IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Development of Speaker-Independent Automatic Speech Recognition System for Marathi Language

[1]Anushka Chaudhari, [2]Vina M. Lomte, [3]Tejas Lanjekar, [4]Bhushan Chaudhari, [5]Mrunmai Chinchwade,

[2]Head of Dept., Computer Engineering, RMD Sinhgad School of Engineering, Savitribai Phule Pune University

[1,3,4,5]Students, Computer Engineering, RMD Sinhgad School of Engineering, Savitribai Phule Pune University

Abstract: The research in the area of speech recognition for the English language and European Languages has reached up to a critical level to be used for a real communication tool. On the other hand, the research for Indian languages has not yet reached to develop an application. The research for Indian languages has been carried out now in various institutes and research labs but, the research is more concentrated towards the development of applications in Tamil, Telugu, and Hindi. Our aim to build a speech recognition system for Marathi language which is the third most popular language in India and fifteenth most popular in the world.

IndexTerms **- Continuous speech, Feature Extraction, Transformer, Wav2Vec2, WER.**

INTRODUCTION

In today's Scenario, we can see speech recognition for the English language is being done the most and has a lot of outcomes each with a unique efficiency and result. But as we know, with technological advancements people are able to learn more than 2 languages. In short, people are able to speak multiple languages, so it becomes natural that even the machines that are learning should learn all the languages. There has been some great development even in other languages like Chinese, Hindi, etc. and all with successful results. There are still certain areas which are not yet entirely focused or successfully researched in the field of speech recognition; and those fields are of the local languages, India is a very big country which speaks a lot of languages. There are 22 official languages in India and creating a successful automatic speech recognition model for all of them is nearly impossible – mostly due to the fact that not enough data is available on the unscheduled languages and that not a lot of people really use the language. A large number of people in India use English words or sentences in their day to day lives. Marathi language is one of the most commonly spoken language in India. Marathi speakers make up the third largest group in India, after Hindi and Bengali speakers, and rank fifteenth in the world as a whole. Despite having such a big population, speech recognition research is relatively scarce for this language. There isn't a specific speech recognition system for the language. Therefore, our research is focused on building continuous speech recognition system for the Marathi language.

LITERATURE SURVEY

Sr No	Publication details	Tech used	Dataset	Accuracy	Research Gap Identified
1	Development of Speaker-Independent Automatic Speech Recognition System for Kannada Language Praveen Kumar, H S Jayanna	Modelling techniques such as monophone, triphone, deep neural network (DNN)-hidden Markov model (HMM) and Gaussian Mixture Model (GMM)-HMM-based models were implemented in Kaldi toolkit and used for continuous Kannada speech recognition (CKSR). To extract feature Mel frequency Cepstral (MFCC) is used.	For this work dataset from Bharat Sanchar Nigam Limited (BSNL) is used. BSNL dataset offers an Integrated Responsive System (IVRS) call flow telephone service.	WER'S for monophonic - 8.36%, triphone1 - 6.22%, triphone2 - 5.38%, triphone3 - 5.12%, combination of SGMM and MMI - 4.84%, combination of DNN and HMM - 4.98%,	These least WER models (SGMM and DNN-based models) could be used to build a stable ASR framework. the precision of the ASR system can be expanded to allow implementation of noise reduction

				HMM, combination of DNN and Subspace Gaussian - 4.05%,	methods more effectively.
2	Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Is ¸ın Demirs ¸ahin, Cibu Johny, Martin Jansche†, Supheakmungkol Sarin, Knot Pipatsrisawat	To construct/develop this project four main resource components are required: a speech corpus, a phonolog-ical inventory, a pronunciation lexicon and a text normaliza-tion front-end. Among these four components, speech cor-pora are usually the most expensive to develop.	There are 6 datasets used for this work: Gujarati (Google, 2019a), Kannada (Google, 2019b), Malayalam (Google, 2019c), Marathi (Google, 2019d), Telugu (Google, 2019f) and Tamil (Google, 2019e).	95%	In the future, tools such as the one described by Podsi-adlo and Ungureanu (2018) could be used to streamline the recording script design process. Additional venues to explore include collecting high quality data for other low-resource languages of India, such as Sindhi (Cole, 2006), which can be used in transfer learning scenarios, where a reasonably small amount of data from a target language is used as adaptation data.
3	ACOUSTIC SPEECH RECOGNITION FOR MARATHI LANGUAGE USING SPHINX Aman Ankit, Sonu Kumar Mishra, Rinaz Shaikh, Chandraketu Kumar Gupta, Prakhar Mathur, Soudamini Pawar and Anil Cherukuri	For this project Sphinx4 library is used. Sphinx-4 is a state-of-the-art, speaker-independent, continuous speech recognition system written entirely in the Java programming language. The design of Sphinx-4 is based on patterns that have emerged from the design of past systems as well as new requirements based on areas that researchers currently want to explore. To exercise this framework, and to provide researchers with a "research-ready" system. Front end, Linguist, Acoustic model, Dictionary, and	Audio is recorded of 8 speakers (4 – Females and 4 - Males).	50%	Factors like local dialect on the language, variation of accent, etc affected the development. On a better note of achieving high efficiency, experiment to gather phonetic structure of same word spoken by different type of people and working on DNNs can further improve results.

		Language Model are the four blocks of the architecture of this project.			
4	Automatic Speech Recognition For Task Oriented IVRS In Marathi Manasi Ram Baheti, Bharti W. Gawali , S.C. Mehrotra	Feature Extraction: Mel Frequency Cepstrum Coefficient (MFCC). Mel-Frequency Cepstral Coefficients (MFCCs) is used for a variety of problems in signal processing. Different temporal and spectral analysis is done on the sound signals to extract the use- full features, the most important of them being the Mel Frequency Cepstrum Coefficients (MFCC). Recognition / Matching: Once feature vectors generated using MFCC, the next step is to find the optimal match. For this, the technique is DTW techniques has been used.The simplest way to recognize sentence sample is to compare it against a number of stored templates and determine the best match .	The recording was done in the ordinary room without noisy sound and effect of echo.	Speaker Dependent – 1) Offline: 93.43% 2) Online: 90.93%, Speaker Independent – 1)Offline: 88.12% 2)Online: 85.62%	The accuracy rate was affected due to the noise in the audio data recording. The variations in the performance rate depends upon the distance of speaker from or to the microphone, speed of the utterance, level of literacy etc. This observation leads to make use of high quality of microphone as to give input.
5	Design and Development of Continuous Marathi Speech Recognition System for Agriculture Purpose Pratik Kurzekar, Shriniwas Darshane, Nikhil Salvithal, Nitin Maske	Feature Extraction: Mel Frequency Cepstrum Coefficients (MFCC), development of text corpus, Pre-emphasizing, Framing and Windowing, Fast Fourier Transform.	Audio sample data was collected from various speakers which was grammatically correct and phonetically rich.	92.5%	Lack of availability of the database.
6	Marathi Speech Recognition System Using Hidden Markov Model Toolkit Sangramsing N. Kayte.	Preprocessing: speech-input is digitized using recognizer, Feature Extraction: , Model Generation: The model is generated using Hidden Markov Model (HMM).	Voices of eight people (5 male and 3 female) are used to train the system. The data is recorded using unidirectional microphones. Distance of approximately 5-10 cm is used between mouth of the speaker and microphone. Recording is carried out at room environment. Sounds are	Accuracy: 94.63% WRE: 5.37	The future Work can involve around the development of system for more vocabulary size and to improve the accuracy of the system.

			recorded at a sampling rate of 16000 Hz.		
7	Novel approach based feature extraction for Marathi continuous speech Recognition Santosh Kashinath Gaikwad, Dr.Bharti W Gawali, Suresh Mehrotra.	Feature Extraction: Mel Frequency Cepstrum Coefficient, Training: MFPCA (MFCC+PCA) MFDWT (MFCC+DWT) MFLDA (MFCC+LDA) MFPDWT MFCC + PCA + DWT) MFLDWT (of MFCC + LDA + DWT),	Recorded in a normal room without noise cancelation. The sampling frequency of all recording was 11025 Hz at the room temperature 27 degree. The speaker were seating in front of microphone about 10-12 cm. The database was collected in natural manner.	95.06%	It was observed that LDA with MFCC and DWT performance was found to be best with accuracy but time period is slightly greater.
8	Speakar-Independent Isolated Word Recognition using HTK for Varhadi – a Dialect of Marathi. Sunil B. Patil, Nita V. Patil, Ajay S. Patil.	Features Extraction: Mel-frequency cepstral coefficients (MFCC), Acoustical Model Generation: Hidden Markov Model (HMM) - HTK tool & Gaussian Mixture Model (GMM).	This system includes 83 isolated varhadi words with 1170 speech files. Each word is recorded in Audacity 1.3 Beta Unicode sound editor toolbox and further manually labeled in wavesurfer-1.8.8p5 toolbox.	92.77%	The system is only for Varhadi a Dialect of Marathi.
9	Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM Ms.Vimala.C, Dr.V.Radha	Pre-Processing: Mel-frequency cepstral coefficients (MFCC), Post-Processing using (HMM): Acoustic Model Phonetic Lexicon Language Model, Sphinx4.		88%	As the vocabulary used is small the system gives minimum word error rate for this system. In future, medium or large vocabulary isolated speech and continuous speech can be put into practice and can be experienced for Tamil language.
10	Development of Automatic Speech Recognition of Marathi Numerals - A Review Yogesh K. Gedam, Sujata S. Magare, Amrapali C. Dabhade, Ratnadeep	Feature extraction (Mel Frequency Cepstral Coefficients [MFCC]), feature-matching (Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization),	Audio was recorded from various speakers from the native place of Aurangabad region. Then a text corpus was developed.	48.75%	Further work can be focused on increasing the accuracy and also performance of speech recognition for Marathi digits.

	R. Deshmukh	LBG Algorithm			
11	Marathi Digit Recognition System based on MFCC and LPC features Pukhraj P. Shrishrimal , Ratnadeep R. Deshmukh , Ganesh B. Janwale , Devyani S. Kulkarni	Feature extraction - Mel Frequency Cepstral Coefficient [MFCC], Linear Predictive Coding [LPC]	The speech samples were collected from 100 native Marathi speakers.	Recognition - 99.75% Accuracy at word level - 48.75%	The noise in the audio data that was recorded for this system affected the accuracy at word level.

III) PROPOSED METHODOLOGY:

In the proposed system an voice is taken as an input, the input voice is then recognized and the predicted result is given as the output of the system. We train the dataset by using XLSR-Wav2Vec2 model. The resulting output, along with the ground truth text, is passed through the CTC layer to obtain the trained model. This trained model is then utilized for speech recognition in the input voice.

XLSR-Wav2Vec2 is fine-tuned using Connectionist Temporal Classification (CTC), which is an algorithm that is used to train neural networks for sequence-to-sequence problems and mainly in Automatic Speech Recognition.

The method of converting speech to text in marathi by combining CTC and XLSR-Wav2Vec2 techniques to train the neural network model has been effectively proposed and implemented.

3.1) Connectionist Temporal Classification (CTC)

Convolutional Neural Networks (CNNs) have proven to be highly effective in handwritten text recognition tasks. They have been widely employed for tasks like optical character recognition (OCR) and handwritten digit recognition. By leveraging its ability to learn hierarchical features from raw pixel data, a CNN effectively recognizes handwritten characters, enabling applications such as automatic transcription, form processing, and text extraction from images. This network receives an image as input, extracts feature from the image, and subsequently discerns distinct features from one another. Drawing inspiration from the interconnectedness of neurons in the human brain, CNN's learning process is driven by the training data it receives, enabling the acquisition of knowledge within the feature detection layer.

3.2) XLSR-Wav2Vec2: XLSR stands for cross-lingual speech representations and refers to XLSR-Wav2Vec2`s ability to learn speech representations that are useful across multiple languages. XLSR-Wav2Vec2 learns powerful speech representations from hundreds of thousands of hours of speech in more than 50 languages of unlabelled speech. Similar, to BERT's masked language modeling, the model learns contextualized speech representations by randomly masking feature vectors before passing them to a transformer network.

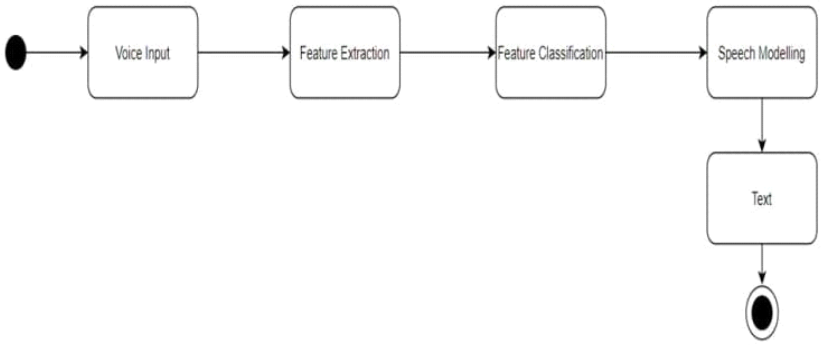
3.4) Proposed Model:

During the process of speech recognition, the user selects required voice from the device. The fact that ASR models convert speech to text necessitates the use of both a feature extractor, which converts the speech signal into the model's input format, such as a feature vector, and a tokenizer, which converts the model's output format into text.

In Transformers, the XLSR-Wav2Vec2 model is thus accompanied by both a tokenizer, called Wav2Vec2CTCTokenizer, and a feature extractor, called Wav2Vec2FeatureExtractor.

The pretrained Wav2Vec2 checkpoint maps the speech signal to a sequence of context. A fine-tuned XLSR-Wav2Vec2 checkpoint needs to map this sequence of context representations to its corresponding transcription so that a linear layer has to be added on top of the transformer block.

The output size of this layer corresponds to the number of tokens in the vocabulary, which does not depend on XLSR-Wav2Vec2's pretraining task, but only on the labeled dataset used for fine-tuning.



3.5) Dataset:

The data set contains transcribed high-quality audio of Marathi sentences recorded by volunteers. When using this model, the speech input is sampled at 16kHz.The model is trained using data containing only female voices but the model works well for male voices too.

3.6) Recognition:

After training the neural network model, it can be utilized for speech to text in marathi. The initial step involves passing the input voice through the CNN layers. After this the Transformers containing feature extractor extract the feature. These features are then mapped and the model is trained. Further the tokenizer processes the output format to text.

3.7) Software Requirements:

3.7.1) Software Requirements:

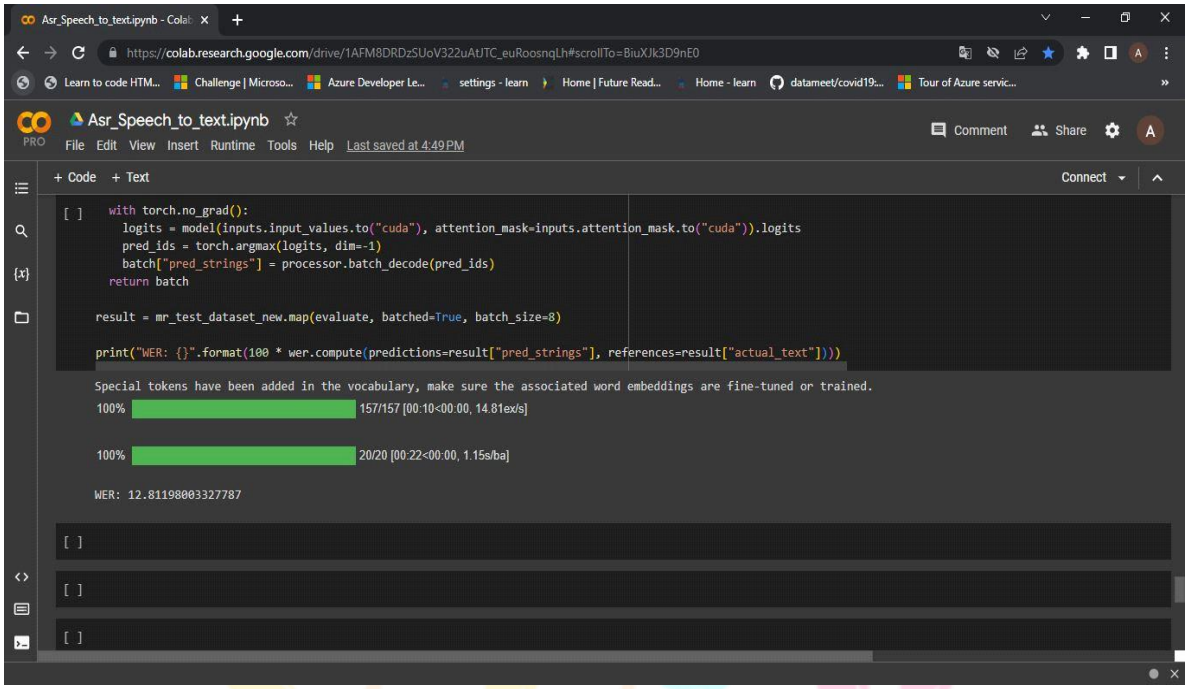
- Python: You will need Python programming language to implement and run the CTC and Wave2Vec2 model.
- Libraries: Torchaudio and librosa package to load audio files. Jiwer to evaluate our fine-tuned model using the word error rate (WER) metric.
- Google Colab: To implement the model with such large dataset using neural network Google colab is required.

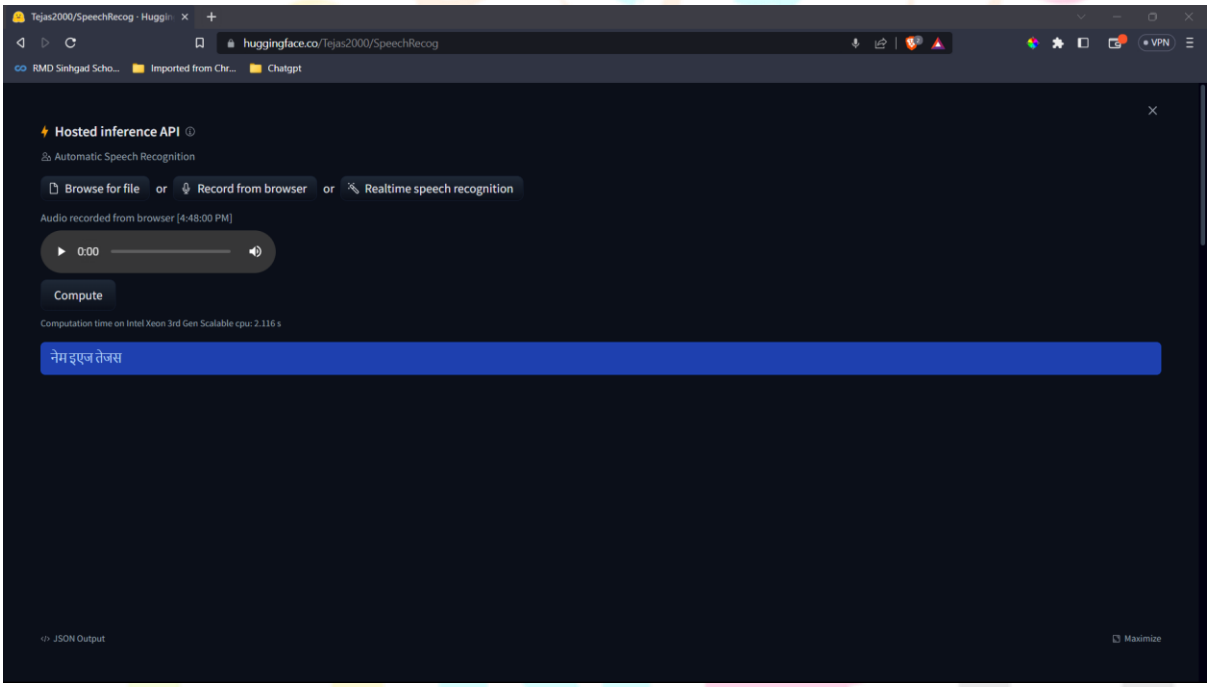
3.7.2) Hardware Requirements:

- CPU (Central Processing Unit): Having a powerful CPU helps with tasks such as data preparation, model evaluation, and deployment.
- GPU (Graphics Processing Unit): Training deep learning models, especially using transformer can be computationally intensive. Having a GPU significantly speeds up the training process.
- Sufficient RAM: Deep learning models can consume a significant amount of memory, especially when working with large datasets. Enough RAM to accommodate the data and model requirements is necessary.

IV) RESULTS

After training the model, the model is evaluated using word error metrics. The word error rate of the model is upto 12.8%.





V) CONCLUSION:

The most common and practical method of inter-person communication is speech. English, Hindi, Urdu, Arabic, Japanese, and other languages are the subject of extensive ASR research. However, Marathi, which is our mother tongue, is still at the beginning level in this area. As a result, we made an effort to research this area and provide some tools for understanding Marathi. Through the research, we tried to talk about our goals, the many tools we employed, and the speech recognition process. However, the method we have used is still at a very early stage. Numerous improvements are needed to create a good and full identification application, such as the necessity for a large training database, numerous speakers, audio with low noise levels, etc. Speech recognition software is still lacking.

VI) FUTURE SCOPE:

The proposed method can be extended to analyse multiple audio files by parallel approach. Audios in different languages can be analyzed. Various other features can be used and the method can be combined with other methods for improving the quality of speech recognition. The project can further be a base while researching the Marathi dialect. The output of this project can be further used as input to several systems.

VII) Acknowledgment

We would like to thank all the authors of the papers mentioned for their valuable information and would also like to thank all the faculty of the Computer Engineering department of RMD Sinhgad School of Engineering for their unwavering support and help during the entire journey of the project. Last, but not the least, we would like to acknowledge the invaluable love and support of our loving parents and great friends.

VIII) References

[1] Praveen Kumar, H S Jayanna, 'Development of Speaker-Independent Automatic Speech Recognition System for Kannada Language'.

[2] Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Is ¸ın Demirs ¸ahin, Cibu Johny, Martin Jansche†, Supheakmungkol Sarin, Knot Pipatsrisawat, 'Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems'.

[3] Aman Ankit, Sonu Kumar Mishra, Rinaz Shaikh, Chandraketu Kumar Gupta, Prakhar Mathur, Soudamini Pawar and Anil Cherukuri, 'ACOUSTIC SPEECH RECOGNITION FOR MARATHI LANGUAGE USING SPHINX'.

[4] Manasi Ram Baheti, Bharti W. Gawali, S.C. Mehrotra, 'Automatic Speech Recognition For Task Oriented IVRS In Marathi'.

[5] Pratik Kurzekar, Shriniwas Darshane, Nikhil Salvithal, Nitin Maske, 'Design and Development of Continuous Marathi Speech Recognition System for Agriculture Purpose'.

[6] Sangramsing N. Kayte, 'Marathi Speech Recognition System Using Hidden Markov Model Toolkit'**.**

[7] Santosh Kashinath Gaikwad, Dr.Bharti W Gawali, Suresh Mehrotra, 'Novel approach based feature extraction for Marathi continuous speech Recognition'.

[8] Sunil B. Patil, Nita V. Patil, Ajay S. Patil, 'Speakar-Independent Isolated Word Recognition using HTK for Varhadi – a Dialect of Marathi'.

[9] Ms.Vimala.C, Dr.V.Radha, 'Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM'.

[10] Devyani S. Kulkarni, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, 'A Review of Speech Signal Enhancement Techniques'.

[11] Yogesh K. Gedam, Sujata S. Magare, Amrapali C. Dabhade, Ratnadeep R. Deshmukh, 'Development of Automatic Speech Recognition of Marathi Numerals - A Review'**.**

[12] Pukhraj P. Shrishrimal, Ratnadeep R. Deshmukh, Ganesh B. Janwale, Devyani S. Kulkarni, 'Marathi Digit Recognition System based on MFCC and LPC features'.

[13] Shaikh Naziya, R.R. Deshmukh, 'Speech Recognition System – A Review'.