ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(4), DOI: 10.17485/ijst/2015/v9i4/83894, January 2016* ISSN (Online) : 0974-5645

A Voice Identification System using Hidden Markov Model

T. K. Das[1*] and Khalid M. O. Nahar[2]

[1]SITE, VIT University, Vellore – 632014, Tamil Nadu, India; tapandas05@gmail.com
[2]Department of Computer Science, Yarmouk University, Irbid – 21163, Jordan; khalids@yu.edu.jo

Abstract

Background/Objectives: Voice Identification System refers to a system which comprises of hardware, software and it is used to identify voice for several applications. The aim of the research is to develop a small scale system that incorporate both speaker recognition and speech recognition and can show specific visual information to a user. **Methods:** To this end, we have developed a system based on the technique of Hidden Markov Model. The Hidden Markov Model is a stochastic approach which models the algorithm as a double stochastic process in which the observed data is thought to be the result of having passed a hidden process through second process. Both processes are characterized only through one that is observed. A database of voice information is created. To extract features from voice signals, Mel-Frequency Cepstral Coefficients (MFCC) technique has been applied producing a set of feature vectors. Subsequently, the system uses The Vector Quantization (VQ) for features training and classification. **Findings:** The designed system has been tested with multiple speakers as reference. Speech recognition based on Hidden Markov Model is achieved successfully for the conversion of speech to text. In this proposed research, speech recognition is achieved with accuracy about 90%. **Applications:** The system has potential to be used in music industry, crime investigation, personal assistant and in hi-tech devices.

Keywords: Hidden Markov Model, Mel-Frequency Cepstrum Coefficients (MFCC), Speech Recognition, Vector Quantization, Voice Identification

1. Introduction

Voice refers to the sound produced in a person's larynx and verbalized through the mouth, as speech or song. It is used to express a particular opinion or interest by using specific words[1]. Voice identification plays a pivotal role in the field of forensics, security and biometric authentication for verifying or identifying the voice of a speaker from the group of speakers. The wide range of applications and the potential need in varied industries has been a great motivation for developing the system. A system is developed that can recognize a speaker who in turn can help in crime industry or to differentiate various singers in music industry or may be used in the biometric system or others voice activated system. Apart from identifying a speaker, the system can instantly recognize isolated words and moreover, continuous words, i.e. a speech recognition. Speech recognition system along with voice or speaker recognition is called voice identification system. We would like to develop a voice identification system, a small-scale system that incorporates both speaker and speech recognition features and can show specific visual information to a user.

Speech refers to the ability to express thoughts and feelings by articulate sounds. The detection of speech by comparing with some person is called speech recognition[2]. Computer captures the words spoken by a human with a help of a microphone. These words are later recognized by speech recognizer and in the end, system outputs the recognized words. Speech recognition technology plays important role in many applications such as speech-to-text, language translation and speech input interface[3]. Speaker recognition is defined as to make sure that if the person is the same person he claims to be or not[4]. This

* *Author for correspondence*

technique is one of the biometric recognition techniques useful in almost all areas where security is a concern[5]. There are two main stages in this technique, feature extraction and feature matching[6]. In voice identification system we have combined both the speaker recognition and speech recognition technology.

Several works has been done in these fields in projects like CMU sphynix and HMM toolkit which is used in system like Google Now, Apple Siri or Microsoft Cortana. Voice recognition has been used in top technical system like NASA MARS Rover and has been successful to a degree in crime detection or for security purposes[7].

2. Hidden Markov Model

A Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states[8]. A HMM can be considered the simplest dynamic Bayesian network. In simpler Markov models (like a Markov chain), the state is directly visible to the observer and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output. 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[9].

Markov processes are stochastic processes where in a Markov model we have physical observation states while a HMM refers to double stochastic process in which we have physical observation linked to hidden states. This is shown in Figure 1.

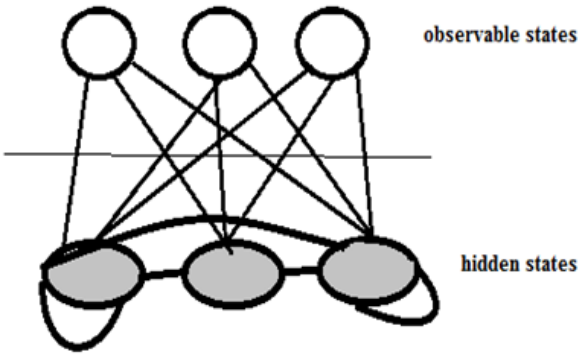

Figure 1. Basic HMM architecture.

3. System Architecture

The system is divided into two parts broadly – voice recognition and speech recognition

3.1 High Level Architecture

Figure 2 depicts the inner working of sub-systems within the voice identification system. Figure 3 explains the detailed operations in voice recognition.

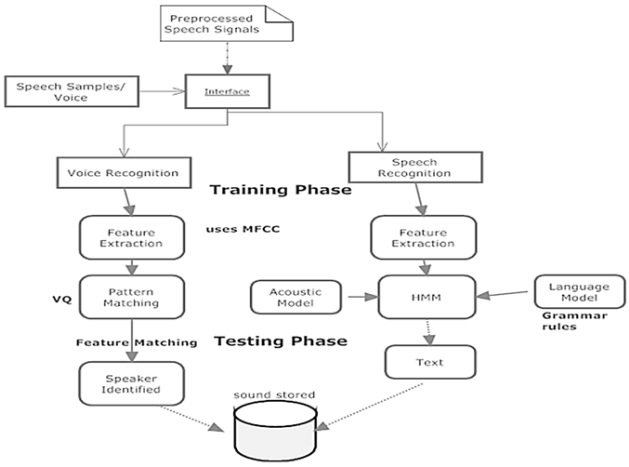

Figure 2. Architecture of voice identification system.

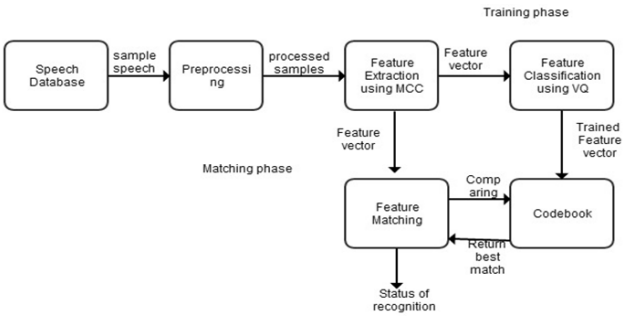

Figure 3. Phases of voice recognition.

3.2 Architecture of Voice Recognition

Speech recognition architecture is portrayed in Figure 4. Architecture of speech recognition.

4. Methodology

The following methodology clearly specifies the development of the system.

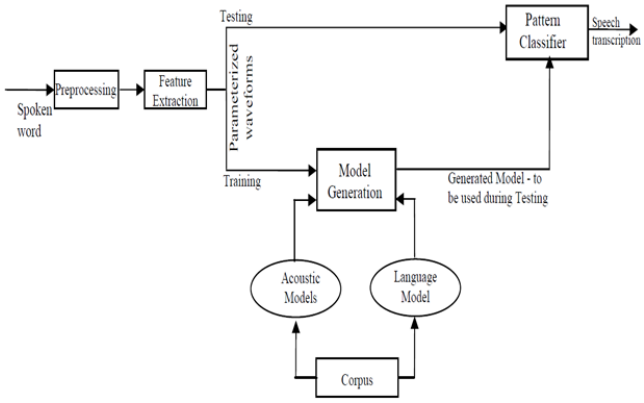

Figure 4. Phases of speech recognition.

The system is first divided into two major modules namely, voice recognition and speech recognition and then later on subdivided into other modules.

4.1 Voice Recognizer

It identifies the speaker and authorizes or authenticates its sound beside a database of voices. First the system is trained using specific voices and then it is tested with the unknown voice and the system recognizes the speaker or to whom the unknown voice belongs.

Speaker recognition systems fall into two categories: text-dependent[10] and text-independent.

4.1.1 Text-Dependent

If the text must be the same for enrollment and verification this is called text-dependent recognition. In a text-dependent system, prompts can either be common across all speakers (e.g.: a common pass phrase) or unique.

4.1.2 Text-Independent

Text-independent systems are most often used for speaker identification as they require very little cooperation by the speaker. In this case the text during enrollment and test is different. However, the enrollment may happen without the user's knowledge, as in the case for many forensic applications.

Sub-modules**:**

- Train: Play each sound file in the TRAIN folder. Record this result so that it can be used later to be compared with the computer's performance of our system. Both of us seem to be unable to recognize random people just by listening at their voice. Our success rates for the provided samples were 1 person out of 8 each.
- Test: Play each sound file from the Test folder. Record this result and compares it. The trained system finally founds the sound.
- Plot: Compute the power spectrum and plot it out using the images command. Note that it is better to view the power spectrum on the log scale. Locate the region in the plot that contains most of the energy. Translate this location into the actual ranges in time (msec) and frequency (in Hz) of the input speech signal.
- MFCC: Major functions used for Speech Recognition are MFCC[11] and VQ[12].

The detailed MFCC process has been shown in Figure 5.

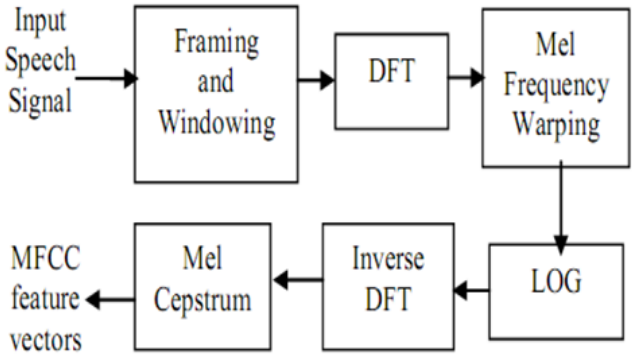

Figure 5. Detailed MFCC process.

The following algorithm explains the detailed MFCC process.

Inputs: s contains the signal to analyze fs is the sampling rate of the signal.

Output: r contains the transformed signal

```
begin
function r = mfcc (s, fs)
m = 100;
n = 256;
frame = block Frames (s, fs, m, n);
m = melfb (20, n, fs);
n2 = 1 + floor (n / 2);
z = m * abs (frame(1:n2, :)).^2;
r = dct (log(z));
end.
```

Cepstrum signal transformation steps has been represented in Figure 6.

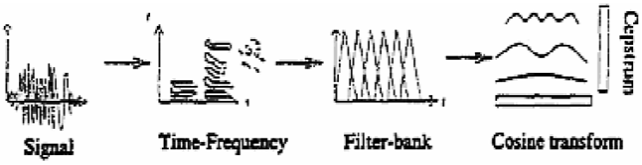

Figure 6. Cepstrum signal and their conversion.

- Vector Quantization (VQ) is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm. In the earlier days, the design of a Vector Quantizer (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. In 1980, Linde, Buzo and Gray (LBG) proposed a VQ design algorithm based on a training sequence. The use of a training sequence bypasses the need for multi-dimensional integration. A VQ that is designed using this algorithm are referred to as LBG-VQ. The algorithm requires an initial codebook. This initial codebook is obtained by the splitting method. In this method, an initial code-vector is set as the average of the entire training sequence. This code-vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code-vectors are split into four and the process is repeated until the desired number of code-vectors is obtained[13]. The algorithm guarantees a locally optimal solution. Steps of vector quantization are represented in Figure 7.

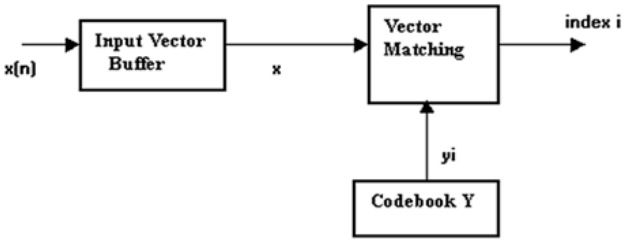

Figure 7. Vector quantization process steps.

4.2 Speech Recognizer

It identifies the speech i.e. the specific word being spoken. First the system is trained here too, on sample of voices and then the system is given a specific speech or sound it recognize the sound based on establish vocabulary and stored grammar rules applied to the dictionary. It uses HMM or hidden Markov model at its base to identify the sound.

4.2.1 Sub-Modules

- Train: Play each sound file in the TRAIN folder. Record this result so that it can be used later to be compared with the computer's performance of our system. The sound files of both the recognizer modules differ and it consists of sound of specific word.
- Test: Command like HVite from HTK toolkit is being used. Show the specific word of speech recognized.
- Analysis: The speech recognition tools cannot process directly on speech waveforms. These have to be represented in a more compact and efficient way. This step is called acoustical analysis and configuration file, target file and HCopy command is used. It is represented in Figure 8.

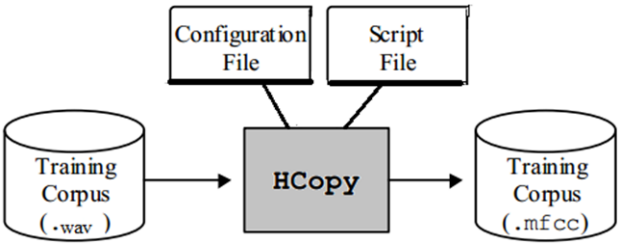

Figure 8. Detailed analysis of sound.

- Modeling: HMM is basically used in Acoustic Modeling but it can also be used in many purposes other than this acoustic modeling. HMM is based on the well-known Markov chains from probability theory. Each HMM state may have a set of output symbols known as output probabilities and having a finite number of states Q = {q1, q2,..., qn}. One process is related to the transitions among the states which are controlled by a set of probabilities called transition probabilities. At each distinct instance of time, one process is assumed to be in some state and an observation is produced by the other process representing the current state.

For training part, we use Baum-Welch algorithm to determine the parameters of a

HMM. To compute maximum likelihood scores, we use Viterbi algorithm in recognition part.

5. Experimental Results

We have plotted one dimensional power spectrum using linespac MatLab function and functions like VQ, melfb. The power spectrum of a signal is the power of that signal at each frequency that it contains. For example, white noise, which contains all frequencies at the same power, has a flat power spectrum. Figure 9 represents Mel-space and VQ Code words. In order to make the interface as much user friendly as possible, we have designed the

interface of the system which is being represented in Figure 10.

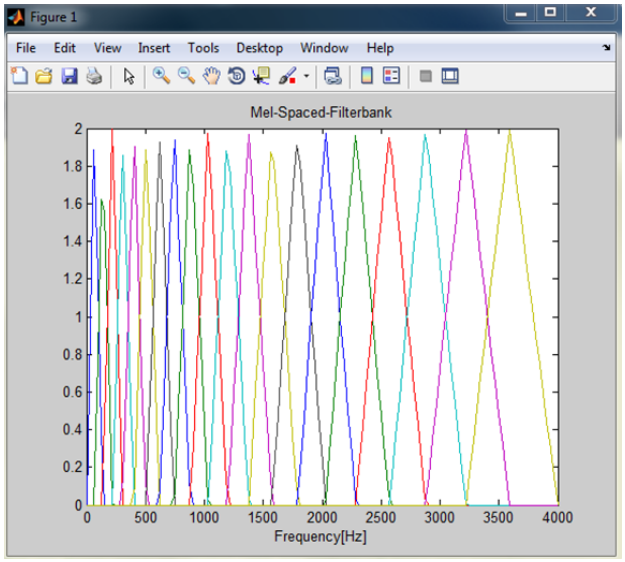

Figure 9. Plots representing Mel-space and VQ code words.

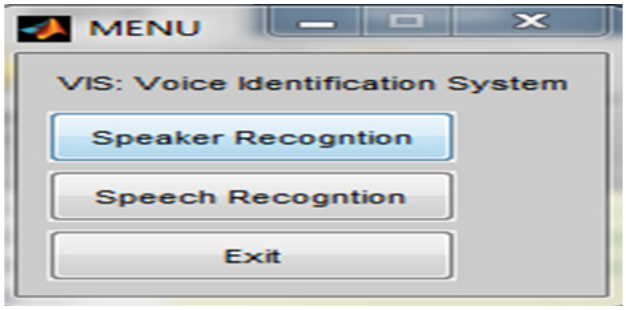

Figure 10. User interface of the system.

6. Conclusion

We have designed a system that would recognize isolated words and identify the voice of different speakers. We have used Hidden Markov Models for isolated word recognizer, i.e. speech recognition. However MFCC and VQ haves been used for speaker recognition. The system has been implemented in MatLab as it directly accesses speaker recognizer module and for speech recognition it accesses command prompt where commands regarding HTK toolkits and description text files along with wave file sounds are used. In this research work, speech recognition has been accomplished successfully by using Hidden Markov Model with about 90% accuracy. In future, the work can be extended for continuous speech recognition. It can be made practical and developed for an application that can be used in various fields like music industry, health care or as a personal assistant.

7. References

1. Wang L, Minami K, Yamamoto K, Nakagaw S. Speaker identification by combining MFCC and phase information in noisy environments. Proceeding of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP); Dallas, TX, USA. 2010 Mar 14-19. p. 4502–5.
2. Sharma K, Singh P. Speech recognition of Punjabi numerals using synergic HMM and DTW approach. Indian Journal of Science and Technology. 2015 Oct; 8(27):1–6.
3. Chaiwongsai J, Chiracharit W, Chamnongthai K, Miyanaga Y. An architecture of HMM-based isolated-word speech recognition with tone detection function. Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'2008); Bangkok, Thailand. 2009 Feb 8-11. pp.1–4.
4. Biswas S, Ahmad S, Molla I. Speaker identification using Cepstral based features and discrete Hidden Markov Model. Proceeding of IEEE International Conference on Information and Communication Technology (ICICT'07); Dhaka, Bangladesh. 2007 Mar 7-9. p. 303–6.
5. Shah HNM, Rashid MZA, Abdollah MF, Kamarudin MN, Lin CK, Kamis Z. Biometric voice recognition in security system. Indian Journal of Science and Technology. 2014 Feb; 7(2):104–12.
6. Abushariah AAM, Gunawan TS, Chebil J, Abushariah MAM. Voice based automatic person identification system using vector quantization. Proceedings of the International Conference on Computer and Communication Engineering (ICCCE 2012); Kuala Lumpur, Malaysia. 2012 Jul 3-5. p. 549–54.
7. Eriksson T, Kim S, Kang HG, Lee C. An information-theoretic perspective on feature selection in speaker recognition. IEEE Signal Processing Letters. 2005 Jul; 12(7):500–3.
8. Rabiner LR. A tutorial on Hidden Markov Model and selected applications. Proceedings of the IEEE. 1989 Feb; 77(2):257–86.
9. Prakash A, Chandrasekar C. An optimized multiple Semi-Hidden Markov Model for credit card fraud detection. Indian Journal of Science and Technology. 2015 Jan; 8(2):165–71.
10. Ranjan R, Singh SK, Shukla A, Tiwari R. Text-dependent multilingual speaker identification for Indian languages using artificial neural network. Proceeding of IEEE International Conference on Emerging Trends in Engineering and Technology (ICETET); Goa, 2010 Nov 19-21. p. 632–5.
11. Kari B, Muthulakshmi S. Real time implementation of speaker recognition system with MFCC and neural networks on FPGA. Indian Journal of Science and Technology. 2015 Aug; 8(19):1–11.

12. Zulfiqar A, Muhammad A, Enriquez AMM. A speaker identification system using MFCC features with VQ technique. IEEE 3rd International Symposium on Intelligent Information Technology Application; Nanchang. 2009 Nov 21-22. p. 115–8.

13. Singh S, Rajan EG. MFCC VQ based speaker recognition and its accuracy affecting factors. International Journal of Computer Applications. 2011 May; 21(6):1–6.