International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Digit Speech Recognition using Hidden Markov Model Toolkit

Chayan Paul, Pronami Bora

Abstract: ***Digit speech recognition refers to the task of identifying the English digit spoken in a particular utterance by an unknown speaker. The conventional methods used for the recognition of digits in speech are based on robust pattern recognition techniques which deal with the statistical parameters of speech. HMM, GMM and dynamic programming techniques are some of the methods. This paper presents recognition of digits using HTK Toolkit which is based on Hidden Markov Model using MFCC. The digit speech database of this work was collected in real time from both male and female speakers and the transcription of the total collected data was done using Wavesurfer.***

Keywords: HMM, MFCC,

I. INTRODUCTION

Speech can be defined as the communication or expression of thoughts in spoken words. Human speech can be categorized into voiced, unvoiced and silence region. When input excitation seems to be periodic impulse sequence, then visual representation of the corresponding speech can be seen as periodic when visualized, and hence known as voiced speech. Contrary, in case the input seems to be random, then the visualization is also expected to be random without any periodic nature, thus these are called unvoiced speech. The process of speech production includes creating both voiced and unvoiced speech one after another. In this process the voiced and unvoiced speech are separated by regions called silent regions [1]. In the silent region, generally excitation signals are absent in the vocal cord, which eventually results in absence of speech output. Still the silence regions remain an important part of the signal, especially between voiced and unvoiced signals. In absence of the silence regions, processing of speech signal will not be possible. The major tasks in speech processing is to develop computer program or algorithms, which can be utilized for classifying words or phrases in spoken language and then convert them to a digital format. Speech processing has got a wide variety of application and mainly being used to replace the mundane tasks like, typing or selecting in any other ways [2]. Speech processing can facilitate software and systems which are more user friendly and efficiently they can perform their tasks in the respective fields.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Chayan Paul, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.
Pronami Bora, Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.

As the sharing of information across the globe is increasing drastically, the demand for breaking the language barrier across the different regions is also increasing. Hence it gives rise to a new challenging area of study called Automated Speech Recognition. One of the important criteria in processing the speech by machine is that the machine or the algorithm must be able to process the individual digits. Hence digit recognition remains as one of the important sub area of speech recognition. The major challenge in digit recognition is to develop efficient system or algorithms which can recognize the digits in spoken words.

One of the methods which is very widely used in the area of extracting spectral features is Mel-Frequency Cepstral Coefficients (MFCC)[3]. MFCC takes human perception sensitivity with respect to frequencies in consideration, and therefore are best for speech recognition[6]. This paper presents the use of Hidden Markov Model (HMM) and is implemented using the Hidden Markov Model Toolkit. A hidden Markov model (HMM) is a Bayesian network which have mostly been applied for digit recognition in speech signals. It is a Markov process which is used for modeling of systems that are changed randomly. Markov model takes into consideration only the current events to generate a new sequence of random but related events. The basic idea of HMM theory was developed by Baum et al. in 1967. HMM has two types namely; left to right HMM and Ergodic as shown in figure 1 and figure 2.

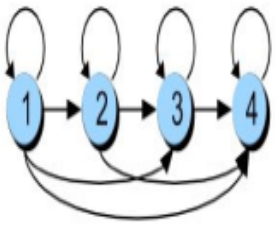

Figure 1: Left to right HMM

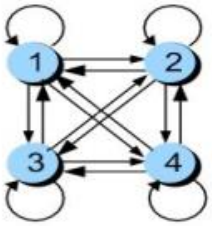

Figure 2: Ergodic HMM

Hidden Markov Model Tool Kit (HTK) is a portative toolkit used for Hidden Markov models training and testing. It was designed by the Machine Intelligence Lab in Cambridge University Engineering Department. HTK has its uses also in the areas of speech synthesis, character recognition and DNA sequencing. HTK comes with a bundle of libraries and tool written and available in C which allows advanced facilities for analysis of speech[4]. There has been a considerable amount of improvement in the accuracy and acceptance of speech recognition system in the recent past and this has got a wide application in the modern day tools and techniques.

Speech recognizers will be beneficial and will find a large number of ap-plication especially with the people with disabilities and impairment. Some of the important applications of speech recognizer are passwords recognized by voice,
Call distribution by voice commands, speech to text processing, automated data entry, in car systems, medical documentation, high performance aircraft, hands-free-computing, home automation, telematics etc.

II. SPEECH RECOGNITION SYSTEM

A. Classification of Speech Recognition System

Speech recognition is very complex and difficult task due to variability in signals and can be classified into different types of classes depending on various channel types, speaker models, vocabularies and speech utterances. Depending upon types of speech utterances they are isolated words, connected words, continuous speech and spontaneous speech. Connected word system appear to be similar to isolated words except it allows separate expressions to be run-together with some minimum pause between them. While computer determines the contents, continuous speech recognizers let users to speak almost naturally. As detecting word boundary is difficult, hence it is difficult to develop continuous speech recognition system. As vocabulary grows larger, confusion between different word sequence grows. Spontaneous speech may include mispronunciations, false-starts and non-words. Isolated words recognizers generally require every word to be quiet on both sides of sample window and this makes its implementation easy because of observable boundaries of words and requirement of clear pronunciation of words. The general processing steps involved in development of digit speech recognition are Feature extraction, Training Phase and Testing phase, and Performance Analysis.

B. Feature Extraction Process

It is the process of extracting the features from a speech signal which can help in recognizing the speaker in speech recognition technique. Some of the criteria which should be satisfied by the extracted feature while dealing with speech signal are:

- Easy to measure extracted speech features.
- Should not be susceptible to mimicry.
- Should be stable over time.
- Should occur frequently and naturally in speech.

Linear predictive coding, mel frequency spectrum, mel frequency cepstrum coefficients, rasta filtering are some of the conventional techniques used for feature extraction. In this paper, MFCC has been used as it gives higher accuracy than the time domain characteristics and it extracts both linear and non-linear factors. Training model is prepared from the extracted features which may be different for different system designs. The true identification of the spoken digits and the feature vectors of speech signals which are sampled are the simplest form of training. Sampled speech segments from certain target digits are used to train a model that represents the target language.

C. Performance Analysis

Speed and accuracy are the most widely used criteria for determining the performance of a speech recognition system. Accuracy can be measured by the commonly used metric called as the word error rate (WER), and speed is measured by the real time factor. WER can be expressed using the following formula.

$$\text{WER} = \frac{S+D+I}{N}$$

where S is the number of substitutions, D is the number of the deletions, I is the number of insertions and N is the number of words in the reference

III. METHODOLOGY

A. Extraction of Mel Frequency Cepstrum Coefficient

One of the most widely used methods to extract spectral features is Mel Frequency Cepstral Coefficients (MFCC). This is a frequency domain feature representation and it has a lot more accuracy than time domain characteristics. The different steps involved in this process can be represented by figure 3.

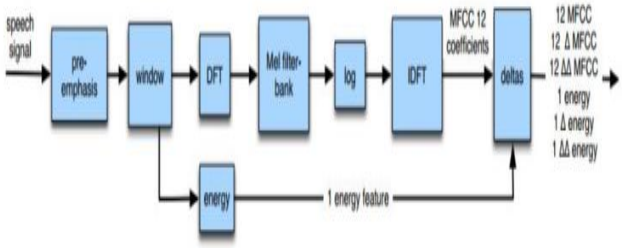

Figure 3: MFCC Extraction Process

In this paper, the pre-processing of speech signal is done where it is segmented into successive frames of 25ms each with a frame shift of 10ms overlapping on each other. This overlapping of the frames makes a smooth transition from one frame to another and each frame is multiplied by a window function to eliminate discontinuous at the edges. The waveforms after pre-processing, pre emphasis and windowing can be shown in figure 4 and figure 5.

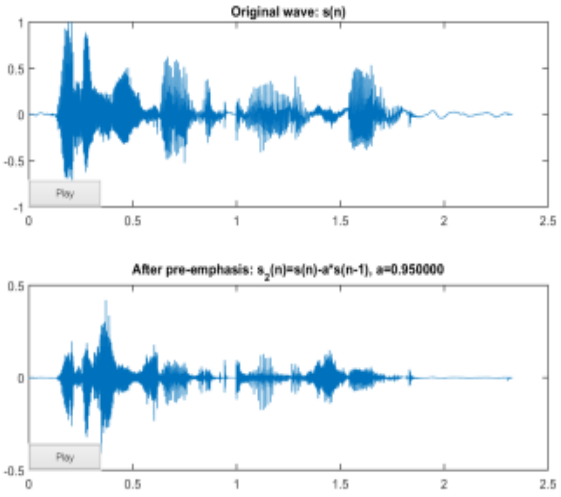

Figure 4: Speech Signal After Pre Processing

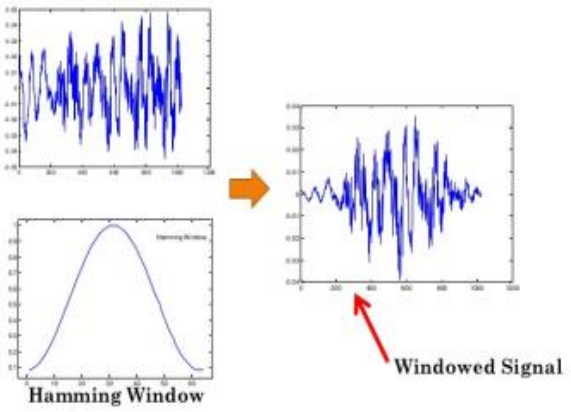

Figure 5: Windowed speech signal

The next step is estimation of discrete cosine transform (DCT) of the output of the filter bank followed by calculation the MFCC for each frame. In this process we get a set of coefficient, known as acoustic vector. The acoustic vector characterizes the phonetic characteristics of speech. This vector has significance for further analysis and processing.

B. HMM for Digit Recognizer

Hidden Markov Models are generally used in most of the speech recognition research as acoustic models. One of the main reasons is, speech signals vary a lot in time and signal and HMMs are found to be best suited to represent this type of signal. It was found that the acoustic units of any language is corresponds better with the states of the HMM and state transitions of HMM are found to match with the spectral vector representation of any language. Reasons of HMMs popularity is because training an HMM can be done automatically and HHMs are relatively simple. The main advantage of HMM is that it reduces the time and computational complexity of recognition process, while training large vocabularies [5].

While training, each of the speakers is represented by a single HMM and training feature vectors are used. We represent the parameters of HMM by state-transition probability distribution. This can be stated by A={aij} where {aij}=P[qt+1=j|qt=i], 1≤i, j≤N represents the probability of transition from state i to j at time t. The following matrix represents a 3 state left right model state transition matrix:

$$A=\{a_{ij}\}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

The state transition matrix of three state ergodic model is given by

$$A=\{a_{ij}\}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

B = {bj(k)} represents the observation probability where $b_j(k) = P[o_t = v_k | q_t = j]$, and M>=k>=1 represents the symbol distribution in state j, j = 1,2…N.

$\pi = \{\pi_i\}$ represents the initial state distribution, where $\pi_i = P[q_1 = i]$, $1 \le i \le N$. N represents number of states, and q_t represents the state at time t, M represents number of distinct observation symbols every state, and o_t is the observation symbol at time t.

During testing, P (O/λ) for each model is calculated, where $O = (o_1, o_2, \ldots o_T)$ is the sequence of the test feature vectors. We can represent the model parameters using $\lambda = (A_i, B_i, \pi_i)$ for i = 1,2,…M. In a digit recognizer model, each of the digit can be represented using an HMM. Each of the digits is referred to by the λ value of each model. Main aim is to obtain the probability of a model given a particular digit. If a model gets the highest probability, this digit is identified by that digit.

C. Development of Digit Recognizer using HTK

Hidden Markov Model Tool Kit (HTK) is a tool commonly used in the area of speech recognition research. Use of HTK can also be found in number of other applications such as speech synthesis, character recognition and DNA sequencing. The general processing steps involved in development of digit recognizer are:

- Data preparation and task definition.
- Acoustic analysis
- Training phase, and testing phase.

The first step in digit recognizer is the transcription of collected speech data using IPA symbols followed by merging the transcribed data which have similar sounds and less occurance. After merging digits of similar sounds, the digits should be assigned to ASCII codes, while the silence symbol can be denoted by sil. Using these ASCII codes altogether with sil, the basic architecture of the recogniser that is the model (the task grammar) and pronunciation model (task dictionary) can be created. The HTK recogniser requires the task grammar in Standard Lattice Format (SLF). The Task grammar is converted to Task Network which is in SLF using HTK Tool HParse as represented in figure 6. During the acoustic analysis (figure7), the signal is segmented in successive frames of 25ms with a frame shift of 10ms. Each frame is then multiplied by a Hamming window. From each window frame a vector of acoustic coefficients is extracted, which gives a compact representation of the spectral properties. Each feature vector consists of an energy coefficient, 12 MFCCs, 13 delta coefficients and another 13 acceleration coefficients respectively. These 39 coefficients give the vocal tract information of the speaker. The configuration file (.conf) is the text file which is used for setting the parameter for extraction the MFCCs coefficient.

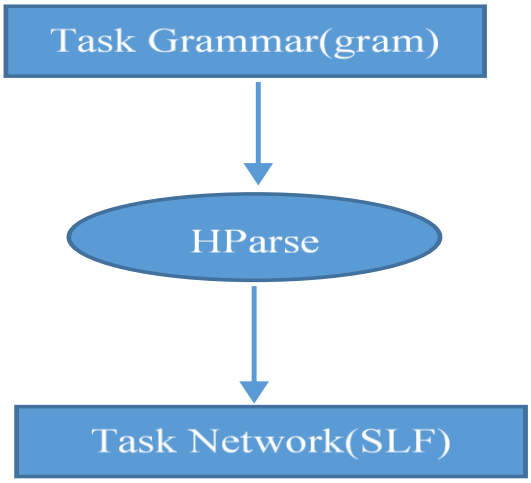

Figure 6: Conversion of Task Grammar to Task Network

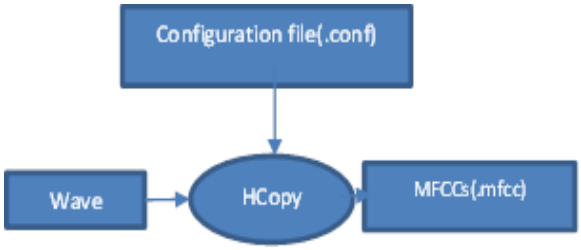

Figure 7: Acoustic Analysis of Digit

In the training phase as shown in figure 8, 5 states left to right HMM with 32 mixture continuous density diagonal covariance Gaussian mixture model (GMM) per state for each phonetic unit is used. The first and last states are non-emitting states and remaining 3 states are emitting states. The pre-defined prototype along with acoustic vectors and transcription of training data are used for initialization which calculates the global speech mean and variance of HMMs per state. Once an initial set of models has been created, the optimal values for the HMM parameters (transition probability, mean and variance vectors for each observation function) are estimated. To make the system more accurate, the HMMs are refined by incrementing the mixture size. After increasing the mixture size, the optimal values of all the total HMMs are re-estimated. This re-estimation process can stop when the performance of the system remain constant after some iteration.

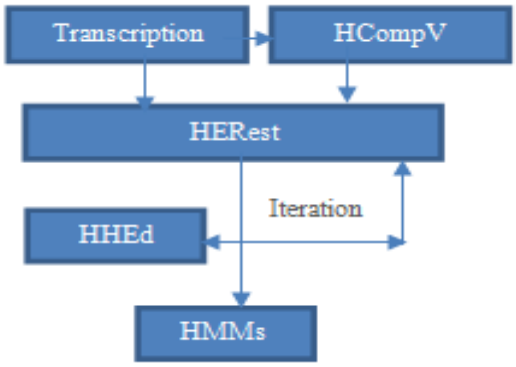

Figure 8: Training phase of Digit Recognizer

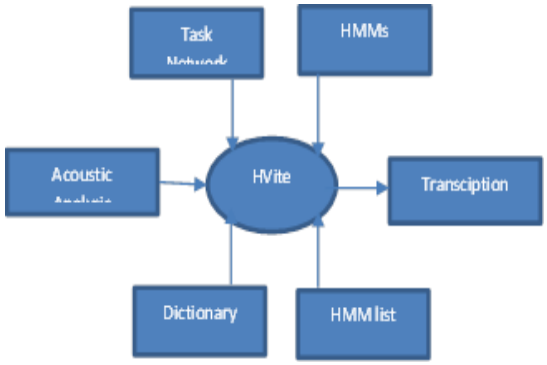

Figure 9: Testing phase of Digit Recognizer

Transformation of data into a series of acoustic vectors is done in the testing phase and the acoustic vectors with HMMs definition, task network, dictionary and HMM lists are processed. The test data transcription produced is shown in figure 9. HMM list is the text file which lists the names of the models. In our case, the HMM list consist names of all the total HMMs.

D. Database Collection

For preparing the database of our system, 5354 speech data was collected. The total data belonged to 120 male and 71 female speakers. Each speaker had given voice from zero to nine respectively. 5254 samples were used for training. 100 samples were used for testing.

IV. RESULTS

After training and testing the system, the performance of the speech system is mentioned in below equation:

$$Percentage\ Accuracy(PA) = \frac{N - D - S - I}{N} * 100$$

Where is N is the number of phones in test set, D is the number of deletion, S is the number of substitution and I is the number of insertion and PA gives accuracy rate.
Performance of Digit Recognizer:-

Digit recognizer	Accuracy
Digits	99.67%

```
------------------------ Overall Results ----------------
SENT: %Correct=99.00 [H=99, S=1, N=100]
WORD: %Corr=99.67, Acc=99.67 [H=299, D=0, S=1, I=0, N=300]
------------------------ Confusion Matrix ----------------
         o   t   t   f   f   s   s   e   n   z   s
         n   w   h   o   i   i   e   i   i   e   i
         e   o   r   u   v   x   v   g   n   r   l
                 e   r   e       e   h   e   o
                 e               n   t               Del [ %c
 one    10   0   0   0   0   0   0   0   0   0   0
 two     0   9   0   1   0   0   0   0   0   0   0 [90.
thre     0   0  10   0   0   0   0   0   0   0   0
four     0   0   0  10   0   0   0   0   0   0   0
five     0   0   0   0  10   0   0   0   0   0   0
 six     0   0   0   0   0  10   0   0   0   0   0
seve     0   0   0   0   0   0  10   0   0   0   0
eigh     0   0   0   0   0   0   0  10   0   0   0
nine     0   0   0   0   0   0   0   0  10   0   0
zero     0   0   0   0   0   0   0   0   0  10   0
 sil     0   0   0   0   0   0   0   0   0   0 200   0
Ins      0   0   0   0   0   0   0   0   0   0   0
```

The accuracy of 99.67% was achieved which is quite enough without any segmentation. Another performance metric and also be derived from the percentage accuracy which is known as digit error rate(DE) and it can be given as:

DE= 100%-PA which in our case is 0.43%.

V. CONCLUSION

This paper presents the recognition of digits, based on Hidden Markov Model and implemented using the portable HTK toolkit. The results obtained reveal that the performance of this digit recognizer is quite satisfactory. The proposed system provides 99.67% accuracy which is quite significant in terms of accuracy. This work can be further extended to larger vocabularies and increment in dictionary size in order to incorporate larger amount of data for testing and training.

REFERENCES

1. L.R. Rabiner , "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol. 77, No- 2,pp. 257-286,1989.
2. R. Rabiner, and B. H. Huang, "An introduction to hidden markov models", IEEE Acoustic. Speech Signal Processing Mag., pp. 4-16, 1986.

3. Dave, N "Feature extraction Methods LPC,PLP,MFCC in speech recognition", International Journal for advance research in Engineering and Technology,vol-1, pp.2-4,2013..
4. Salam Nandakishor, S. K. Dutta and L. Joyprakash Singh, "An HMM based SemiAutomatic syllable labeling system for Manipuri language", ICCCA 2015, May 15-16, 2015 at Galgotias University, UP, IEEE Proceeding, 2015.
5. Adami, A.G., Hermansky, H, "Segmentation of Speech for Speaker and Language Recognition" Proc. Eurospeech'03, pp. 841-844, September 2003.
6. Singh P.P and Rani P, "An Approach to Extract Feature using MFCC", IOSRJEN, vol. 04, Issue 08 pp. 21-25, Aug 2014
7. Steve Young, "HMMs And Related Speech Recognition Technologies", Springer Handbook on Speech Processing and Speech Communication.
8. Preeti Saini, Parneet Kaur, "Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology, vol-4 Issue 2- 2013
9. Claude C, Farzin Deravi, "A Review of Speech Based Bimodal Recognition", IEEE Transactions on multimedia, vol-4, no-1, 2002.
10. P. Mohanty and A. Nayak, "Isolated Odia Digit Recognition using HTK: An Implementation View", IEEE Proceedings of Second International Conference on Data Science and Business Analytics, pp-30-35, 2018