iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VI **Month of publication:** June 2019

DOI: http://doi.org/10.22214/ijraset.2019.6076

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

Intelligence Agent Device for E-Learning

Prof. V. S. Gaikwad, Saakshi Dixit[1], Vidya Borkar[2], Madhvi Kokil[3], Mohit Lulla[4]
[1, 2, 3, 4]*Computer Engineering Department, Savitribai Phule Pune University*

Abstract: Artificial Intelligence aims to shape our digital daily life from several aspects in the future. In this topic, we present a device which would work like an Intelligent device. The futuristic objective of this device is supposed to be made for pre-primary kids for their learning. The language preferred with this device is Marathi.
The comparative study of different techniques is done as per stages . This paper concludes with the decision on feature for developing technique in human computer interface system in either of Marathi or English language.
Keywords: Automatic speech recognition, HMM, MFCC, Language Modeling, Acoustic modeling.

I. INTRODUCTION

After years of research and development the accuracy of automatic speech recognition (ASR) remains one of the most important research challenges e.g. speaker and language variability, vocabulary size and domain, noise. The design of speech recognition system require careful attentions to the challenges or issue such as various types of speech classes, speech representation, feature extraction techniques, database and performance evaluation. This paper presents a study of basic approaches to speech recognition and their results shows better accuracy. This paper also presents what research has been done around for dealing with the problem of speech recognition.
The speech is most prominent and primary mode of communication among human being. The communication among human computer interaction is called human computer interface. This project subject gives an overview of fundamental process of speech recognition. By converting spoken audio into text, speech recognition technology let users to control digital devices by speaking instead of using conventional tools such as keystrokes, buttons, keyboards etc. This helps in recognization of human speech via microphone and helps in E- learning . That is, this device roled for E- learning and works as music player.

II. MOTIVATION

In todays technological world it is convenient for one to use the technology to manage their tasks. Our project makes user aware of solving a problem like playing a song without using **Internet.** Also we know there are some online music applications like Saavn Music, Gaana music, Wynk Music, Amazon Prime Music, etc. has millions of songs including poems and live radio which must need a strong Internet. So as lot of students or children availing these online services so this would help them to obtain the required output offline. The purpose of this project is to study the speech recognition system. Overall our goal for this project sytems is to overcome this Internet problem and to provide useful informative solution for e-learning inside house or school which will contain all the necessary poem for children and student in order to ensure easy, accurate, learning to reach their desired output without any inconvenience

III. AUTOMATIC SPEECH RECOGNITION

Speech recognition is also known as automatic speech recognition or computer speech recognition which means understanding voice of the computer and performing any required task or the ability to match a voice against a provided or acquired vocabulary. The task is to getting a computer to understand spoken language. By “understand” we mean to react appropriately and convert the input speech into another medium e.g. text. Speech recognition is therefore sometimes referred to as speech-to-text (STT). A speech recognition system consists of a microphone, for the person to speak into; speech recognition software; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation.

A. Mathematical Representation of ASR

In statistical based ASR systems an utterance is represented by some sequence of acoustic feature observations O, derived from the sequence of words W. The recognition system needs to find the most likely word sequence, and given the observed acoustic signal is formulated by:

$$W = \text{argmax}_W P(W|O) \quad \ldots. (i)$$

In "equation (i)", the argument P(W|O) i.e. the word sequence W is found which shows maximum probability, given the observation vector O. Using Baye’s rule it can be written as:

$$W = \text{argmax}_W P(W|O). P(W)/P(O) \ldots. \text{(ii)}$$

In "equation (ii)" [4], P(O) is the probability of observation sequence and is not considered as it is a constant w.r.t. W. Hence,

$$W = \text{argmax}_W P(W|O) P(W) \ldots.\text{(iii)}$$

In "equation (iii)" [4], P(W) is determined by a language model like grammar based model and P(O|W) is the observation likelihood and is evaluated based on an acoustic model. Among the various models, Hidden Markov Model (HMM) is so far the most widely used technique due to its efficient algorithm for training and recognition.

B. Typology of Speech Recognition Systems

1) Speaker Dependent: systems that require a user to train the system according to his or her voice.
2) Speaker Independent: systems that do not require a user to train the system i.e. they are developed to operate for any speaker.
3) Isolated word recognizers: accept one word at a time. These recognition systems allow us to speak naturally continuous.
4) Connected word systems allow speaker to speak slowly and distinctly each word with a short pause i.e. planned speech.
5) Spontaneous recognition systems allow us to speak spontaneously.

IV. SYSTEM ARCHITECTURE

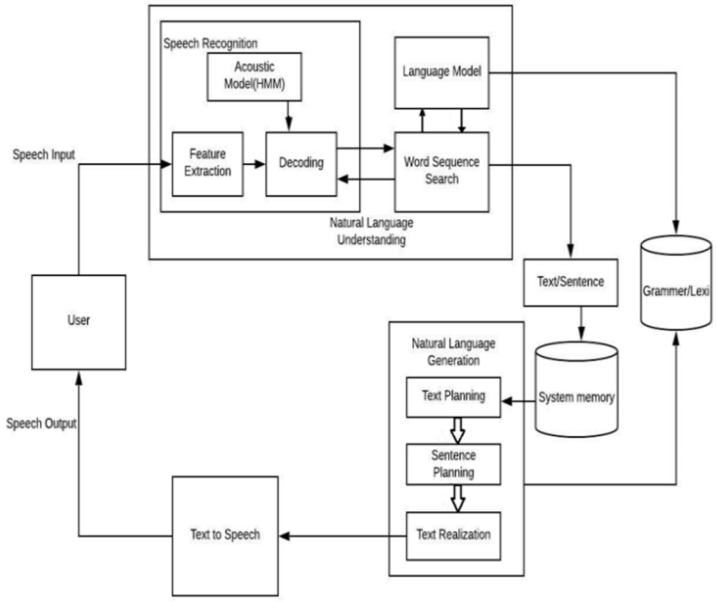


Fig. Proposed System Architecture

This architecture diagram gives us the flow and overall functionality of the system.
A speech recognition consists of four blocks : Feature Extraction, Acoustic modeling, Language modeling, Decoder. The process of speech recognition begins with a speaker creating an utterance which consists of the sound waves. These sound waves are then captured by a microphone and converted into electrical signals.
Speech signal is then converted into discrete sequence of feature vectors, which is assumed to contain only the relevant information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification such as information about speaker(e.g. fundamental frequency)and information about transmission channel (e.g. characteristic of a microphone).

1) Sensor: Using Microphone Sensor , we collect spoken audio data, and send it to the system memory via natural language understanding module.
2) System Memory: Collected data is given to the system memory for searching the song. We access it from the system memory, and use it.

A. *Feature Extraction*

First of all, recording of various speech samples of each word of the vocabulary is done by different speakers. After the speech samples are collected, they are converted from analog to digital form by sampling at a frequency of 16 kHz. Sampling means recording the speech signals at a regular interval. The collected data is now quantized if required to eliminate noise in speech samples. The collected speech samples are then passed through the feature extraction, feature training & feature testing stages. Feature extraction transforms the incoming sound into an internal representation such that it is possible to reconstruct the original signal from it. There are various techniques to extract features like MFCC, PLP, RAST, LPCC, but mostly used is MFCC.

B. *MFCC Approach*

The purpose of this module is to convert the speech waveform to some type of parametric representation. MFCC is used to extract the unique features of speech samples. It represents the short term power spectrum of human speech. The MFCC technique makes use of two types of filters, namely, linearly spaced filters and logarithmically spaced filters.

To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units mel.

The Mel scale is normally a linear mapping below 1000 Hz and logarithmically spaced above 1000 Hz.

MFCC consists of six computational steps.

1) *Step 1: Pre-Emphasis:* This step processes the passing of signal through a filter which emphasizes higher frequency in the band of frequencies the magnitude of some higher frequencies with respect to magnitude of other lower frequencies in order to improve the overall SNR. It increases the energy of signal at higher frequency.
2) *Step 2: Framing:* The process of segmenting the sampled speech samples into a small frames. The speech signal is divided into frames of N samples. Adjacent frames are being separated by M(M¡N). Typical values used are M=100 and N=256(which is equivalent to 30 m sec windowing).
3) *Step 3: Hamming windowing:* Each individual frame is windowed so as to minimize the signal discontinuities at the beginning and end of each frame. Hamming window is used as window and it integrates all the closest frequency lines.
4) *Step 4: Fast Fourier Transform:* To convert each frame of N samples from time domain into frequency domain FFT is applied.
5) *Step 5: Mel Filter Bank Processing:* The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale is performed.
6) *Step 6: Discrete Cosine Transform:* This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficients. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector. The spacing of the filter bank is inspired by the human auditory system and the relation between linear frequency and Mel frequency can be described with the mathematical equation as:

$$\text{Mel}(f) = 2595 * \log10(1 + f / 700)$$

Here, f is normal frequency and Mel (f) is the output frequency of Mel scale. Usually, it covers the frequency range from 156 Hz to 6844 Hz and it is logarithmic above 1 KHz, and linear below the range of 1 KHz frequency. Normally, first 13 coefficients are wide enough to represent the speech signal.

C. *Decoding*

It is the most important step in the speech recognition process. Decoding is performed for finding the best match for the incoming feature vectors using the knowledge base. A decoder performs the actual decision about recognition of a speech utterance by combining and optimizing the information conveyed by the acoustic and language models. This uses theory from statistics in order to(sort of) arrange our feature vectors into a Markov matrix (chains) that stores probabilities of state transitions. That is, if each of our code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state.

D. *Acoustic Modeling*

An acoustic model is implemented using different approaches such as HMM, ANNs, dynamic Bayesian networks (DBN), support vector machines (SVM). HMM is used in some form or the other in every state of the art speech and speech recognition system.

E. Hidden Markov Modeling Approach

A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters; the challenge is to determine the hidden parameters from the observable data. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.HMM creates stochastic models from known utterances and compares the probability that the unknown utterance was generated by each model. This uses theory from statistics in order to (sort of) arrange our feature vectors into a Markov matrix (chains) that stores probabilities of state transitions. That is, if each of our code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state.

HMMs are more popular because they can be trained automatically and are simple and computationally feasible to use HMM considers the speech signal as quasistatic for short durations and models these frames for recognition. It breaks the feature vector of the signal into a number of states and finds the probability of a signal to transit from one state to another. HMMs are simple networks that can generate speech (sequences of cepstral vectors) using a number of states for each model and modeling the short-term spectra associated with each state with, usually, mixtures of multivariate Gaussian distributions (the state output distributions). The parameters of the model are the state transition probabilities and the means, variances and mixture weights that characterize the state output distributions

F. Language Modeling

Language models are used to guide the search correct word sequence by predicting the likelihood of nth word using (n-1) preceding words.

Language models can be classified into:

1) *Uniform Model:* each word has equal probability of occurrence.
2) *Stochastic Model:* probability of occurrence of a word depends on the word preceding it.
3) *Finite state Languages:* languages use a finite3 state network to define the allowed word sequences.
4) *Context free Grammar:* It can be used to encode which kind of sentences are allowed.

V. APPROACHES TO SPEECH RECOGNITION

There are three types of approaches to ASR. They are:

A. Acoustic Phonetic Approach

Acoustic phonetic approach is also known as rule-based approach. This approach uses knowledge of phonetics & linguistics to guide search process. There are usually some rules which are defined expressing everything or anything that might help to decode based in “blackboard” architecture i.e. at each decision point it lays out the possibilities and apply rules to determine which sequences are permitted. It has poor performance due to difficulty to express rules, to improve the system. This approach identifies individual phonemes, words, sentence structure and/or meaning.

B. Pattern Recognition Approach

This method has two steps i.e. training of speech patterns and recognition of pattern by way of pattern comparison. In the parameter measurement phase (filter bank), a sequence of measurements is made on the input signal to define the “test pattern”. The unknown test pattern is then compared with each sound reference pattern and a measure of similarity between the test pattern & reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase (dynamic time warping).

C. Artificial Intelligence Recognition Approach

This approach is a combination of the acoustic phonetic approach & the pattern recognition approach. In the AI, an expert system implemented by neural networks is used to classify sounds. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand.

VI. CONCLUSION

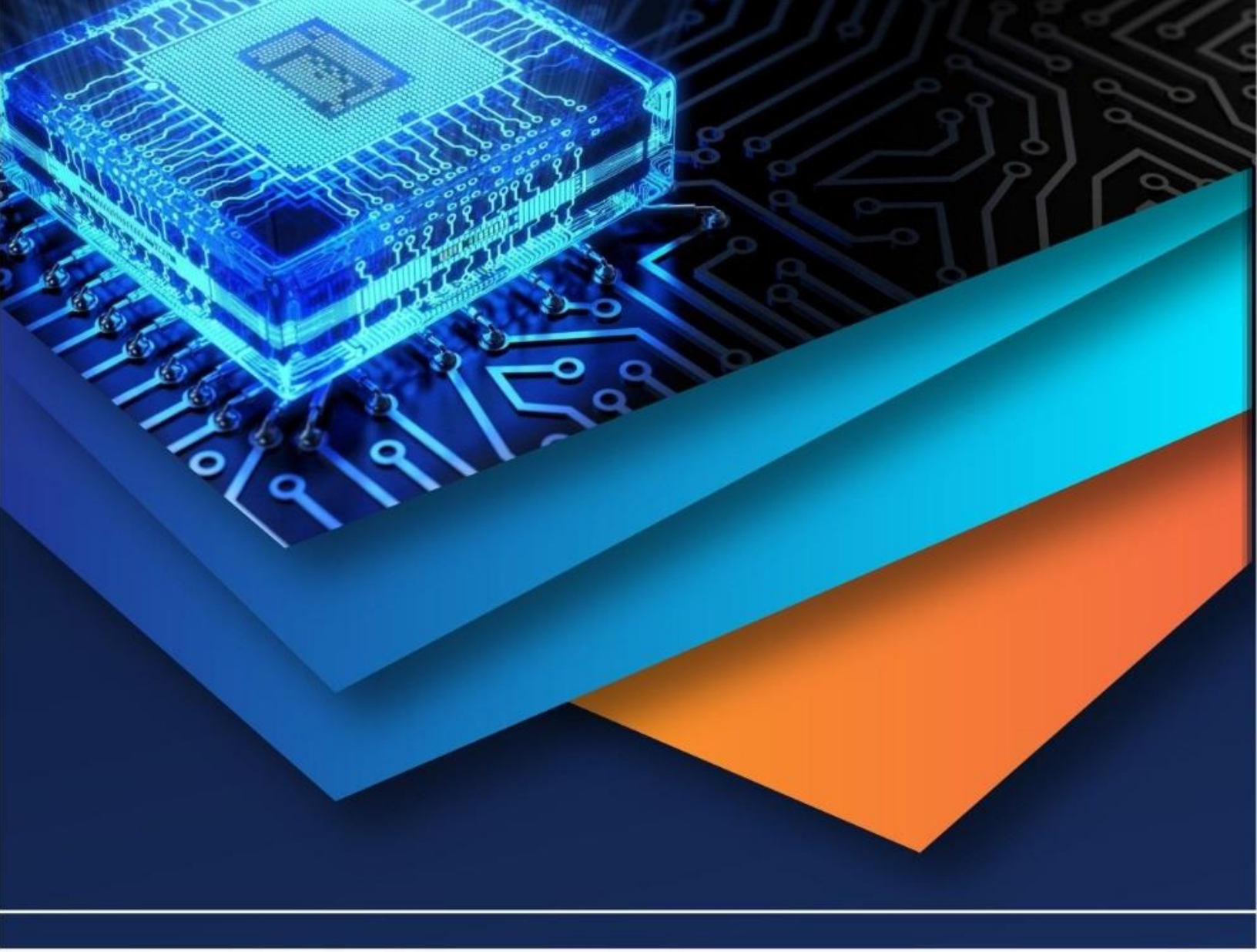
Speech recognition is a challenging problem to deal with. We have attempted in this project to provide a review of how much this technology has progressed in the previous years.

The proposed system performs actions based on the user's voice input and performs the actions to put the best relevant data in terms of audio. There are many advantages with the proposed system when compared with the traditional system. The advantage include less cost, low power, high accuracy, low power consumption and less analysis time.

The Internet also makes it much easier for students to cheat on their studies, find others on the Internet to write reports, listen songs, watch videos, do their homework. We can say students or also children can access the Internet and go through the search engine via recorder (microphone) by speaking name of any song or poem to get desired output. But our project system is providing useful informative solution for e-learning inside house or school which will contain all the necessary poem for children and students in order to ensure easy, accurate learning to reach their expected output without spending a lot of frivolous time on Internet where they will be focused on studies.

REFERENCES

[1] Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition, Ji Ming ,Member, IEEE, Danny Crookes, 2016
[2] Ahmad A. M. Abushariah,Teddy S. Gunawan, Othman O. Khalifa"English Digits Speech Recognition System Based on Hidden Markov Models", International Islamic University Malaysia, International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia
[3] Crosslingual amd Multilingual Speech Recognition Based on speech manifold, Reza Sahraeian and Dirk Van Compernolle, 2017.
[4] Automatic Speech Recognition, Preeti Saini, Parneet Kaur, 2013.
[5] Speech RecognitionUsingHMM, Ms. Rupali S Chavan, Dr. Ganesh. S Sable, 2013.
[6] Ibrahim Patel,Dr.Y.Srinivasa Rao, , "Speech recognition using Hidden Markov Model With MFCCSubband Technique." 2010 International Conference on Recent Trends in Information,Telecommunication and Computing.
[7] Vimala C, Dr.V.Radha, "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 1, 1-7, 2012
[8] D. O'shaughnessy, "Acoustic analysis for automatic speech recognition," Proceeding of the IEEE, vol. 101, no. 5, May 2013.
[9] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, pp. 1283–1294, 2015
[10] J. Ming and D. Crookes, "Wide mathcing – an approach to improving noise robustness for speech enhancement," in Proceedings Int. Conf. Acoust., Speech, Signal Process. IEEE, 2016, pp. 5910–5914.
[11] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, pp. 967–977, 2016.
[12] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, pp. 982–992, 2015.
[13] M. Delcroix et al., "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," Comput. Speech Lang., vol. 27, pp. 851–873, 2013.
[14] Ankit kuamr, Mohit Dua and Tripti Choudhary, " Continuous Hindi Speech Recognition Using Gaussian Mixture HMM ",2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY