Realization of Hidden Markov Model for English Digit Recognition

Ganesh S Pawar
Department of E &TC Engg.
SNJB's KBJ COE
Chandwad (Nashik)

Sunil S Morade
Department of E &TC Engg.
KKW IEER
Nashik

ABSTRACT

The objective of the work described here is to compare the isolated English language digit speech recognition using Hidden Markov Model for speaker independent system. Two different datasets were collected of audio recordings for the said comparison of isolated digits of English language. Speakers here read numeric digits 0 to 9 i.e. ZERO to NINE. One corpus is self recorded signals and other is standard CUAVE dataset (36 speakers, each uttered 10 words). The training and testing samples are separated for speaker dependent and speaker independent systems. 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 modeled structure. The trained system was tested on training data as well as test data and results shown that most of the speech samples were correctly recognized.

The system was tested for speaker independent and dependent way, to check the changes in the recognition rate. Further this 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. If clean database is available, further this can be generalized to recognize words of any language. Continuous speech can also be recognized using study of this system.

General Terms

Algorithm, Confusion matrix, Database, Speech Recognition.

Keywords

CUAVE, HTK, HMM, Isolated digits, Speaker Independent

1. INTRODUCTION

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]. In this system, at Bell laboratories a single speaker was involved. It was purely speaker dependent system. 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].

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 modeling 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 for speaker dependent and independent approach.

This paper has been organized in to various sections. Section 2 discusses the Hidden Markov Model (HMM) and MFCC. 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.

2. HMM AND FEATURE EXTRACTION

This section is going to explain the basic concepts and required things to be known to design a HMM and use of MFCC as feature extraction technique.

2.1 Hidden Markov Model (HMM)

HMM is very powerful mathematical tool for modeling 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 observable evidences [7]. Mathematically Hidden Markov Model contains five elements.

1. Internal States: These states are hidden and give the flexibility to model different applications. Although they are hidden, usually there is some kind of relation between the physical significance to hidden states.

2. Output: O = {O_1, O_2, O_3, ... , O_n} an output observation alphabet.

3. Transition Probability Distribution: A = aij is a matrix. The matrix defines what the probability to transition from one state to another is.

4. Output Observation: Probability Distribution B = bi (k) is probability of generating observation symbol o (k) while entering to state i is entered.

5. The initial state distribution (π = { πi }) is the distribution of states before jumping into any state.

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 λ = (A, B, π). 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.

2.2 FEATURE EXTRACTION

MFCC’s are based on the known variation of the human ear’s critical bandwidths with frequency. Filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. Psychological studies have shown that, human perception of the frequency content of sounds, either for tones or for speech signals, does not follow linear scale. This research has led to the idea of defining subjective pitch of pure tones. Thus for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called Mel scale. Feature extraction using Mel Frequency Cepstral Coefficients (MFCC) utilizes power spectrum of which the center frequency and bandwidth are scaled by subjective Mel measure. The cepstral coefficients are then computed by taking the logarithm of the power spectrum and transforming this log spectrum to the cepstral domain using an Inverse Discrete Fourier Transform (IDFT) [10]. Cepstral coefficients can be seen as a parametric representation of the spectrum. 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. 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 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 by just setting MFCC to MFCC_0_D_A.

3. METHODOLOGY Using HTK

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. The methodology used is nothing but the use of modeling 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 get recognition results and confusion matrix for the given dataset. The complete diagram of the said methodology has been shown in the figure 1. It shows the different tools used at every different stage of training with necessary input and output routes to various other blocks of the system.

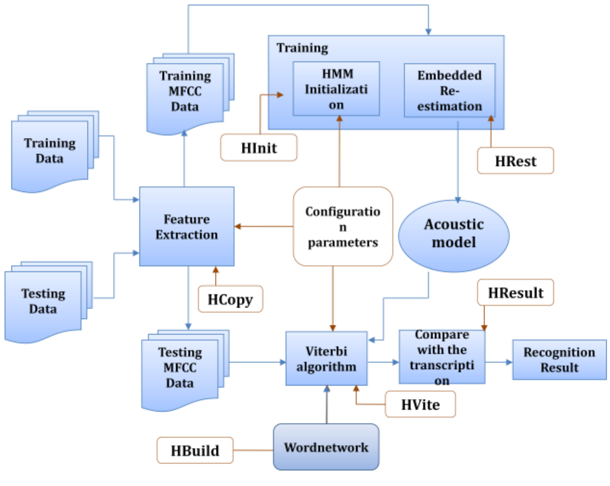

Figure 1-System realization using HTK

1. Create the two different folders. These folders will contain the training and test database.

2. Create all the necessary files manually like Configuration file, dictionary file, grammar file and models file. Download the file named ‘proto’ from the web address, i.e. http://htk.eng.cam.ac.uk.

3. Convert data files into parametric format using HCopy i.e. the MFCC files

4. Initialize HMM models using proto definitions and .mfc files. Create generic definitions for each word model and initialize it for every single digit.

5. Re-estimate parameters for word models using data files using HRest. This step can be done in multiple iterations, here we are going for 10 re-estimations.

6. Perform the recognition on test data using HVite. Analyze the recognition results for Accuracy on both training and test data using HResults

7. Now carry out the same process for speaker independent system. In this case, use the samples of unknown speakers who were not involved in the training dataset.

4. 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 has been given with a graph 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 [15]. 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 36, each uttered 10 words. The speakers consist of Males, Females from different age group. The system has been tested for speaker dependent system and yields maximum accuracy for CUAVE dataset and moderate accuracy for the own dataset. Also if we go for speaker independent system, where speakers involved in the training were different than those involved during testing. Here the recognition drops to some extent. Still we have up to 86% recognition for CUAVE dataset. The statistics have been shown in the graph. The sample result windows for CUAVE dataset generated by the HTK are also given below.

```
================HTK Results Analysis =============
  Date: Sat Jun 28 13:27:06 2014
  Ref : test.mlf
  Rec : testout.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=86.00 [H=43, S=7, N=50]
WORD: %Corr=86.00, Acc=86.00 [H=43, D=0, S=7, I=0, N=50]
===============================================
```

Figure 2- Results for speaker independent system

```
==============HTK Results Analysis ==============
  Date: Sat Jun  28 13:55:39 2014
  Ref : train.mlf
  Rec : testout1.mlf
------------------------ Overall Results ----------------------------------
SENT: %Correct=95 [H=380, S=20, N=400]
WORD: %Corr=95, Acc=95 [H=380, D=0, S=20, I=0, N=400]
==============================================
```

Figure 3- Results for speaker dependent system

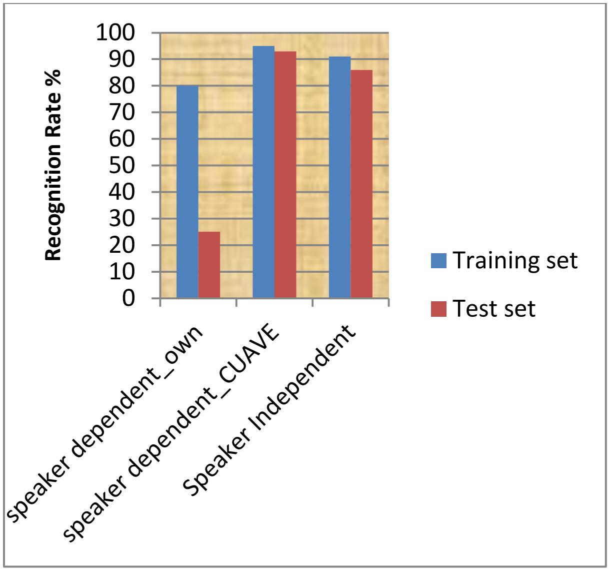

Figure 4- Comparison of results for different datasets and with speaker independent system.

The system is relatively successful, as it can identify the spoken digit at an accuracy of 95% for speaker dependent approach, which is relatively high. Further it is observed that for speaker independent approach, accuracy drops. It is due to the variations in the uttered speech by speakers, who were not involved in training, nature of utterance, environmental conditions, difference between speakers due to age, sex, accent etc. etc.. There is possibility that by using more robust model, we can increase the accuracy further. The confusion matrix generated is shown below in figure 5 [16].

```
===================== 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]
=============================================================
```

Figure 5- Confusion matrix for CUAVE dataset

A confusion matrix which tells us about the recognition rate for every individual digit checked against all other digit and to itself. It was observed that digit 4 and digit 6 gives up to 100% recognition. Digit 1 gets confused with digit 9 with confusion rate 12.50%. Digit 0 gets confused with digit 2, digit7 gets confused with 9. Also digit 9 gets confused with digit 1, 5.

5. CONCLUSION

5.1 Discussion

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. For the speaker independent approach; environmental conditions, speakers invariability, way of utterance should be considered and then accordingly system can be made more robust by more training the HMMs to give proper recognition.

5.2 Conclusion

The system was tested using testing corpus data and the system scored up to 95% word recognition for speaker dependent approach and up to 86% for speaker independent approach. 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, so the large test data could be generated and trained. The system can be enhanced to a larger vocabulary including alphabets and commonly used words. The system can be made robust by using larger database for training.

6. ACKNOWLEDGEMENT

Authors would like to thank the experts who have contributed towards development of work and its implementation as described in the paper and not to forget the staff members of KKW IEER, Nashik for providing the necessary infrastructure to carry out the experiments.

7. 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 (Retrieved Jan 2, 2013) from: http://htk.eng.cam.ac.uk.

[8] Roux, J.C., Botha, E.C., Du Preez, J.A., "Developing a Multilingual Telephone Based Information System in African Languages", Proceedings of the 2nd International Language Resources and Evaluation Conference, Athens, Greece : ELRA (2),975-980, 2000.

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

[10] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun – "An Efficient MFCC Extraction Method in Speech Recognition", Department of Electronics Engineering, The Chinese University of Hong Kong, Hong, IEEE – ISCAS, 2006.

[11] 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.

[12] 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.

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

[14] www.myfit.edu/~vkepuska/HTK/HTK-basic-tutorial.pdf

[15] E. K. Patterson, S. Gurbuz, Z. tufekci, and J. N. Gowdy, "CUAVE: A New Audio-visual Database for Multimodal Human Computer Interface Research", Clemson University, USA.

[16] Ganesh S Pawar, Sunil S Morade, "Isolated English Language Digit Recognition Using Hidden Markov Model Toolkit", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 4, Issue 6, ISSN: 2277 128X, June 2014.