Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Isolated English Language Digit Recognition Using Hidden Markov Model Toolkit

Ganesh S Pawar*
Department of E&TC, University of Pune
India

Sunil S Morade
Department of E&TC, University of Pune
India

Abstract— The main purpose of the study was to develop a speech recognition system for isolated digits of English language using HTK. Speech, in addition to being a tool of communication, is also a symbol of identity and authorization. Two different corpora were collected of audio recordings of isolated digits of English language speakers, in which speakers read numeric digits. Both of the collected corpora contained the training data and the other testing data. One corpus is self recorded signals and other is standard CUAVE dataset (50 speakers, each uttered 10 words). The system has been implemented using the HMM toolkit i.e. HTK by training HMMs of the words making the vocabulary on the training data. Different HMMs for individual digits have been initialized and trained to have well modelled structure. The trained system was tested on training data as well as test data and results revealed that 95% of the data was correctly recognized.
The developed system can be used by developers and researchers interested in speech recognition for English language not only for isolated digits but also for other words of English language. The findings of the study can be generalized to cater for large vocabularies and for continuous speech recognition.

Keywords— CUAVE, HTK, HMM, Isolated digits, MFCC

I. INTRODUCTION

Communication has been the integral aspect of human life; a strong tool for sharing and building the knowledge that is passed on from generation to generation. Speech recognition is a field of computer science that deals with designing computer systems that recognize spoken words. As the new generation of computing technology, ASR (Automatic Speech Recognition) comes as the next major innovation in man-machine interaction, after functionality of text-to-speech (TTS), supporting interactive voice response (IVR) systems. The first attempts (during the 1950s) to develop techniques in ASR, which were based on the direct conversion of speech signal into a sequence of phoneme-like units, failed [1]. The first positive results of spoken word recognition came into existence in the 1970s, when general pattern matching techniques were introduced. R Kumar [2] implemented an experimental, speaker dependent, real time, isolated word recognizer for the regional language like Punjabi and further extended his work to compare the performance of speech recognition system for small vocabulary of speaker dependent isolated spoken words using the Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) technique. Further a group of researchers have developed a connected words speech recognition system for regional language like Hindi. The system was developed using Hidden Markov Model Tool Kit (HTK) and the system was trained to recognize any sequence of words selected from vocabulary of 102 words [3]. It got reasonably good success rate to recognize the words. However, early systems were expensive hardware devices that could only recognize a few isolated words (i.e. words with pauses between them), and needed to be trained by users repeating each of the vocabulary words several times. There are several methods of feature extraction, out of which we are using MFCC (Mel Frequency Cepstrum Coefficients) with delta and acceleration coefficients technique due to its advantages over other methods and availability of simpler framework in the Hidden Markov Model Tool Kit i.e. HTK [4].

Currently, the statistical techniques prevail over ASR applications. Common speech recognition systems these days can recognize thousands of words. There are different approaches for recognition or classification based on different techniques like Template based, Statistics based, Learning based, Knowledge based and Artificial Intelligence based. In this work, we are going to use HMM which is on Statistical approach. The HMM is popular statistical tool for modelling a wide range of time series data [5]. The last decade has witnessed dramatic improvement in speech recognition technology, to the extent that high performance algorithms and systems are becoming available. The reason for the evolution of ASR, hence improved is that it has a lot of applications in many aspects of our daily life, for example, telephone applications, applications for the physically handicapped and illiterates and many others in the area of computer science [6]. The aim of this work is therefore to design and train a speech recognition system which will recognize the speech signals (that is test signals) of isolated digits of English language in Linux environment using HTK and MFCC as the feature of extraction with delta and acceleration coefficients on standard database like CUAVE and self recorded database. The digits used are 0 to 9.

This paper is organized in to various sections. Section 2 discusses the Hidden Markov Model (HMM) and HTK. Section 3 explains methodology to be adopted to complete the recognition task. Output result has been compared in section 4. Lastly, conclusions are discussed the section 5.

II. HMM AND HTK

This section is going to explain the basic concepts and required things to be known to design a HMM using HTK for isolated digit speech recognition of English language (0-9) and use of MFCC as feature extraction technique.

A. Hidden Markov Model

HMM is very powerful mathematical tool for modelling time series. It provides efficient algorithms for state and parameter estimation, and it automatically performs dynamic time warping of signals that are locally stretched. Hidden Markov models are based on the well known chains from probability theory that can be used to model a sequence of events in time. Markov chain is deterministically an observable event. The most likely word with the largest probability is produced as the result of the given speech waveform. A natural extension of Markov chain is Hidden Markov Model (HMM), the extension where the internal states are hidden and any state produces observable symbols or evidences [7].

Here all three symbols represents probability distributions i.e. A, B and π. The probability distributions A, B and π are usually written in HMM as a compact form denoted by lambda as $\lambda = (A, B, \pi)$. Generally we are using Left – Right Architecture without state skipping. In our work, we are using the same kind of HMM. In this, transition from 1st state to 2nd state and to itself is allowed. It uses 7 states in which 1st and 7th state are non-emitting, other 5 are emitting states. PDF will be single Gaussian with diagonal co-variance matrix. We are using a file called 'proto' which contains all necessary information and specifications. This file has been taken as it is from the HTK book.

B. HTK- Hidden Markov Model ToolKit

HTK is one of the most widely used tools for speech recognition research and teaching-learning. The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research although it has been used for numerous other applications including research into speech synthesis, character recognition and DNA sequencing. The HMM Toolkit was originated in Machine Intelligence Laboratory in the Cambridge University Engineering Department. A currently available stable release is version HTK 3.4.1.

C. MFCC

The speech recognition tools of HTK cannot directly process on the speech waveform. These have to be represented in more efficient and compact form. The original waveform must be converted to such form or vectors. The computational steps of calculating MFCC include:

Framing: The signal is segmented in successive frames overlapping with each other.

Windowing: Each frame is multiplied by a windowing function like Hamming window.

Discrete Fourier Transform (DFT), Mel Scaling: From the data obtained above, and performing DFT, log scaling and conversion to Mel scale can be done.

All these processes and conversions can be done in single step using HTK. We need to just use a configuration file consisting of all such specifications and used it with HCopy tool to extract MFCC features. We can compute delta and acceleration coefficients by just setting MFCC to MFCC_0_D_A. It also includes details like no. of cepstral coefficients, sampling frequency, length of time frame, frame rate, source format, target format, whether normalization is required or not and other necessary details. Acoustic vectors (i.e. .mfc files) are used in both training and decoding phase of the system. In signal processing, these steps need to be done sequentially, whereas using HTK only setting the configuration file and giving proper input, user can easily get the MFCC coefficients in the way they want.

III. METHODOLOGY USING HTK

The methodology used is nothing but the use of modelling technique like HMM with special tool like HTK. Here according to Rabiner [9], we are adopting the regular procedure for isolated digit speech recognition of English language. Initially data preparation tools of HTK (like HParse, HCopy) are used to prepare the language model, dictionary and word network. Then parameter estimation tools of HTK (HInit) are used to define the HMMs of every digit and then to initialize the model. Further training tools like HRest are used for re-estimation to have proper and robust training. Finally recognition tools like HVite, HResults are used to have recognition results and confusion matrix for the given dataset. In this work we are using MFCC as the feature of extraction with delta and acceleration coefficients due to its advantage over other feature extraction methods. The main requirement of the work is, we must have necessary sample speech signals for training and testing purpose, specifically in .wav format and our system (i.e. PC or Laptop) must have Linux installed on it. In order to specify to HTK the nature of the audio data (format, sample rate, etc.) and feature extraction parameters (type of feature, window length, pre-emphasis, etc.), a configuration file (config.txt) was created as said in HTK book. Here the steps are given to carry out the complete process. Here only the step 6 is manual.

Step: 1 Create the two different folders or directories separately on the HOME drive of our Linux system. These folders will contain the training and test database.

Step: 2 Install the HTK 3.4.1 on our system along with HDecode package.

Step: 3 Here we will use SOX to convert all *.wav files to *.raw files.

Step: 4 Create all the necessary files manually like Configuration file, dictionary file, grammar file and models file. Not to forget a file called 'proto' must be taken as it is from the HTK book, as provided on the website from where we have downloaded HTK i.e. http://htk.eng.cam.ac.uk.

Step: 5 Convert data files (.raw) into parametric format using HCopy.

Step: 6 Initialize HMM models using proto definitions and .mfc files. Create generic definitions for each word model and rename it manually using HInit.

Step: 7 Re-estimate parameters for word models using data files using HRest. This step can be done in multiple iterations, here we are going for 3 re-estimations.

Step: 8 Perform the recognition on test data using HVite. Analyse the recognition results for Accuracy on both training and test data using HResults.

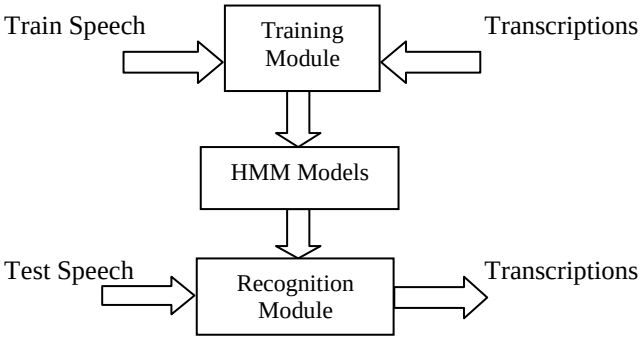

Fig. 1 HMM Based Speech Recognition System

IV. RESULTS

The evaluation of the performance of the speech recognition system can be done by using HTK tool HResults. It is clear that for the self recorded dataset, the recognition results are poor as compared that with standard dataset like CUAVE. As the recording conditions will play major role in making the signal noise-free. The comparison of the results for CUAVE and self recorded dataset have been given with a graph and result windows along with confusion matrix generated.

The database used here is a CUAVE database. CUAVE (Clemson University Audio Visual Experiments) was recorded by E.K. Patterson of Department of Electrical and Computer Engineering, Clemson University, US. The database was recorded in an isolated sound booth. This database is a speaker-independent database consisting of connected and continuous digits spoken in different situations. It contains mixture of speaker with white and black skin. Database digits are continuous and with pause. Total no. of speakers were 50, each uttered 10 words. So total speech samples were 500. Out of which 400 were used for training and 100 were used for testing purpose. The speakers consist of Males, Females from different age group.

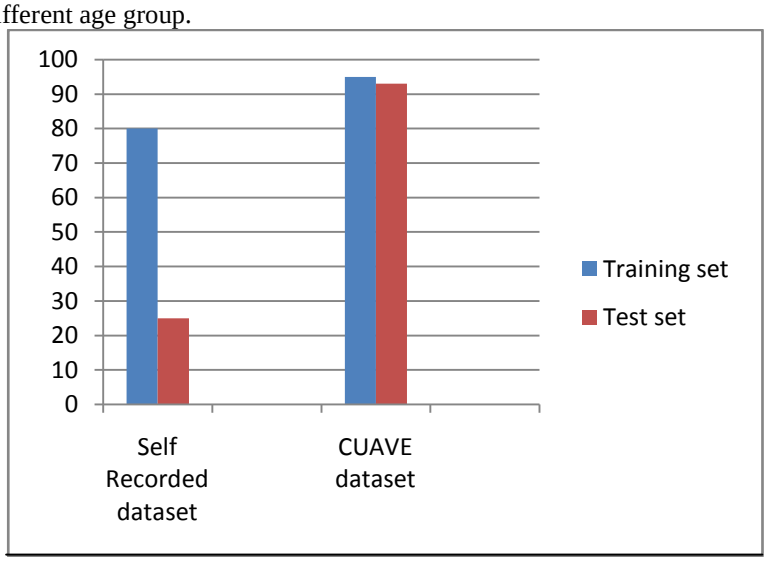

Fig. 2 Comparison of results of 2 different datasets using HTK

The system is relatively successful, as it can identify the spoken digit at an accuracy of 95%, which is relatively high. Presently, system can be used only for isolated word single digit recognition and it can be extended to perform user authentication based on speech and also accommodate connected digits. The confusion matrix generated is shown below in figure 4.2.


```
====================== HTK Results Analysis ================
  Date: Mon Jun 16 13:24:38 2014
  Ref : train.mlf
  Rec : testout1.mlf
---------------------- Overall Results --------------------
SENT: %Correct=95.00 [H=380, S=20, N=400]
WORD: %Corr=95.00, Acc=95.00 [H=380, D=0, S=20, I=0, N=400]
---------------------- Confusion Matrix -------------------
       Z   O   T   T   F   F   S   S   E   N
       E   N   W   H   O   I   I   E   I   I
       R   E   O   R   U   V   X   V   G   N
       O           E   R   E       E   H   E
                   E               N   T       [ %c / %e]
ZERO  36   0   4   0   0   0   0   0   0   0   [90.0/1.0]
 ONE   0  35   0   0   0   1   0   0   0   4   [87.5/1.2]
 TWO   2   0  38   0   0   0   0   0   0   0   [95.0/0.5]
THRE   0   0   0  39   0   0   0   0   1   0   [97.5/0.2]
FOUR   0   0   0   0  40   0   0   0   0   0
FIVE   0   0   0   0   1  39   0   0   0   0   [97.5/0.2]
 SIX   0   0   0   0   0   0  40   0   0   0
SEVE   0   0   0   0   0   0   0  38   0   2   [95.0/0.5]
EIGH   0   1   0   1   0   0   0   0  38   0   [95.0/0.5]
NINE   0   1   0   0   0   2   0   0   0  37   [92.5/0.8]
============================================================
```

```
====================== HTK Results Analysis =============
  Date: Mon Jun 16 13:34:03 2014
  Ref : train.mlf
  Rec : testout1.mlf
---------------------- Overall Results -----------------
SENT: %Correct=80.00 [H=44, S=11, N=55]
WORD: %Corr=80.00, Acc=80.00 [H=44, D=0, S=11, I=0, N=55]
---------------------- Confusion Matrix ----------------
       Z   O   T   T   F   F   S   S   E   N
       E   N   W   H   O   I   I   E   I   I
       R   E   O   R   U   V   X   V   G   N
       O           E   R   E       E   H   E
                   E               N   T       [ %c / %e]
ZERO   2   0   0   1   1   0   0   1   0   0   [40.0/5.5]
 ONE   0   5   0   0   0   0   0   0   0   0
 TWO   0   0   3   0   2   0   0   0   0   0   [60.0/3.6]
THRE   0   0   0   5   0   0   0   0   0   0
FOUR   0   1   0   0   4   0   0   0   0   0   [80.0/1.8]
FIVE   0   1   0   0   1   3   0   0   0   0   [60.0/3.6]
 SIX   0   0   0   0   0   0   4   0   1   0   [80.0/1.8]
SEVE   0   0   0   0   0   0   0   5   0   0
EIGH   0   0   0   1   0   0   0   0   4   0   [80.0/1.8]
NINE   0   0   0   0   0   0   0   0   0   5
============================================================
```

Fig. 3 Confusion Matrix for CUAVE and Self recorded dataset

V. CONCLUSIONS

HTK was used for the implementation of the recognizer. HTK was used because it is free and has been used by many researchers all over the world. HTK supports both isolated whole word recognition and sub-word or phone based recognition. A limited grammar and dictionary were constructed to be used by the recognizer. The experiments/test carried out showed that a higher level of accuracy can be achieved if the language model was designed for limited dictionary and trained the word model with a large set of speech data from the user.

The system was tested using testing corpus data and the system scored up to 95% word recognition. The work is however not all conclusive as it has catered for only an Isolated Digit Speech data. As much as it has created a basis for research, this work can be expanded to cater for more extensive language models and larger vocabularies. For the work presented, digit-pronunciation was limited to the English language only. The model could be further developed to incorporate digits from other languages; most preferably the local languages of the clients. Furthermore, the dictionary size could be increased using alphabets and commonly used words, so the large test data could be generated and trained. The system can be made robust by using larger database for training.

ACKNOWLEDGMENT

Authors would like to thank the experts who have contributed towards development of work and its implementation as described in the paper.

REFERENCES

[1] R. Klevansand, R. Rodman, "*Voice Recognition*", Artech House, Boston, London 1997.

[2] R. Kumar, "*Comparison of HMM and DTW for Isolated Word Recognition of Punjabi Language*", In Proceedings of Progress in Pattern Recognition, Image Analysis, Computer Vision and Applications, Sao Paulo, Brazil. Vol. 6419 of LNCS, pp. 244-252, Springer Verlag, November 8-11, 2010.

[3] K. Kumar, R. K. Aggrawal, A Jain, "*A Hindi speech recognition system for connected words using HTK*", International Journal of Computational Systems Engineering, Vol. 1, No. 1, 2012.

[4] Santosh K Gaikwad, Bharti W Gawali, Pravin Yannawar, "*A Review on Speech Recognition Techniques*", International Journal of Computer Applications (0975-8887), Vol. 10, No. 3, November 2010.

[5] Ibrahim patel, Dr. Y shrinivas rao, "*Speech recognition using HMM with MFCC – an analysis using frequency spectral decomposition technique*", Signal and Image Processing: An International Journal (SIPIJ), Vol. 1, No. 2, December 2010.

[6] Rabiner L.R., S.E. Levinson, "*Isolated and connected word recognition - Theory and selected applications*", IEEE Trans. COM-29, pp.621-629, 1981.

[7] Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, "*The HTK Book"*, 2002 from: http://htk.eng.cam.ac.uk.

[8] Kritika Nimje, Madhu Shandilya, "*Automatic isolated digit recognition system: an approach using HMM*", Journal of Scientific and Industrial Research, Vol.70, pp. 270-272, April 2011.

[9] Juang B, Rabiner L, "*Hidden Markov Models for speech recognition*", Technometrics, 33 (1991), 251-272.

[10] Dipmoy Gupta, Radha Mounima C., Navya Manjunath, Manoj P.B., "*Isolated word speech recognition using VQ*", International Journal of Advanced Research in Computer science and Software Engineering, Vol. 2, Issue 5, ISSN: 2277 128X, May 2012.

[11] Kritika Nimje, Madhu Shandilya, "*Automatic isolated digit recognition system: an approach using HMM*", Journal of Scientific and Industrial Research, Vol.70, pp. 270-272, April 2011.

[12] Mohit Dua, R. K. aggarwal, Virender Kadyan, Shelza Dua, "*Punjabi Automatic Speech Recognition using HTK*", IJCSI, Vol. 9, Issue 4, No. 1, July 2012.

