SPEECH RECOGNITION BASED LEARNING SYSTEM

Lavin Jalan[I], Rahul Masrani[II], Roshan Jadhav[III] Tejaswini Palav[IV]

*[#]VIDYAVARDHINI'S COLLEGE OF ENGINEERING AND TECHNOLOGY-MUMBAI UNIVERSITY
VIDYAVARDHINI'S CAMPUS ,K.T Marg, Vasai road(west), Thane-401202*

***Abstract*— This paper presents an English dictionary consisting of accents and meanings of different words, which is voice operated i.e. operated on speech input from user which will be in the form of individual alphabets. On spelling the individual alphabets, the user will be provided with the accent and meaning of the word formed from the alphabets spelled by him, which will be in the audio format.**

***Keywords*— Mel frequency cepstral coefficients (MFCC), Vector quantization (VQ), End point detection (epd), codebook, automatic speech recognition (ASR).**

I. INTRODUCTION

English is an internationally accepted language with large number of phonemes and is being used with various accents in different parts of the world. To understand the language irrespective of the accents, the study implements a dictionary based on speech recognition using isolated characters to provide the exact meaning of the spoken word (in terms of isolated characters) with high accuracy. The entire process of speech recognition is carried out in MATLAB. The entire process to identify the pronounced character is performed in three steps. The first step performs the endpoint detection by using short-term temporal analysis. The second step includes speech feature extraction using MFCC (Mel-Frequency Cepstral Coefficients) parameters and third step is codebook generation of each of the characters using Vector Quantization LBG algorithm [Linde, Buzo and Gray] where we get the output in the form of characters on the matlab command window. These characters will be combined together to form a meaningful word. And that word will be compared with the pre-prepared database to get the audio output.

This will be highly useful in learning. One of the main advantages of our project is that it will be very useful for blind people. It can also be viewed as a unique approach in the varied domain of human computer interface.

II. SPEECH RECOGNITION ALGORITHMS[1][2][4]

Today's vocal recognition systems are based on the general principles of forms' recognition. The methods and algorithms that have been used so far can be divided into four large classes:

Discriminant Analysis Methods based on Bayesian discrimination;

1. Hidden Markov Models (HMM);
2. Dynamic Programming –Dynamic Time Warping algorithm (DTW);
3. Neural Networks.

A. *Hidden Markov Models (HMM):*

It's a mathematical model derived from a Markov Model. Speech recognition uses a slightly adopted Markov Model. Speech is split into the smallest audible entities (not only vowels and consonants but also conjugated sound like ou,ea,au,…).All these entities are represented as states in the Markov Model. As a word enters a Hidden Markov Model it is compared to the best suited model (entity).

B. Dynamic Time Warping algorithm (DTW):

Dynamic Time Warping algorithm (DTW) [Sakoe, H. & S. Chiba-8] is an algorithm that calculates an optimal warping path between two time series. The algorithm calculates both warping path values between the two series and the distance between them. The algorithm starts with local distances calculation between the elements of the two sequences using different types of distances. The most frequent used method for distance calculation is the absolute distance between the values of the two elements (Euclidian distance). That results in a matrix of distances having n lines and m columns of general term:

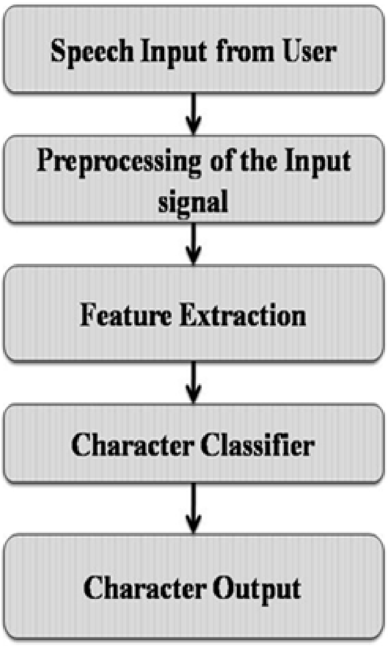
$$d_{ij} = \left|a_i - b_j\right|, \ i = \overline{1,n}, \ j = \overline{1,m}.$$

C. *Neural Networks:*

Neural networks have much similarity with Markov Models. Both are statistical models represented as graphs. Where Markov Models use the probability for state transitions, neural networks use connection strengths and functions. A key difference is that neural networks are fundamentally parallel while Markov chains are serial. Frequencies and speech occur in parallel, while syllable series and words are essentially serial. This means that both techniques are very powerful in different aspects.

As in the neural networks, the challenge is to set the appropriate weights of the connections, the Markov Model challenges is finding the appropriate transitions and observations possibilities. In many speech recognition systems both techniques are implemented together and work in symbiotic relationship. Neural networks perform very well at learning phoneme probability from highly parallel audio input, while Markov Model can use the phoneme observations probabilities that the neural networks provide to produce the likeliest phoneme sequence or words. This is at the core of a hybrid approach to natural language understanding.

III. OUR PROJECT---DESIGN AND DEVELOPMENT



I. Speech Processing Model [1]

Speech recognition or speech to text conversion involves capturing and digitizing the sound waves, converting them into basic language units or phonemes, constructing characters from phonemes, and contextually analyzing the characters to ensure correct detection of characters that sound like.

This process can be elaborated in following steps:

A. Input From The User:

User will spell out the individual alphabets of the word through the microphone. This is the word for which the user wants to know the accent and the meaning. Pronouncing the word in form of character limits the probability of false identification of any desired word and thus improves the accuracy.

B. *Preprocessing[1][5]:*

Pre-processing is the important step which makes the signal suitable for further processing. A speech file of the phonemes pronounced by the user is extracted from the audio file and is down sampled to 11 kHz.

C .Feature Extraction [1][5]:

Feature extraction involves the mining of useful amount of information required to describe a large set of data accurately. Apart from the message content, the speech signal also carries variability such as speaker characteristics, emotions and background noise.

A method of generating feature signals from speech signals comprises of the following steps:

1. Receive the speech signal which is in analog format and stored as a .wav file.
2. Convert the .wav file to a data file in MATLAB using the *wavread* function for further processing.
3. Block the signal into frames.
4. Form the frequency domain representation of the said blocked speech signal.
5. Pass the said frequency domain representation through Mel-filter banks.

Also, the system is designed to be **speaker independent**. Hence, it provides a greater flexibility in terms of usage by any person and eliminates the time required for training. But at the same time it provides a reduced accuracy and limits the recognition to a few limited characters/words.

D. End-Point Detection [2][3]:

After accepting the speech input from user, the next step is to process it. An important problem in speech processing is to detect the presence of speech in a background of noise. This problem is often referred to as the end point location problem. To determine the end point of the speech signal, we have considered the 'Teager' energy approach which considers both energy and frequency characteristics of the speech signal. It can be very effective when the speech signal is very weak but has frequency components. different than background noise. Therefore, it is a very useful tool to locate the beginning and ending point for an utterance [1, 2, 3, 5].

E. Generation of Mel Frequency cepstral coefficients (MFCCs)[4][5]:

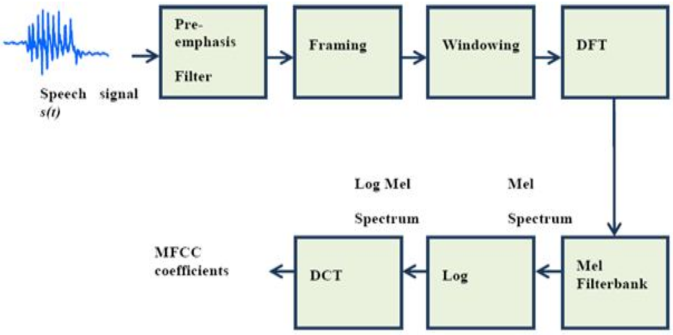
Considering the known variation of the ear's critical band-width frequency, for speech recognition filters spaced **linearly at low frequencies** and **logarithmically at high frequencies** have been used to capture the phonetically important characteristics of speech. This representation would be provided by a set of Mel-Frequency Cepstrum Coefficients (MFCC). These coefficients collectively make up an MFC (Mel Frequency Cepstrum) (a nonlinear "spectrum-of-a-spectrum").

In MFCC, the main advantage is that it uses Mel frequency scaling which is very approximate to the human auditory system.

The sequence of processing includes for each chunk of data and basic steps are

1. Window the data with a hamming window.
2. Shift it into FFT order.
3. Find the magnitude of FFT.
4. Convert the FFT data into filter bank outputs.
5. Find the log base 10.
6. Find the cosine transform to reduce dimensionality.

A block diagram of the structure of an MFCC process is shown in the figure below.



II. Mel Frequency Cepstral Coefficients [6]

F. Frame Blocking [4][5]:

In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N -M samples. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100.

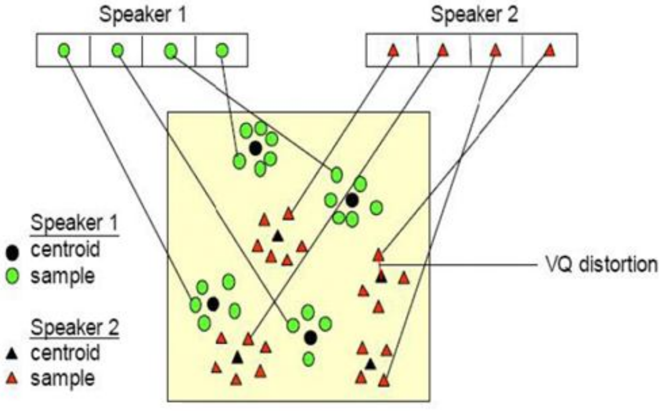
G. Windowing[4][5]:

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. Typically the Hamming Window is used. The coefficients of a Hamming window are computed from the following equation:

W[k+1]=0.54-0.46cos(2*pi*(k/(n-1)),k=0,1…,n-1.

IV .VECTOR QUANTIZATION[5][6]:

The next step after the generation of MFCCs is identification of the character pronounced by the user. The state-of-the-art in feature matching techniques used in speaker recognition or character detection includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). Of these techniques, VQ is easiest to implement and provides high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook.



III. Vector Quantization [6]

The figure above shows a conceptual diagram to illustrate the recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result code words (centroids) are shown in the figure by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

V. OUTPUT DATABASE [2]

The sound database i.e. predefined wave files consisting of the accent and meaning of different words can be made in “JAVA” or “MY-SQL”. By importing the wave file in MATLAB one can easily get the accent and meaning o spoken word in the audio format. This can be very helpful for blind people. Our work consists of the database of words, their accents and meaning but the dictionary can be defined as per user requirement. It can be extended as and when required. This flexibility regarding the size and the type of dictionary provides a unique touch to our idea.

VI. APPLICATIONS:

A. SPEECH RECOGNITION BASED ENGLISH DICTIONARY:

One way to communicate with your computer is to speak to it. The successful implementation of our project will form software with some time and patience to operate it. We are training our software to recognize a text which we speak and command that we issue. Success in using this software depends upon suitable hardware, training and technique. In this software, user will spell out the alphabets of a word and our training on a software will convert these voice formatted isolated alphabets into a text word. The main reason behind dealing with isolated alphabets instead of direct word is to avoid glitches in the pronunciation of similar sounding words like "there", "their" and "they're", "here" and "hear", and even "youth in Asia" and "euthanasia", and 1,000 other examples which all sound very similar. Also as the name of our project suggested it's a learning system. We are assuming that user doesn't know the meaning and pronunciation of a word and with this reason user is using our learning system. After forming a word it will display the meaning of the same also its proper pronunciation along with its meaning in the voice format. So, it will be user friendly software for visually impaired learners and the best application for blind people.
This combination does not give us completely hands-free use of our computer but can offer substantial relief from keyboarding. We are also working on the different techniques which will make our software user-friendly for physically impaired people also. We are trying to eliminate the use of keyboard.

B. SPEECH RECOGNITION BASED TALKING CALCULATOR:

The main idea behind this application originates from a question, Is there anything that makes calculation easy for blind people for blind people?
We are trying to make a talking calculator in which user will get arithmetic calculations in the form of voice. User will say a number (it can be of 4or 5 digits) and operator (+,-,*, /). The answer of the calculation will get converted into an audio signal and software will give the answer to the user in the form of audio along with the text. This will be an excellent application for the visually impaired people who can't use normal calculator.

C. OTHER THAN LEARNING SYSTEM:

Voice controlled robot is an interesting application, mainly used for industrial and surveillance applications. It gives exact concept of controlling a robot by a voice command. Robot is capable of understanding and synthesizing human speech for communication. A voice recognition unit built around a high speed processor that ensures various operations of the system to be performed by voice commands. Few commands recommended for operation which we can train are listed as: START, STOP, FORWARD, REVERSE, RIGHT, LEFT, SLOW, FAST, OK, UP, DOWN, CLOCK, ANTICLOCK, CLOSE and OPEN.

VII. FUTURE SCOPE:

Speech recognition, also referred to as speech-to-text or voice recognition, is technology that recognizes speech, allowing voice to serve as the "main interface between the human and the computer". Our project discusses how current speech recognition technology facilitates student learning, as well as how the technology can develop to advance learning in the future.
Although speech recognition has a potential benefit for students with physical disabilities and severe learning disabilities, the technology has been inconsistently implemented in the classroom over the years. As the technology continues to improve, however, many of the issues are being addressed. If one hasn't used speech recognition with the students lately, it may be time to take another look. Both Microsoft and Apple have built speech recognition capabilities into their operating systems, so one can easily try out these features with your students to find out whether speech recognition might be right for them.
Though our project aims on building a dictionary of English words its scope can be further extended by developing a dictionary of sentences which can be the definitions of scientific terminologies or meaning of "tedious looking" jargons. This can be viewed as an extension to our project in terms of easy computer based learning mechanisms. A student can have a quick revision of all the technical terms with its meaning by just spelling the alphabets of the desired word or a jargon. By pre-preparing the wave files of the required words, the student actually creates a revision module for his own reference. In today's "technology dominant" world of varied human computer based learning systems this can be a major contribution. Now-a-days education systems are more inclined to computer based learning and e-learning based can have a quick revision of all the technical terms with its meaning by just spelling the alphabets of the desired word or a jargon. By pre-preparing the wave files of the required words, the student actually creates a revision module for his own reference. In

today's "technology dominant" world of varied human computer based learning systems this can be a major contribution. Also, according to the researches done by American universities "learning by listening" is the most effective method and the development of such a creative dictionary can be a major and an innovative contribution. As it is rightly said," technologies are developed by simplest of the minds", our simple idea of developing an simple English based dictionary can be a booster in the field of creative learning methodologies and unique touch to the wide and varied domain of speech recognition and human computer interfacing.

VIII. A UNIQUE TOUCH

The project is an asset to blind people as the output is in audio format. Also the dictionary is flexible. The work done in this field was for dictionary of words and their meanings but our dictionary can be modified as per user requirement. A student may create a dictionary consisting of meaning or explanation of certain scientific terminologies. The next time when he refers it he just has to spell the alphabets of the desired words and he is provided with the explanation. The dictionary thereby provides a unique way of learning in the interactive voice learning systems. The dictionary can also be developed for different languages like FRENCH, SPANISH etc. Similar systems can also serve as a "query assistant" for airports answering the queries of passengers regarding their flights. This can save a lot of time and the passengers can get immediate response by just spelling their flight name alphabetically. Thus our learning system is helpful and different not only for normal people but also for physically disabled people, especially blind people. Further research can be done for implementing this learning module as an application for "android phones" thereby making it user friendly.

IX. REFERENCES:

1. S. F. Surui," Speaker independent isolated word recognizer using dynamic features of speech spectrum", IEEE Trans.ASSP,volume-34,1986.
2. J.R.Deller,Ja.G.Proakis, J.H.L.Hanson, "Discrete Time Processing of speech signal,MaccMilan Publishing Company 1993.
3. Lingyun Gu and Stephen A. Zahorian, a new robust algorithm for isolated word endpoint detection.
4. Y.EphirimA.Dembo L.R.Rabinar,"A minimum discrimination approach for hidden markov models",IEEE Transactions on Information Theory vol.35 September 1989.
5. Kumbharana, Chandresh K., 2007, "Speech Pattern Recognition for Speech to Text Conversion", thesis PhD, Saurashtra University.
6. Linde Y., Buzo A. and Gray A. M., "An algorithm for Vector Quantization", IEEE Transactions on Communication, vol 28., No. 1, 1980.