ISSN:0975-9646

Speech Recognition System; Challenges and Techniques

Parneet Kaur, Parminder Singh, Vidushi Garg

Dept. of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab 141006, India

***Abstract* -Pattern matching is an emerging subject for doing identification of various objects including speech and its parts. Initially these pattern matching techniques has been limited to identifying the speech patterns of the words that are separated with enough silence. It is easy to segment each word and extract its features which machine algorithm like Neural Network can learn and simulate, but if a sufficient interval of silence is not there, then there are connected words between each pause based on persons rate of speech, if for example person has spoken 5 types of words the connected combination might reached to 120 which would ultimately lead to a great challenge for pattern matching algorithm. In this paper we are reviewing such methodologies.**

Keywords*-*Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Artificial Neural Network (ANN)

I. INTRODUCTION

Speech is one of the most natural ways to interact. People are so comfortable with speech that we would also like to interact with our computers via speech, rather than having to resort to primitive interfaces such as keyboards and pointing devices and when it comes to computers it is no different. Speech recognition is a problem in the field of pattern recognition, which estimates the probability density function of each pattern to be recognized and then with the help of "Bayes theorem" identify the pattern which gives the highest likelihood for the observed speech data. Speech is a natural way of communication for people. The human vocal tract and articulators are biological organs with nonlinear properties, whose operation are not just under conscious control but also affected with factors varying from gender to upbringing to emotional state. As a result, vocalizations can vary with various factors such as their accent, pronunciation, articulation, roughness, nasality, pitch, volume, speed and many other features.

In other words Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands & control, data entry, and document preparation.

Speech Recognition and Voice Recognition are two different but these two terms are often used interchangeably, but they really should not be. They have distinct meanings. Imagine you answer the telephone, listen for a few seconds and then say "Caroline, can you call me back? We have a bad connection. I can barely hear you." You recognized your friend Caroline's voice. That is Voice Recognition. Speech Recognition is trying to understand the words being spoken.

Speech recognition system aims following:

Speech Recognition: aims to know the contents of the speech.

Speaker Recognition: aims to know the person who is talking.

Language Identification: aims to know the spoken language.

A. Types of Speech Recognition System

Speech System is divided into various types depending upon following:

Speaking Mode: Its means that how the words are spoken whether in isolated or in connected. An isolated word speech recognition system requires that the speaker pause briefly between words. It means single word. A connected word speech recognition system does not require that the speaker pause briefly between words. It means full sentences in which words are artificially separated by silence.

Speaking Style: It includes that whether the speech is continuous or spontaneous. Continuous means naturally spoken words. Systems can be evaluated on speech read from prepared scripts whereas spontaneous or extemporaneously generated, speech contains disfluencies, and is much more difficult to recognize than speech read from script. It is vastly more difficult because it tends to be peppered with disfluencies like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, laughter and moreover vocabulary is essentially unlimited, so the system must be able to deal intelligently with unknown words.

Vocabulary: It is simple to discriminate a small set of words, but error rates increase as the vocabulary size increases. example, 10 digits 0 to 9 can be recognized perfectly on the other hand vocabulary sizes of 200,5000 or 20000 have error rates of 3%,7% or 45%.even small vocabulary can be hard to recognize if it contains confusable words .

Enrollment: Enrollment is by two ways one is speaker dependent and speaker independent. In speaker dependent a user must provide samples of his or her speech before using them, a speaker dependent system is meant for use of a single speaker whereas speaker independent system is intended for use by any speaker.

B. Challenges of Speech Recognition system

Within Speaker variability: It means that various in the single speaker's way of speaking because of timing vary and speaking style.

Between Speaker variability: It means that various when two persons speaking with each other the way be vary with respect to accent variation, Voice Quality variation and individual characteristic variation.

Environment Variability: Speech Recognition is also affected by the environment factors such as Background noise and Microphone /Channel.

1.3 Speech-Recognition Engines Match a Detected Word to a Known Word Using One of the following Techniques [7].

Complete-word matching: The comparison between the incoming digital-audio signal and prerecorded template of the word. The preprocessing required by this technique is less than the sub word but it needs the pre record of all words that are to be recognized. It require large amount of memory to store the whole word.

Sub-word matching: The engine looks for phonemes after that it does pattern recognition on those. In order to process a phoneme a less amount of storage is required but processing time is less. In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand [7].

II. VARIOUS TECHNIQUES OF SPEECH RECOGNITION

A. Dynamic Time Warping

Dynamic time warping is a statistical approach, previously used to recognize speech but its use is displaced by more powerful and successful techniques such as Hidden Markov Model. Any type of data that can be represented linearly can be analyzed with the help of Dynamic time warping. Dynamic time warping algorithm is powerful for measuring similarity between two time series which may vary in time or speed [2].To find optimal match between two sequences i.e. input and reference template in that case DTW is powerful algorithm. The main principle of DTW is to compare two dynamic patterns and measure its similarity by calculating a minimum distance between them [5].In order to calculate minimum cost between input frame f_i and transition costs T_{xy}. A formula is applied to find the cost is by calculating the distance between the reference frame and the input frame.

$$\mathbf{D\ (i, j) = d\ (i, j) + min\ (D\ (i, j) + T_{10}, D\ (i, j) + T_{11}, D\ (i, j) + T_{01})} \qquad \mathbf{(1)}$$

Where D (i, j) is lowest cost to i, j
D (i, j) is lowest match cost.

The steps included in Dynamic time warping are as follow:
Record, Parameterize and store vocabulary of reference words.
Record test word to be recognized and parameterize.
Measure distance between test word and each reference word.
Choose reference word closest to test word.

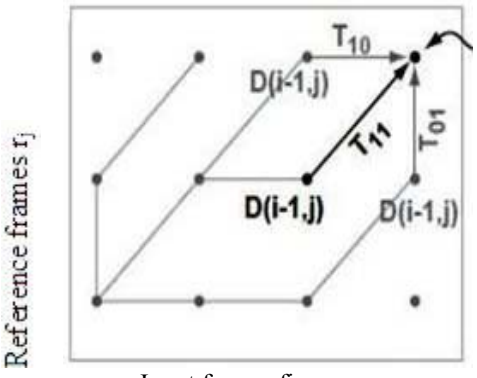

Fig 1 Optimal Alignment between Input Frame and Reference Frame.

The main problem in Dynamic time warping is to prepare the reference template. Previously it was prepared by choosing an example of each word that is to be recognized. It is considered as reference template but a single template is not sufficient as it is not possible to repetitively speak same word in similar manner as previously by the speaker, so to avoid this crossword reference template is used for creation of crossword reference template is with multiple examples. Then the average length of the extracted template is calculated. Next template with length nearest to average length is chosen to be best template [11].Initial reference is later template. Dynamic time wrapping time align the other templates. Final reference template is obtained by performing the average time aligned template across frame.

Dynamic time warping is used in small scale embedded speech recognition system such as those embedded in cell phones. The reason for this is its simplicity of hardware implementation of the Dynamic time warping engine which makes it suitable for many mobile devices. The training procedure in Dynamic time warping is very simple and fast as compared to the Hidden Markov Model and Artificial Neural Network.

B. Hidden Markov Model

A Hidden Markov Model is a statistical model of a sequence of feature vector observations. In HMM state sequences are hidden and the observations are probabilistic functions of the state. A Hidden Markov Model is a collection of states connected by transitions. The choice of transition and output symbol are both random, governed by probability distributions. The sequence of output symbols generated over time is observable, but the sequence of states visited over time is hidden from view.

HMM model can be described by these three set of parameters a, b and π and the model of N states and M observations referred to by:

$$\boldsymbol{\lambda = (A, B, \pi)} \qquad \mathbf{(2)}$$

Where A = $\{ a_{ij} \}$, B = $\{ b_j (w_k)\}$ and $1< I, j\leq N$ and $1\leq K \leq M$.

The three cases for HMM are Evaluation, Decoding and Training. The purpose of evaluation is to compute the probability of given sequence $O = O_1, O_2, O_3……..O_{t-1}, O_t$ with given HMM $\lambda = (A, B, \pi)$ that λ has generated the sequence O. Decoding calculates the most likely sequence of hidden states S_i of $O = O_1, O_2, O_3……..O_{t-1}, O_t$ that produced this observation sequence O. In learning the HMM parameters $\lambda = (A, B, \pi)$ are adjusted to maximize the probability to get the best model that represent certain set of observations.

The Strengths of HMM is its mathematical framework and its implementation structure. HMM method is fast in its initial training, and when a new voice is used in the training process to create a new HMM model [8].

The figure shows the general architecture of HMM. Oval shape represents random variable that can adopt a number of values. X (t) random variable whose value is hidden at variable time.

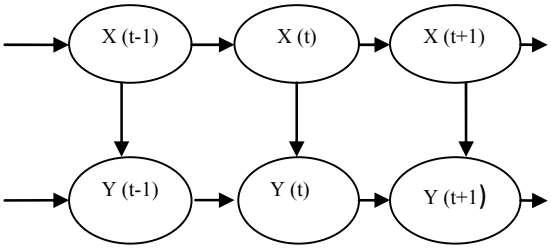

Fig 2 HMM Architecture

C. Artificial Neural Network

Neural Network has emerged as a field of study of Artificial Intelligence and engineering via the collaborative efforts of engineers, physicist's mathematicians, computer scientists and neuroscientists [4]. An Artificial Neural Network is a model that is composed of inputs, weights and outputs. It has a number of neurons which are connected or linked to each other with labeled edge is called weights. The main motive of Artificial Neural Network is to convert input to meaningful output. An Artificial neural network is a parallel system because all the neurons present in a layer perform computation at their own level and these computations allow for an entire layer to be generalized in single operation. Artificial Neural network is also called an adaptive system because of its property of changing its structure depending upon the external or internal information that flow through network.

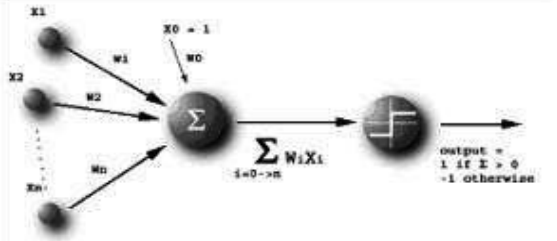

Fig 3 Model of a Perceptron [10]

A set of input connections multiplied by weight on edge then enter to the activation function to give output. The values w_1, w_2, w_3, w_4…. w_n are weight to determine the strength of input vector $X=[x_1, x_2, x_3……x_n]^T$. Each is multiplied by associated weight of the neuron connection X_tW. The positive weight excites and negative inhibits the output. The activation function f performs a mathematical operation on the signal output. It should be taken based upon the type of problem solved by the network. Artificial neural network has number of layers i.e. input, output and hidden layer. The main layers in ANN are input layer, output layer and hidden layers. Hidden layers are the layer between input layer and output layer

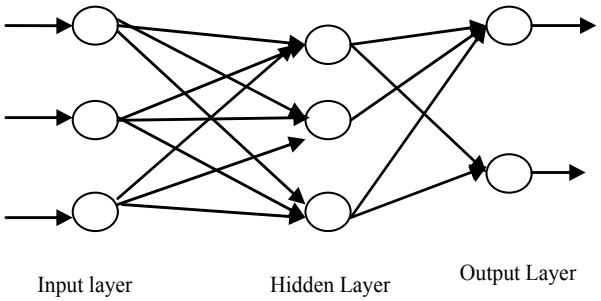

Fig 4 Architecture of Artificial Neural Network

Artificial neural network is classified as Supervised Learning, Unsupervised Learning and Reinforcement learning

Supervised Learning- It is a learning Procedure based on error. In it with respect to input output is known. If the output does not match with the desired output then modification should be made. The main motive is to minimize the difference between answer of network and expected value.

Unsupervised Learning-It is a learning in which no teacher is present. It depends upon the output. It does not have any external agent to adjust weight of common link to their neurons. It is capable of self organizing.

Reinforcement Learning-In this type of learning the learner not knows about the action to be taken but discover which action is to take to yield most reward by trial method. Action may affect the immediate reward and also to next situation and through that it affects all next rewards. The two characteristics i.e. trial and error search and delayed reward distinguish features of reinforcement learning.

III. CONCLUSION

After conducting a review of such methodologies we have found that lot of work needs to be done on regional languages in which the local dialect has necessarily connected words as parts of speech. Therefore we need robust algorithm which would help us to achieve this objective based on profile voice of each person. We can also design a database and interface to print the recognized words spoken by the user.

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] D.G. Bhalke, C.B.R. Rao, and D. S. Bormane, "Dynamic Time Warping Technique for Musical Instrument Recognition for Isolated Notes", IEEE International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), 2011, Tamil Nadu, pp.768-771.

[3] G.T. Tsenov, and V.M. Mladenov, "Speech Recognition Using Neural Networks", in the proc. of IEEE 10th Symposium on Neural Network Applications in Electrical Engineering, 2010, Serbia, Vol. 20, pp. 181-186.

[4] J.S. Sengar, and N. Sharma, "Design a Neural Network Based on Hebbian Learning and ART", International Journal of Computer Science and Technology, 2011, Vol. 2, Issue 4, pp.157-160.

[5] 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.

[6] N. Seman, Z.A. Bakar, and N.A. Bakar, "Measuring The Performance Of Isolated Spoken Malay Speech Recognition Using Multi-Layer Neural Networks", in proc. of IEEE International Conference on Science and Social Research(CSSR), Kuala Lumpur, Malaysia, 2010, pp. 182-186.

[7] R. Tadeusiewicz, "Speech In Human Interaction", in proc. of IEEE 3rd Conference on Human System Interactions (HSI), Poland, 2010, pp. 1-12.

[8] 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.

[9] Thiang and S. Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot", in proc. of IEEE International Conference on Information and Electronics Engineering (ICIEE), Singapore, 2011, vol. 6, pp. 179-183.

[10] V.A. Keturi, "Speech Recognition Based on Artificial Neural Networks", Helsinki University of Technology, Finland, 2004.

[11] 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.