VOICE BASED SPEAKER IDENTIFICATION SYSTEM

Vidushi Garg[1], Parminder Singh[2], Parneet Kaur[3]

[1]*Dept. of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India*
technovidushi@yahoo.com
[2]*Dept. of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India*
parminder2u@gmail.com
[3] *Dept. of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India*
parneet.sidhu@ymail.com

Abstract

In today's context security is a main concern of most of the organization. Therefore for their own employee and customer they need a very reliable and robust security system which authenticates people based on a criteria which does not waste the time of people in due cause of authentication. For this it is suggested a human voice is a best time savor factor for doing authentication and identification of a person. The person can simply speak out and its voice is identified and recognized in due authentication process. In our research work we are conducting a survey of all techniques which help us to get a clear view of methodologies used to achieve high success rate of true positives in authentication.

Keywords: *Dynamic time warping (DTW), Hidden Markov model (HMM), Artificial Neural Network (ANN).*

----------***----------

1. Introduction

Speech recognition is a process of converting speech signal to a sequence of word. Speech recognition is the ability of a machine or a program to identify words and phrases in spoken and convert them to a machine-readable format. The only limitation of Speech recognition is limited vocabulary of words and phrases and that may only identify these if they are spoken very clearly. Speech recognition applications include call routing, speech-to-text, voice dialing and search. The terms "speech recognition" and "voice recognition" are sometimes used synonymously.

However, the two terms have different meanings. Speech recognition is used to identify words in spoken language. Voice recognition is a biometric technology used to identify a particular individual's voice. The performance of a speech recognition system (SRS) is usually specified in terms of accuracy and speed. Accuracy is measured with the word error rate, whereas speed is measured with the real time factor.

The speech recognition process is performed by a software component known as the speech recognition engine. Speech recognition engine is to process the spoken input and then translate it into a text that an application can understand.

Number of approaches has been used for recognition of speech but here we discussed only two of them one is dynamic programming and other is neural network. Speech recognition
The speech recognition process is performed by a software component known as the speech recognition engine. The primary function of the speech recognition engine is to process the spoken input and then translate it into a text that an application can understand.

Number of approaches has been used for recognition of speech but here we discussed only two of them one is dynamic programming and other is neural network. Speech recognition basically consists of two main modules that are feature extraction and feature matching. The main objective of feature extraction module is to convert speech waveform to some type of representation for further analysis and processing, this extracted information is known as feature vector. The process of converting voice signal to feature vector is done by signal-processing front end module.

Converting speech into a textual representation requires several stages. First, a microphone converts the acoustic vibrations into an analog signal. This analog signal is then altered to eliminate the high frequency components of the signal that lie outside the range of frequencies that the human ear can detect. The altered signal is then digitized using a sampling and quantization phase. The digitized waveform is

then partitioned into fixed-duration time-slices called frames, which are further compressed using one of several encoding schemes, to yield a stream of feature vectors. At this point, pre processing is complete, and recognition techniques can be applied to this representation of the audio input. These typically involve a search to determine the optimal path through a graph, and constitute by far the most time-consuming and complex stage of the process.

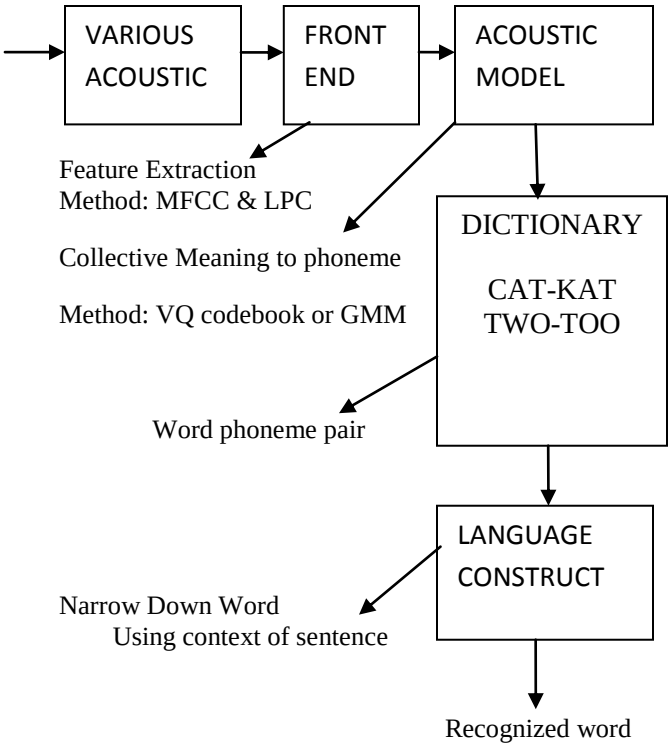

Fig-1: Block Diagram of Speech Recognition

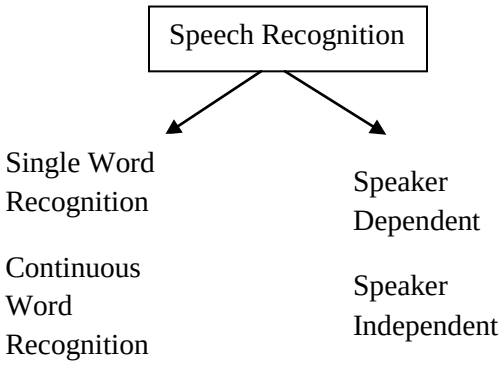

Fig-2: Classification of Speech Recognition

Isolated word recognizers usually require each utterance to have quiet on both sides of the sample window. It doesn't mean that it accepts single words, but requires a single utterance at a time.

Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content.

Speaker dependent systems are designed around specific speaker. Generally they are more accurate for the correct speaker, but much less accurate for other speakers. The system requires speaker consistent voice and tempo.

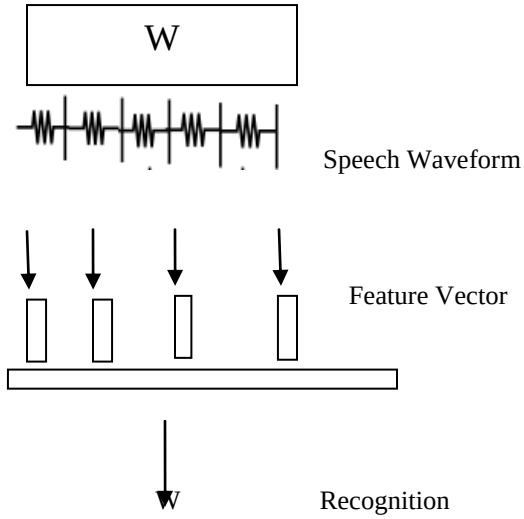

Fig-3: Concept: Sequence of Symbol

Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as a speaker independent systems and utilize the training techniques to adapt to the speaker to increase their recognition accuracy.

2. DYNAMIC TIME WARPING (DTW)

Dynamic time warping (DTW) is simple and effective algorithm in case of small to medium vocabulary. Dynamic time wrapping is dynamic programming means divide the problem and solve independently to obtain the solutions of sub problem at the end combine the result.
The purpose of Dynamic time warping is to produce warping function that minimizes the total distance between the respective points of the signal **[**8**].**

Dynamic time Warping can deal with different speaking speeds. DTW is used to compute the best possible alignment warp Q between T and R and associated distortion D (T, R).

An optimal alignment path between variable length sequence $T = \{t_1, t_2 \ldots\ldots t_n\}$ and $R = \{r_1, r_2 \ldots\ldots r_n\}$.

A particular alignment wrap q, align T and R via point to point mapping $Q = (Q_t, Q_r)$ of length K_q.

Optimal alignment minimize overall distortion
$D(T, R) = \min D_p(T,R)$

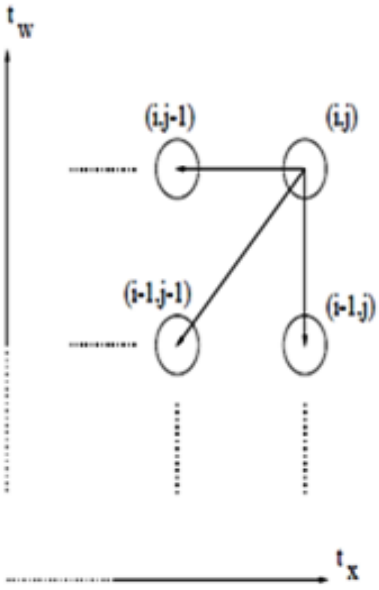

Fig -4: Local Paths for Alterative Grid Points.

$$\delta(i,j) = \min \begin{cases} \delta(i,j-1) + d(\vec{w}i, \vec{x}j) \\ \delta(i-1,j-1) + 2.d(\vec{w}i, \vec{x}j) \\ \delta(i-1,j) + d(\vec{w}i, \vec{x}j) \end{cases} \quad (1)$$

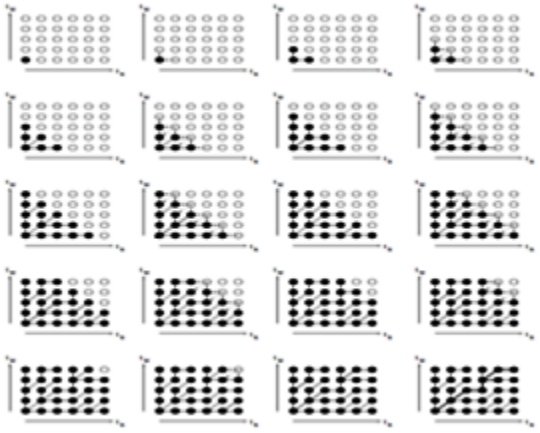

Fig 5: Iteration Paths for Finding Optimal Path

The accuracy of DTW based speech recognition system greatly relies on the quality of the prepared reference templates [12].

The reference template is created by three methods:

Each word's first occurrence in training material is used as template.

In training material number of occurrence of single word are averaged**.** In the averaging the weight of the training pattern is one and that of old template is equal to number of training pattern already created [9].

All occurrence of a word are averaged. During averaging certain amount of endpoint detection relaxation is allowed to compensate for endpoint detection errors [9**].**

DTW is applicable for speaker dependent system. The performance degrades if recording environment changes. DTW is time synchronous search.

3. HIDDEN MARKOV MODEL (HMM)

In speech recognition speech spectrum features are extracted by using speech analysis method as a front end and then speech recognition is carried out by HMM speech recognition with extracted feature vectors **[**11**].**

In HMM whenever a new input comes that is the voice signal, corresponding to that voice feature parameters are generated which is further used in learning process to create a new HMM model.

So with each new HMM model created for every word, during the testing phase, all these models are compared with the test word to find out the matching voice sample**[**5**].**

A.HMM Model

HMM can be defined by a set of N states, K observation symbols, and three probabilistic matrices **[**10**]**

$$\lambda = (A, B, \pi) \quad (2)$$

Where

N- Number of states

$Q = \{q_1, q_2, \ldots\ldots q_T\}$ - set of states

M - The number of symbols (observables)

$O = \{o_1, o_2 \ldots\ldots o_T\}$ - set of symbols

A - The state transition probability matrix

$a_{ij} = P(q_{t+1} = j | q_t = i)$

B- Observation probability distribution

$b_j(k) = P(o_t = k | q_t = j) \quad i \leq k \leq M$

π - the initial state distribution

B. HMM network topologies

Ergodic-In this model every state of model can be reached in a single step from every other state of the model. Generally it is used for recognition. The recognizer outputs a stream of states and each state represent part of speech.

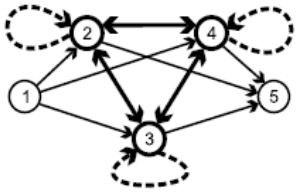

Fig 6: HMM Ergodic model

Bakis-It is also known as left to Right model. In Bakis model as the time increases states proceed from left to right. In this model a state once left cannot be revisited.

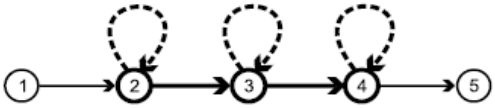

Fig 7 HMM Bakis Model

C. HMM for Speech

Speech is the output of an HMM, the problem is to find the most likely state sequence for a given observation of speech. Each state can be associated with sub phoneme, phoneme and sub word.

One HMM corresponds to one phoneme or word and for each HMM the most likely state sequence is determined. HMM with the highest match to observed speech is chosen.

4. ARTIFICIAL NEURAL NETWORK (ANN)

Artificial neural network is a computational model that has been developed as generalizations of mathematical models of biological nervous systems. It uses a set of processing elements that are called neurons or nodes.

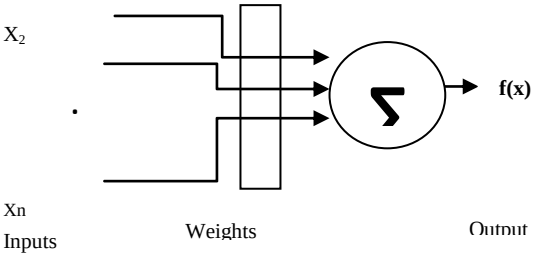

Fig 8: Basic Neuron Diagram

These nodes are interconnected in a network that can identify patterns in data as it is exposed to the data**.**

When the input arrives then for each input a neuron multiplies the input with its corresponding weight value and then sums this value with value of other inputs for this neuron and then passed the result to output function.

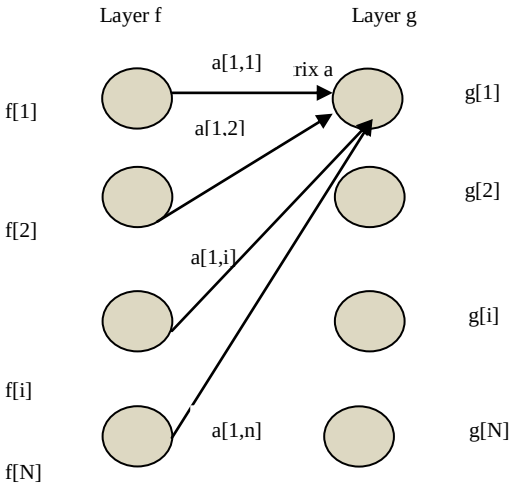

Fig 9: Networks of Neurons

Take a network of neurons which are organized in two distinct layers (called f and g). Each neuron in layer g is a linear unit and its output is the sum of its inputs a linear unit and its output is the sum of its inputs. Following calculations are performed by:

g [i] = a [1, 1] * f [1] +a [1, 2] * f [2].....etc (3)

A. Network Architecture

Network Components-A neuron has three states that are Idle, Ready to learn and committed.

In the initial state neurons are empty. So at the initialization neuron are idle in state. Their status is "Idle" except for the first one which is in "Ready to learn [5]. For every input the neuron are progressively used to store and to build knowledge and then become committed.

Segmentation Per Context -Context can divide the network into several small networks. It is more useful when processing multiple neurons [5].

Speed	3	15	18	28	55
category	6	4	4	2	9

Table 1 Response of Neurons

If choosing the maximum speed you will consider the input vector belonging to category 9; if choosing the minimum speed, you will consider category 6. As for different applications, you can choose different rules.

CONCLUSION

After conducting a thorough survey of these techniques we conclude that the above discussed techniques are taking advantage of technologies which are related to conversion of voice signal to frequency domain which helps to get more information about the signal by using wavelets decomposition and then processing the signal. Other technologies which have been surveyed include Hidden Markov Model (HMM) and Artificial Neural Network (ANN). HMM which take the concept of probabilities which are observable and non-observable for calculating the most probable data point match. On the other hand ANN utilizes the concept of distributed memory and parallel distributed processing.

REFERENCES

[1] A. Abraham, Artificial Neural Networks, Handbook for Measurement Systems Design, Peter Sydenham and Richard Thorn (Eds.), John Wiley and Sons Ltd, London, pp. 901-908, 2005.

[2]A.M. Othman, and M.H. Riadh, "Speech Recognition using Scaly Neural Networks", World Academy of Science, Engineering and Technology, 2008, Vol. 38, pp. 253-258

[3]A.R. Sukumar, A.F. Shah, and P.B. Anto, "Isolated question words recognition from speech queries by using Artificial Neural Networks", in proc. of IEEE 2nd International conference on Computing, Communication and Networking Technologies(ICCCNT), Karur, India, 2010, pp. 1-4.

[4] B. Lu, and Y. Wang, "Speech Recognition System Based on Multiple Neural Networks", in proc. of IEEE Sixth International Conference on Natural Computation (ICNC), Yantai, China, 2010, vol. 1, pp. 48-51.

[5] F.F. Zeng, and P.C. Shi, "Neural Network Design Based on Isolated Words", IEEE International Conference on Machine Learning and Cybernetics (ICMLC), 2011, Guilin, pp. 769-772.

[6] J.D. Tan, and H.N. Ting, "Malay Speaker Identification Using Neural Networks", in proc. of International Conference on Information Science and Technology (ICIST), 2011, Nanjing, pp. 476-479.

[7] J. Zhang, and M. Zhang, "A Speech Recognition Method Based Clustering Neural Network Integration", in proc. of International Conference on Electric Information and Control Engineering (ICEICE), Wuhan, China, 2011, pp. 1120-1122.

[8] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, 2010, Vol. 2, Issue 3, pp. 138-143.

[9] S. Haltsonen, "Recognition of isolated-word sentences from a large vocabulary using dynamic time warping methods", in proc. of IEEE Transactions on Acoustic Speech and Signal Processing (TASSP), Hong Kong, 2003, Vol. 33, Issue 4, pp. 1026-1027.

[10] S.S. Jarng, "HMM Voice Recognition Algorithm Coding", in the proc. of IEEE International Conference on Information Science and Applications (ICISA), Jeju Island, 2011, pp. 1-7.

[11] T. Kinjo, and K. Funaki, "On Hmm Speech Recognition Based On Complex Speech Analysis", in proc. of IEEE 32nd Annual Conference on Industrial Electronics, Paris, 2006, pp. 3477-3480.

[12] W.H. Abdulla, D. Chow, and G. Sin, "Cross Word Reference Template For DTW Based Speech Recognition System", in proc. of International Conference on Convergent Technologies for Asia-Pacific Region (TENCON), Bangalore, 2003,Vol. 4, pp. 1-4.