ISSN:0975-9646

Survey on Speech Recognization Techniques.
(Natural Language Processing)

Anushree R. Pore
Computer Science And Engineering
G.H.R.C.E.M
Amravati, Maharashtra, India

Prof. Amit Sahu
Computer Science And Engineering
G.H.RC.E.M
Amravati, Maharashtra, India

Abstract— **Natural Language Processing is a technique where machine can become more human and thereby reducing the distance between human being and the machine can be reduced. Therefore in simple sense NLP makes human to communicate with the machine easily. There are many applications developed in past few decades in NLP. Spoken language recognition refers to the automatic process through which we determine or verify the identity of the language spoken in a speech sample. And these are very useful in everyday life for example a machine that takes instructions by voice. Humans are born with the ability of discrimination to discriminate between spoken languages as part of human intelligence There are lots of research groups working on this topic to develop more practical are useful systems. Natural Language Processing holds great promise for making computer interfaces that are easier to use for people, since people will hopefully be able to talk to the computer in their own language, rather than learn a specialized language of computer commands. Today, automatic spoken language recognition is no longer a part of science fiction. We have seen it being deployed for practical uses For programming, however, the necessity of a formal programming language for communicating with a computer has always been taken for granted. We would like to challenge this assumption. Speech is more natural and efficient communication method between human, automatic Speech recognization will continue to find applications. We believe that modern Natural language processing techniques can make possible the use of automatic recognization in audio, speech and language processing to express the programming ideas and thus drastically increases the accessibility of programming to non-expert users. To demonstrate the feasibility of the Natural language Programming and this paper tackles what are perceived to be some of the hardest cases.**

Index Terms—Speech recognization, Dynamic time wrapping, Hidden markov model, Vector Quantization, Egrodic HMM's, Artificial neural network.

INTRODUCTION

Natural Language Processing holds great promise for making computer interfaces that are easier to use for people, since people will be able to talk to the computer in their own language, rather than to learn a specialized language of computer commands. For programming, the necessity of the formal programming language to communicate with a computer has always been taken for granted. We would like to challenge this type of assumption. We believe that modern Natural Language Processing techniques can make possible the use of natural language to (at least partially) express programming ideas, NLP include a number of fields of techniques, such as Dynamic Time Warping (DTW), Hidden Markov Models (HMM) ,Vector Quantization (VQ), Ergodic-HMM's, Artificial Neural Networks (ANN) Long-Term Statistics, machine translation, natural language text proc easing and summarization and some kind of user interfaces, multilingual and cross language information retrieval (CLIR), and so on. thus drastically there are some methods such as Automatic speech recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time and Audio speech language processing . Currently, most speech recognition systems are based on hidden Markov models (HMMs), a statistical framework that supports both acoustic and temporal modeling. Despite their state of arts performances. HMM make a number of suboptimal modeling assumptions that limit their potential effectiveness. ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. ASL(Audio Speech Language) field is its intimate connection ,more than any other technical fields to human-centric information processing and to "artificial intelligence" which is coming of age. Indeed, the tidal wave of human-centric computing is upon us in various forms, including natural user interface between human and machines/devices, and there is a strong need to improve related fundamental technologies in all aspects. Together with the new technological trends in mobility and in social computing, closely tied to audio, speech, and human language processing, we have a unique opportunity to make our publication the best among the best. Having a machine to understand fluently spoken speech has driven speech research for more than 50 years. Although all above technology is not yet at the point where machines understand all speech, in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services. The ultimate goal of ASR research is to allow a computer to recognize in real-time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent. Thus a natural language interface should be able to translate the natural language statements into appropriate actions for the system.

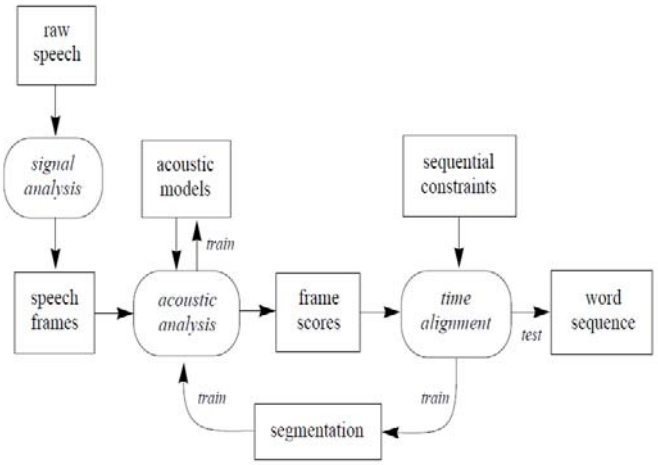


fig : Structure of a standard speech recognition system

SOUND TYPE COMPARISON

Research into speech recognition began by reviewing the literature and finding techniques that had previously been used for speech/speaker recognition. It was found that six techniques are commonly used for speech recognition or have been used for this domain in the past. These are as follows:

- Dynamic Time Warping (DTW)
- Hidden Markov Models (HMM)
- Vector Quantization (VQ)
- Ergodic-HMM's
- Artificial Neural Networks (ANN)

Dynamic Time Warping (DTW)

[2] Bellagarda, J. and Nahamoo states that Dynamic time warping (DTW) is a well-known technique to find an optimal alignment between two given sequences under certain restriction. DTW has been used to compare different speech patterns in automatic speech recognition. we show how DTW can be employed to identify all subsequence within a long data stream that are similar to a given query sequence. One of the earliest approaches to isolated word speech recognition was to store a prototypical version of each word in the vocabulary and compare incoming speech with each word by taking the closest match with it.

Classical DTW

The objective of DTW is to compare two (time-dependent) sequences $X := (x1, x2, \ldots, xN)$ of length $N \in \mathrm{N}$ and $Y := (y1, y2, \ldots, yM)$ of length $M \in \mathrm{N}$. These sequences may be discrete signals (time-series) or, more generally, feature sequences sampled at equidistant points in time. In the following, we fix a *feature space* denoted by F. Then $xn, ym \in F$ for $n \in [1 : N]$ and $m \in [1 : M]$. To compare two different features $x, y \in F$, one needs a *local cost measure*, sometimes also referred to as *local distance measure*, which is defined to be a function $c : F \times F \rightarrow \mathrm{R} \geq 0$. (4.1) Typically, $c(x, y)$ is small (low cost) if x and y are similar to each other, and otherwise $c(x, y)$ is large (high cost). Evaluating the local cost measure for

4 Dynamic Time Warping

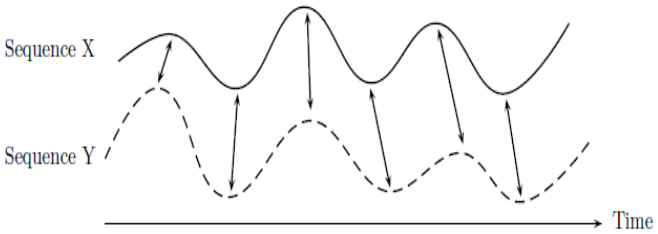
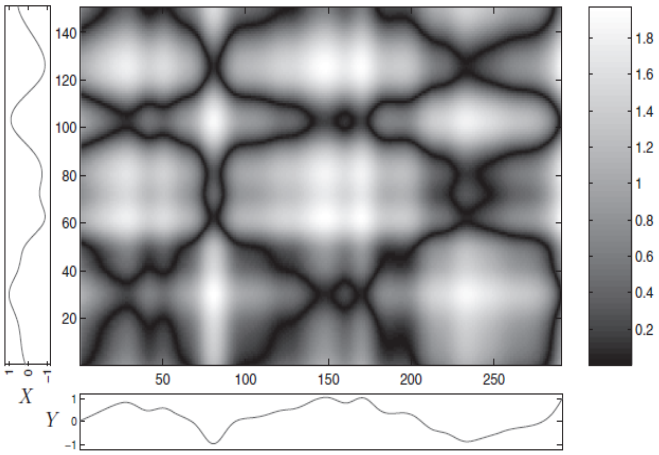


fig: Time alignment of two time-dependent sequences. Aligned points are indicated by the arrows.



Cost matrix of the two real-valued sequences X (*vertical axis*) and Y (*horizontal axis*) using the Manhattan distance (absolute value of the difference) as local cost measure c. Regions of low cost are indicated by *dark colors* and regions of high cost are indicated by light colors.

Cost matrix of the two real-valued sequences X (*vertical axis*) and Y (*horizontal axis*) using the Manhattan distance (absolute value of the difference) as local cost measure c. Regions of low cost are indicated by *dark colors* and regions of high cost are indicated by light colors.

Hidden Markov Models (HMM)

[3]Bahl, L, Brown states that Hidden Markov Models (HMMs) provide a simple and effective framework for modelling time-varying spectral vector sequences. As a consequence, almost all present day large vocabulary continuous speech recognition (LVCSR) systems are based on HMMs. The principal components of a large vocabulary continuous speech recogniser are illustrated. The input audio waveform from a microphone is converted into a sequence of fixed size acoustic vectors $\boldsymbol{Y}1{:}T = \boldsymbol{y}1, \ldots, \boldsymbol{y}T$ in a process called feature extraction. The decoder then attempts to find the sequence of words $\boldsymbol{w}1{:}L = w1,\ldots,wL$ which is most likely to have generated $\boldsymbol{Y}$, i.e. the decoder tries to find $\hat{}\boldsymbol{w} = \arg\max_{\boldsymbol{w}} \{P(\boldsymbol{w}|\boldsymbol{Y})$. However, since $P(\boldsymbol{w}|\boldsymbol{Y})$ is difficult to model directly,1 Bayes' Rule is used to transform into the equivalent problem of finding: $\hat{}\boldsymbol{w} = \arg\max_{\boldsymbol{w}} \{p(\boldsymbol{Y}|\boldsymbol{w})P(\boldsymbol{w})\}$. The likelihood $p(\boldsymbol{Y}|\boldsymbol{w})$ is determined by an *acoustic model* and the prior $P(\boldsymbol{w})$ is determined by a language model.

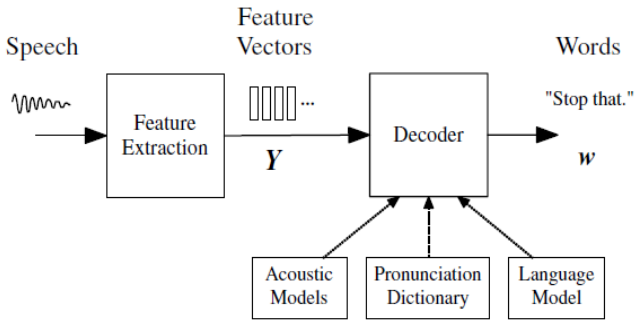


fig: Architecture of HMM-based recogniser.

represented by the acoustic model is the *phone*. For example, the word "bat" is composed of three phones /b/ /ae/ /t/. About 40 such phones are required for English. For any given ***w***, the corresponding acoustic model is synthesised by concatenating phone models to make words as defined by a pronunciation dictionary. The parameters of these phone models are estimated from training data consisting of speech waveforms and their orthographic transcriptions. The language model is typically an *N*-gram model in which the probability of each word is conditioned only on its $N-1$ predecessors. The *N*-gram parameters are estimated by counting *N*-tuples in appropriate text corpora. The decoder operates by searching through all possible word sequences using pruning to remove unlikely hypotheses thereby keeping the search tractable. When the end of the utterance is reached, the most likely word sequence is output. Alternatively, modern decoders can generate lattices containing a compact representation of the most likely hypotheses.

Vector Quantization

Dr. H. B. Kekre, Ms. Vaishali Kulkarni states that the results of a case study carried out while developing an automatic speaker recognition system are presented in this paper. The Vector Quantization (VQ) approach is used for mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook.

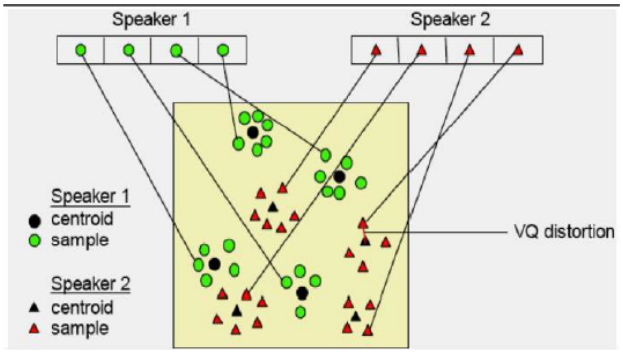


fig: Vector Quantization Form.

Ergodic-HMM's

[7]Shing-Tai P states that, HMM is the best method for modeling the speech signal because this method is represented in the form of states which the characteristics of the speech signal is also represented in the form of states.Left- right model of HMM models commonly used for modeling the isolated words but in this research, we used the ergodic model to modeling speech signal. Ergodic HMM can directly model the sequence of verbal units (phonemes, words) articulated more than left to-rigth model that have been used. During articulation, only certain verbal units may follow each other, but one may progress from one verbal unit if enough intermediate states are allowed. In this paper, genetic algorithm was applied to optimize the Baum-Welch algorithm in Ergodic HMM. The result between HMM system and hybrid HMM-GA was compared to analyze how GA can improve the accuracy in hybrid HMM-GA.

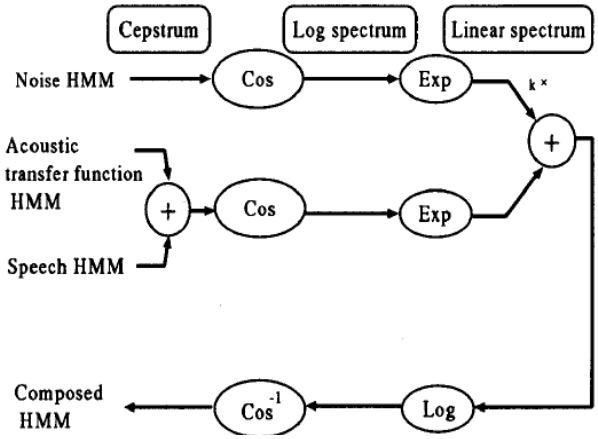


fig: Ergodic HMM Form.

Artificial Neural Networks (Ann)

[8]Bourlard, H., Morgan states that, the brain's impressive superiority at a wide range of cognitive skills, including speech recognition, has motivated research into its novel computational paradigm since the 1940's, on the assumption that brain like models may ultimately lead to brain like performance on many complex tasks. This is an such fascinating research area is now known connectionism, or the study of artificial neural network. Neural networks can indeed form the basis for a general purpose speech recognition system, and that neural networks offer some clear advantages over conventional techniques. Neural networks are usually used to perform static pattern recognition, that is, to statically
map complex inputs to simple outputs, such as an N-ary classification of the input patterns. Moreover, the most common way to train a neural network for this task is via a procedure called *back propagation* (Rumel hart et al, 1986), whereby the network's weights are modified in proportion to their contribution to the observed error in the output unit activations (relative to desired outputs). To date, there have been many successful applications of neural networks trained by back propagation. For instance:

- *NETtalk* (Sejnowski and Rosenberg, 1987) is a neural network that learns how to pronounce English text. Its input is a window of 7 characters (orthographic text symbols), scanning a larger text buffer, and its output is a

phoneme code (relayed to a speech synthesizer) that tells how to pronounce the middle character in that context. During successive cycles of training on 1024 words and their pronunciations, NET talk steadily improved is performance like a child learning how to talk, and it eventually produced quite intelligible speech, even on words that it had never seen before.

- *Neuro gammon* (Tesauro 1989) is a neural network that learns a winning strategy for Backgammon. Its input describes the current position, the dice values, and a possible move, and its output represents the merit of that move, according to a training set of 3000 examples hand-scored by an expert player. After sufficient training, the network generalized well enough to win the gold medal at the computer olympiad in London, 1989, defeating five commercial and two non-commercial programs, although it lost to a human expert.
- *ALVINN* (Pomerleau 1993) is a neural network that learns how to drive a car. Its input is a coarse visual image of the road ahead (provided by a video camera and an imaging laser rangefinder), and its output is a continuous vector that indicates which way to turn the vehicle's wheel. The system learns how to drive by observing how a person drives. ALVINN has successfully driven at speeds of up to 70 miles per hour for more than 90 miles, under variety of different way conditions.
- *Handwriting recognition* (Le Cun et al, 1990) based on neural networks has been used to read ZIP codes on US mail envelopes. Size-normalized images of isolated digits, found by conventional algorithms, are fed to a highly constrained neural network, which transforms each visual image to one of 10 class outputs. This system has achieved 92% digit recognition accuracy on actual mail provided by the US Postal Service. A more elaborate system by Bodenhausen and Manke (1993) has achieved up to 99.5% digit recognition accuracy on another database. Speech recognition, but obvious has been another proving ground for neural networks. Researchers quickly achieved excellent results in such basic tasks as voiced/unvoiced discrimination (Watrous 1988), phoneme recognition (Waibel et al, 1989), and spoken digit recognition (Franzini et al, 1989). However, in 1990, when this thesis was proposed, it still remained to be seen whether neural networks could support a large vocabulary, speaker independent, continuous speech recognition system.

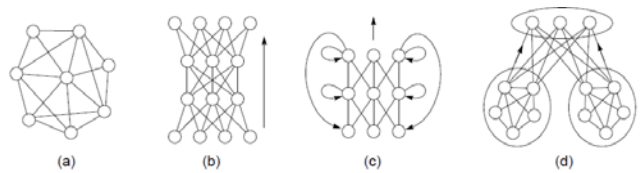


fig: Neural Network Topology

Best Technique

The field of speech recognition has seen tremendous activity in recent years. Hidden Markov Models still dominate the field, but many researchers have begun to explore ways in

which neural networks can enhance the accuracy of HMM-based systems. Researchers into NN-HMM hybrids have explored many techniques (e.g., frame level training, segment level training, word level training, global optimization), many issues (e.g., temporal modeling, parameter sharing, context dependence, speaker independence), and many tasks (e.g., isolated word recognition, continuous speech recognition, word spotting). These explorations have especially proliferated since 1990, when this thesis was proposed, hence it is not surprising that there is a great deal of overlap between this thesis and concurrent developments in the field. The remainder of this thesis will present the results of my own research in the area of NN-HMM hybrids. Neural networks can be trained to compute smoother, nonlinear, and nonparametric functions from any input space to any output space. Two very general types of functions are *prediction* and *classification*. In a predictive network, the inputs are several frames of speech, and the outputs are a prediction of the next frame of speech; by using multiple predictive networks, one for each phone, their prediction errors can be compared, and the one with the least prediction error is considered the best match for that segment of speech. By contrast, in a classification network, the inputs are again several frames of speech, but the outputs directly classify the speech segment into one of the given classes.

Predictions of frame *t* (separate networks) Classification of frames 1...*t*

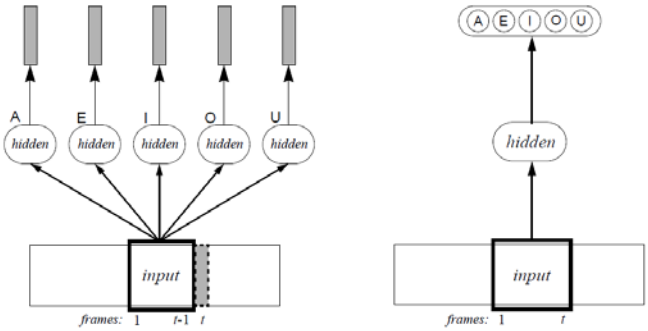


fig : Prediction versus Classification

In the course of our research, we have investigated all of these approaches. Predictive neural networks will be treated as best technique.

CONCLUSION

Future Advances for Voice Recognition**:**

Body Language + Facial Expression + Voice Recognition: Speech recognition allows converting speech into text, making it easier both to create and to use information. Speech is easier to generate, it's intuitive and fast, but listening to speech is slow, it is hard to index speech, and easy to forget. Text is easier to store, process and also consume both for computer and for human but writing text is slow and requires some intention. today, there are having robotic android projects in the works in Japan and in the US, facial expression and mirroring, or more is very

popular. The basic goal is for the human that interfaces with the system to create an emotional attachment with machine. Voice Recognition systems that can read body language and facial expression also be used for threat assessment at lets we can say airports, border crossings and replace human workers at those locations or choke points.

If you smiling at robotic android and it smiles back at you, while you are having a conversation, this up the emotional value of the conversation to the human. Perhaps the system might be get starts complimenting you. If you are persnickety towards the system, maybe it will mirror those responses or reciprocate an angry response or work to diffuse the situation, it all depends on its programming, but you can see the advances and potential applications and the trends going forward.

If you will recall Hal the famous science fiction computer, it said that "I sense hostility in your voice Dave." Perhaps since this was once in a science fiction work, human scientists today are trying to make it so. And right now, we are there, with having this technology, CRM Voice Recognition Software can also sense emotion, hesitation, aggression, hostility, anger, etc. So, within five years we will see these features in more and maximum applications.

Hap tics is another field of science, which lends itself good to merging of Voice Recognition and Facial expression emotional recognition. Perhaps robots of the future will look like humans and mimic their characteristics. A robot that also feels strong handshakes and firm grips along with a self-confident voice of an individual with an earned ego might elevate the trust factor a notch.

Increasing the confidence in the individual's ability to perform - will voice recognition software combined with these other technologies replace corporate Human Resources, Folks are already thinking here; 10-15 years out, but not without ruffling a few options. Being replaced by a computer, robot or machine has caused many a conflict in the past, so plot thickens and more barriers are must be seen.

APPLICATIONS

In the past people mostly imagined speech recognition directly producing the end-result, e.g. a dictated document or a computer performing a command. This is a limited perspective, as availability of speech recognition is likely to make possible much more varied applications. For example, speech recognition will likely be used.

- to send instant messages.
- to annotate and to comment.
- to keep real-time transcripts during conversations.
- to instruct and answer computers in a hands-free environment. (while driving; see Driving Cars, though)
- eventually, for general computer interaction; the Linguistic User Interface (LUI)

One area of application of NLP that has drawn much research attention, but where the results are yet to reach the general public with an acceptable level of performance, is the natural language question-answering system. Based on progress various application system have been developed using dictation and spoken dialogue technology. Using speech recognization technology, broadcast news automatically indexed, producing wide range of capabilities for browsing news archives interactively. One of the most important applications is information extraction and retrieval. Hence, speech recognization system is most efficient and natural method to communicate with the humans.

REFERENCES

[1] Austin, S, Zavaliagkos, G., Makhoul, J, and Schwartz, R. (1992). Speech Recognition Using Segmental Neural Nets. In *Proc. IEEE International Conference on Acoustics,Speech, and Signal Processing, 1992.*

[2] Bellagarda, J. and Nahamoo, D. (1988). Tied-Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1988.*

[3] Bahl, L., Brown, P., De Souza, P., and Mercer, R. (1988). Speech Recognition with Continuous-Parameter Hidden Markov Models. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1988.*

[4] Bourlard, H. and Wellekens, C. (1990). Links Between Markov Models and Multilayer Perceptrons. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(12), December 1990. Originally appeared as Technical Report Manuscript M-263, Philips Research Laboratory, Brussels, Belgium, 1988

[5] Dr. H. B. Kekre, Ms. Vaishali Kulkarni, "*Speaker Identification by using Vector Quantization*", International Journal of Engineering Science and Technology, Vol. 2(5), 2010, 1325-1331

[6] Jesus Savage, Carlos Rivera, Vanessa Aguilar, "*Isolated Word Speech Recognition Using Vector Quantization Techniques and Artificial Neural Networks*", Facultad de Ingenieria ,Departamento de Ingeniería en Computación ,University of Mexico,UNAM,Mexico City C.P. 04510,Mexico

[7] Shing-Tai P, Ching-Fa C, Jian-Hong Z. *Speech Recognition via Hidden Markov Model and Neural Network Trained by Genetic Algorithm*. Ninth International Conference on Machine Learning and Cybernetics. Qingdao, 11-14 July 2010.

[7] Shing-Tai P, hih-Hung W, Shih-Chin L. The Application of Improved Genetic Algorithm on The training of Neural Network for Speech Recognition. *IEEE Transactions on Neural Network*. 2007.

[8] Bourlard, Morgan, N, Wooters, C., and Renals, S. (1992). CDNN: A Context Dependent Neural Network for Continuous Speech Recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992.*

[9] Burr, D. (1988). Experiments on Neural Net Recognition of Spoken and Written Text. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36, 1162-1168.

[10] Doddington, G.(1989). Phonetically Sensitive Discriminants for Improved Speech Recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1989.*

[11] S Kwong, Chau W. Analysis of Parallel Genetic Algorithms on HMM Based Speech Recognition System. City University of Hong Kong. *IEEE Transactions on Consumer Electronics.* 1997; 43

[12] Carpenter, G. and Grossberg, S. (1988). The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network. *Computer* 21(3), March 1988