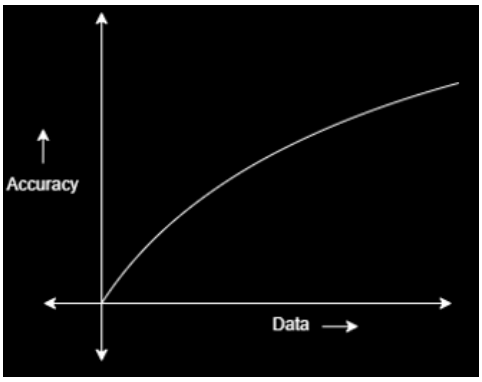

Fig-3 : End-to-End Model

End-to-end models for speech recognition have gained popularity in recent years, as they can simplify the traditional pipeline approach of acoustic modeling, pronunciation modeling, and language modeling. Here's an overview of an end-to-end model for Sanskrit speech recognition:
1. Data preprocessing: The audio signal is preprocessed to extract features, such as Mel-Frequency Cepstral Coefficients (MFCCs) or Filterbank Energies (FBEs).
2. Deep learning model: A deep learning model, such as a Convolutional Neural Network (CNN) or a Recurrent Neural Network (RNN), is used to directly map the audio features to the corresponding text transcription. The model takes in the audio features as input and outputs a sequence of characters or words that correspond to the spoken words.
3. Decoding: The output sequence of the deep learning model is decoded to generate the final transcription. Various decoding algorithms can be used, such as beam search or Connectionist Temporal Classification (CTC).
4. Training: The end-to-end model is trained using a large corpus of aligned audio and text data. The model learns to directly map the audio features to the corresponding text transcription without the need for separate acoustic and language models.
End-to-end models for speech recognition have several advantages over traditional pipeline approaches, including simplicity, faster development times, and the ability to learn complex relationships between audio and text data. However, they also have some limitations, such as a lack of interpretability and the need for large amounts of training data.

2.4 PROPOSED SYSTEM

A block diagram of the automatic speech recognition (ASR) system for the Sanskrit language is shown in Fig. 0. The following sections describe the block diagram in detail.

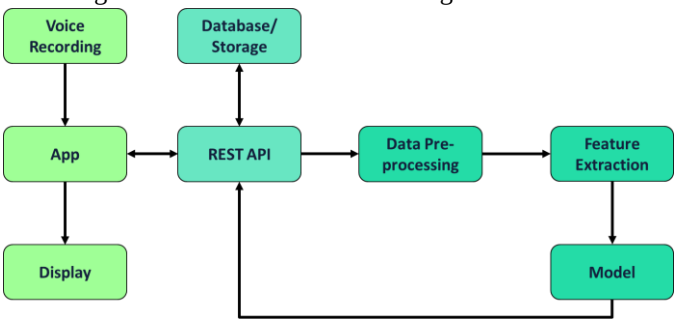

Fig- 4: Proposed System

Initially, the user will record Sanskrit speech input using a built-in voice recording service. The recorded audio file is sent to the back end and stored in the database. The audio file will go through preprocessing after which the spectrogram of each audio file will be computed. The next step is feature extraction which is done through CNN. An RNN layer is used for computing the characters and a CTC loss function to compute Character Error Rate (CER), Word Error Rate (WER), etc. After the computation, the corresponding text will be displayed on the screen of the device.

2.5 WORD ERROR RATE

The Word Error Rate (WER) and Character Error Rate (CER) are two commonly used metrics for evaluating the performance of speech recognition systems, including those based on the Connectionist Temporal Classification (CTC) model for Sanskrit speech recognition.

WER: The WER measures the percentage of words in the transcription that are incorrect. It is calculated by dividing the total number of word errors (insertions, deletions, and substitutions) by the total number of words in the reference transcription. WER is often used to evaluate the performance of speech recognition systems in natural language processing applications, where the goal is to transcribe spoken words accurately.

CER: The CER measures the percentage of characters in the transcription that are incorrect. It is calculated by dividing the total number of character errors (insertions, deletions, and substitutions) by the total number of characters in the reference transcription. CER is often used to evaluate the performance of speech recognition systems in languages with complex scripts or phonetic structures, where the goal is to transcribe individual phonemes or graphemes accurately.

In the case of Sanskrit speech recognition using CTC models, both WER and CER can be used to evaluate the system's performance. However, WER may be a more appropriate metric since Sanskrit has a complex phonetic structure.

In summary, both WER and CER can be used to evaluate the performance of Sanskrit speech recognition systems based

on CTC models, but WER may be more appropriate due to the language's complex phonetic structure and the output of character sequences by CTC models.

2.6 EXPERIMENT AND RESULT

This paper discusses the development of a speech recognition system for Sanskrit language. The proposed system is evaluated using a large corpus of 78 hours out of which the model was trained on approx. 55 hrs of speech corpus and evaluated on the remaining 23 hrs of corpus. Due to the complex structure of this language, we used Word Error Rate (WER) as an evaluation metric. WER for our end-to-end model is 0.44 and we also achieved 0.02 WER is the best score. As any other ASR system, our model is still not able to recognize all voices efficiently. It requires more training and we are working on using techniques likes Speed Perturbation, Spec Augmentation and adding a separate language model to improve speech recognition.

3. CONCLUSIONS

Speech recognition technology, an increasingly popular concept in recent years, has attracted attention from organizations to individuals; the technology is widely used for its various advantages. It brings the ability for a machine to listen and understand what people are talking about or what users are commanding. The research and implementation of an entire speech recognition model for the Sanskrit language provided great insight into the technology and future scope for research. In this work, we have built a Sanskrit ASR where various models like HMM, CNN, and ML Acoustic models were used, and the best results were obtained from the end-to-end CTC model. The system achieves a WER of 7.64%. A speech-to-text integration benefits for translation of the Sanskrit language such as allowing for inputting Sanskrit text without knowing the writing system and digitizing ancient documents. Authors of the Sanskrit language can greatly benefit from a service like this as they can dictate their material and reduce their typing workload to a great extent. The knowledge of Sanskrit is scattered across the globe, and residing in small villages, it is an indispensable need to preserve this knowledge, and an ASR application for Sanskrit feels like an ideal solution.

ACKNOWLEDGEMENT

We owe sincere thanks to our college Atharva College of Engineering for giving us a platform to prepare a project on the topic “Sanskrit ASR” and would like to thank our Principal Dr. Ramesh Kulkarni for instigating within us the need for this research and giving us the opportunities and time to conduct and present research on the topic. We are sincerely grateful for having Prof. Nidhi Sanghavi as our guide and Prof. Suvarna Pansambal, Head of Computer Engineering

Department, for their encouragement, constant support and valuable suggestions. Moreover, the completion of this research would have been impossible without the cooperation, suggestions and help of our friends and family.

REFERENCES

1. Lifa Sun, Shiyin Kang, Kun Li and Helen Meng. (2015). Voice conversion using deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Lifa, Shiyin, Kun Li & Helen, 2015

2. Anoop C. S., A. G. Ramakrishnan. (2021). CTC-Based End-To-End ASR for the Low Resource Sanskrit Language with Spectrogram Augmentation. National Conference on Communications (NCC), 2021, pp. 1-6, doi: 10.1109/NCC52529.2021.9530162. Anoop & Ramakrishnan, 2021

3. Pahini A. Trivedi. (2014). Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks. International Journal of Engineering Development and Research, Volume 2, Issue 4. Pahini, 2014

4. Jitendra Singh Pokhariya, Dr. Sanjay Mathur. (2014). Sanskrit Speech Recognition using Hidden Markov Model Toolkit. International Journal of Engineering Research & Technology (IJERT) IJERT ISSN: 2278-0181 IJERTV3IS100141 Vol. 3 Issue 10. Jitendra & Dr. Sanjay, 2014

5. Amrith Krishna, Preethi Jyoti, Rishab Kumar, Devaraj Adiga. (2021) Automatic Speech Recognition in Sanskrit: A New Speech Corpus and Modelling Insights. Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5039–5050. Amrith, Preethi, Rishab & Devaraj, 2021

6. Piyush Mishra, Jainendra Shukla. (2013). Research Proposal Paper on Sanskrit Voice Engine: Convert Text to Audio in Sanakrit/Hindi. International Journal of Computer Application 70(26):30-34. Piyush & Jainendra, 2013

7. C.S. Anoop., A.G. Ramakrishnan. (2019). Automatic Speech Recognition for Sanskrit. International Conference on Intelligent Computing, Instrumental and Control Technologies (ICICICT), 2019, vol. 1, pp. 1146-1151. Anoop & Ramakrishnan, 2019

8. Jayadev Billa. (2018). ISI ASR System for the Low Resource Speech Recognition Challenge for Indian Languages. INTERSPEECH, 2018, vol. September, pp. 3207–3211. Jayadev, 2018

9. Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97. Geoffrey, Li, Dong, George, Abdel-rahman, Navdeep, Andrew, Vincent, Patric, Tara & Brian, 2012

10. Yajie Miao, Mohammad Gowayyed, Florian Metze. (2015). EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 167–174. Yajie, Mohammad & Florian, 2015

11. T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. (2015). Audio Augmentation for Speech Recognition. INTERSPEECH, 2015, vol. January, pp. 3586–3589. Tom, Vijayaditya, Daniel & Sanjeev, 2015