Efficient Classification of SOM – Based Speech Recognition System

R.L.K.Venkateswarlu
Sasi Institute of Technology and Engineering,
Tadepalligudem, INDIA.

R. Vasantha Kumari
Perunthalaivar Kamarajar Arts College,
Puducherry - 605 107

A.K.V.Nagayya
Sasi Institute of Technology & Engineering,
Tadepalligudem, INDIA.

ABSTRACT

In this paper, an attempt is made to study speech recognition system with classifiers Self Organized Maps, Multilayer Perceptron, Radial Basis Function neural networks, Modular neural network, Time Lagged neural network and to develop SOM based speech recognition system. The training parameters varies as the classifier varies. In this paper a novel SOM - based speech recognition system is developed to find the nearest neighbour classification by using Euclidean and weighted Euclidean weighted distances. In order to find out the outlier of the speakers, Nearest neighbour classification by using Euclidean and Weighted Euclidean distances are developed. The promising results are obtained with good degree of accuracy.

General Terms

Speech Recognition, Nearest Neighbour Classification, Classifiers, Euclidean distance, Weighted Euclidean distance.

Keywords: Pitch, Intensity, Mel-frequency cepstral coefficient, Linear predictive coefficient, Recognition rate.

1. INTRODUCTION

Speech is a human's most efficient communication modality. Beyond efficiency, humans are comfortable and familiar with speech. Other modalities require more concentration, restrict movement and cause body strain due to unnatural positions Research work on English speech recognition, although lagging that other language, is becoming more intensive than before and several researches have been published in the last few years [5]. Automatic speech recognition is a process by which a machine identifies speech. The conventional method of speech recognition insist in representing each word by its feature vector & pattern matching with the statistically available vectors using neural network. The promising technique for speech recognition is the neural network based approach. Artificial Neural Networks, (ANN) are biologically inspired tools for information processing [6]. The classic methods based on multilayer perceptron use the Time Delay Neural Network, it is the first model used by Weibel in the speech recognition domain [Berthold, M.R.,]. But the problem was the hard time processing and the adjustment of parameters that become a laborious stain for the new applications. In the opposite, the RBF networks don't require a special adjustment and the training time becomes shorter with regard to the Time Delay Neural Network. But the problem of RBF is the shift invariant in time [Berthold, M.R.,]. An outlier is a pattern that is in some way unlike the other patterns of its class, A class is represented by prototype. Outliers are rare. The most of the time our class is same as that of our neighbours has been applied to a procedure called the nearest neighbour classification. The more training patterns there are, the more distances have to be calculated, and consequently the computation required increases, thus slowing down the process of classification. The classification which is a measure of speech data analysis predicts categorical labels (classes), predication models continuous-valued functions. Nearest- neighbour classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all of the training tuples are stored in an n-dimensional space

2. SYSTEM CONCEPT

2.1 Dataset

Speech dictation of the word "Jaundice" is given and spelt as "Jaundice", "Zaundice", "Zaundis", "Jaundis", Johndis", Johndice", "Jhondis", "Jaandis", "Zaundees", Jaundees", "Jandice", "Jowndis" depending on the understanding capabilities (of the speakers). These words are then pronounced in a quite environment and uttered by five speakers (3Male, 2Female).

2.2 Pre processing

The speech signals are recorded in a low noise environment with good quality recording equipment. The signals are samples at 11kHz. Reasonable results can be achieved in isolated word recognition when the input data is surrounded by silence.

2.3 Sampling Rate

Samples are chosen with sampling rate 11kHz, which is adequate to represent all speech sounds.

2.4 Windowing

In order to avoid discontinuities at the end of speech segments the signal should be tapered to zero or near zero and hence reduce the mismatch. To the given 12 Mel-Frequency coefficients, and for time 0.005 seconds, a window length of 0.015 is selected by the Praat Object software tool.

2.5 Features Extraction

2.5.1 Linear Predictive Cepstral Coefficients

The goal of feature extraction is to represent speech signal by a finite number of measures of the signal. This is because the entirety of the information in the acoustic signal is too much to process, and not all of the information is relevant for specific tasks. In present Speech Recognition systems, the approach of feature extraction has generally been to find a representation that

is relatively stable for different examples of the same speech sound, despite differences in the speaker or various environmental characteristics, while keeping the part that represents the message in the speech signal relatively intact.

Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive mode. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters.

LPC analyzes the speech signal by estimating the formats, removing their effects from the speech signal, and estimating the Intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modelled signal is called the reside.

The number which describe the Intensity and frequency of the buzz, the formants, and the reside signal, can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal. Use the formants to create a filter (which represents the tube), and run the sources through the filter, resulting in speech.

Because speech signals vary with time, this process is done on short chunks of the speech signal, which are called frames; generally 30 to 50 frames per second give intelligible speech with good compression.

LPC is frequently used for transmitting spectral envelope information, and as such it has to be tolerant of transmission errors. Transmission of the filter coefficients directly is undesirable, since they are very sensitive to errors. In other words, a very small error can distort the whole spectrum, or worse, a small error might make the prediction filter unstable.

LPC is generally used for speech analysis and resynthesis. It is used as a form of voice compression by phone companies, for example in the GSM standard. It is also used for secure wireless, where voice must be digitized, encrypted and sent over a narrow voice channel.

In the LPC analysis one tries to predict x_n on the basis of the p previous samples,

$$x_n^{'} = \sum a_k \, x_{n-k}$$

then $\{a_1, a_2, ..., a_p\}$ can be chosen to minimize the prediction power Q_p where

$$Q_p = E\left[\left| x_n - x_n^{'} \right|^2 \right]$$

Linear Predictive Coding is used to extract the LPCC coefficients from the speech tokens. The LPCC coefficients are then converted to cepstral coefficients. The cepstral coefficients are normalized in between 1 and -1. The speech is blocked into overlapping frames of 20ms every 10ms using Hamming window. LPCC was implemented using the autocorrelation method. A drawback of LPCC estimates is their high sensitivity to quantization noise. Convert LPCC coefficients into cepstral coefficients where the cepstral order is the LPCC order and to decrease the sensitivity of high and low-order cepstral coefficients to noise, the obtained cepstral coefficients are then weighted. 16 Linear Predictive Cepstral Coefficients are considered for windowing. Linear Predictive Coding analysis of speech is based on human perception experiments. Sample the signal with 11 kHz. Frames are obtained for each utterance of the speaker form Linear Predicitve Cepstral Coefficients.

2.5.2 Mel-Frequency Cepstral Coefficients

Feature extraction consists of computing representations of the speech signal that are robust to acoustic variation but sensitive to linguistic content. The Mel-filter is used to find band filtering in the frequency domain with a bank of filters. The filter functions used are triangular in shape on a curvear frequency scale. The filter function depends on three parameters: the lower frequency, the central frequency and higher frequency. On a Mel scale the distances between the lower and the central frequencies and that of the higher and the central frequencies are equal. The filter functions are

$H(f)=0 \;\; for \, f \le f_l \;\; and \;\; f \ge f_h$

$H(f)=(f-f_l)/(f_c-f_l) \;\; for \;\; f_l \le f \le f_c$

$H(f)=(f_h-f)/(f_h-f_c) \;\; for \;\; f_c \le f \le f_h$

Mel - Frequency cepstral coefficients are found from the Discrete Cosine Transform of the

Filter bank spectrum by using the formula given by Davis and Mermelstein[1980].

$$c_i = \sum_{j=1}^{N} P_j \cos(i\pi / N(j-0.5))),$$

Pj denotes the power in dB in the jth filter and N denotes number of samples.

12 Mel- Frequency coefficients are considered for windowing. Mel-Frequency analysis of speech is based on human perception experiments. Sample the signal with 11 kHz, apply the sample speech data to the mel-filter and the filtered signal is trained. Frames are obtained for each utterance of the speaker form Mel-Frequency Cepstral Coefficients.

2.5.3 Pitch

Pitch, in speech, the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation.

Pitch is the property of voice and is determined by the rate of vibration of the vocal cords. The greater the number of vibrations per second, the higher the Pitch. The rate of vibration, in turn, is determined by the length and thickness of the vocal cords and by the tightening or relaxation of these cords.

The voice control is dependent largely upon emotional control. When human get excited or frightened, unconsciously the muscles around your voice box or larynx are tightened. The resulting tension in the vocal cords, according to the science of sound, produces a greater frequency of vibration and consequently a higher Pitch. It is an indication of lack of mental poise if you habitually speak in a voice Pitched too high. Frames are obtained for each utterance of the speaker form Intensity.

2.5.4 Intensity

Vocal Intensity, the major vocal attribute, depends primarily on the amplitude of vocal cord vibrations and thus on the pressure of the subglottic airstream. The greater the expiratory effort, the greater the vocal volume. Another component of vocal Intensity is the radiating efficiency of the sound generator and its

superimposed resonator. The larynx has been compared to the physical shape of a horn. The Intensity (or energy flow) of a sound wave is the power (in energy/sec) transmitted through an area of $1m^2$ oriented perpendicular to (normal to) the propagation direction of the wave. Almost everyone knows that if they move away from a constant sound source, they perceive a decrease in loudness. Consider the following example:

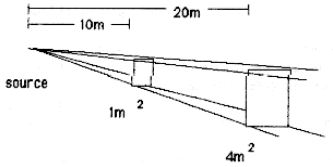

Fig 1: Decrease in loudness over distance

Assume that a sound from a source propagates through $1m^2$ of air at 10m from the source. Looking at the diagram we can see that the power that is concentrated over $1m^2$ at 10m from the source, is spread over a larger area at the distance of 20m. The same amount of energy is spread over a larger area, so the Intensity has decreased. Specifically, the area at 20m is $4m^2$ which is 4 times the area at 10m ($1m^2$), making the energy at 20m 1/4 the Intensity that it was at 10m. That is:

$$I = \frac{1}{\gamma^2}$$

where r is the distance from the source.

The sensation of loudness is determined by the Intensity. The greater the Intensity the greater is the perceived loudness. It is usual to symbolise Intensity as I expressed in watt/m^2.

The energy flow associated with a sound wave is the total mechanical energy (potential and kinetic energies associated with elastic oscillations of the medium) that is transferred during each second through a surface of unit area ($1m^2$) - expressed in Joule/m^2/sec or Watt/m^2. Frames are obtained for each utterance of the speaker form Intensity.

3. SPEECH RECOGNITION SYSTEM

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 and control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding, a subject covered in section. Speech recognition performs a task similar with human brain. Start from phoneme, syllable, word and then the sentence which is an input for speech recognition systems. Many researcher that have been prove to decrease the error and also any disruption while doing the recognition.

Generally, speech recognition process contains three main stages for processing the speech which is acoustic processing, feature extraction and recognition, as shown in Fig.2. The acoustic processing obtains the sequence of input vector that will be used in next stages, feature extraction. For comparison purposes, Linear Prediction Coding and Mel frequency cepstral coefficients are performed.

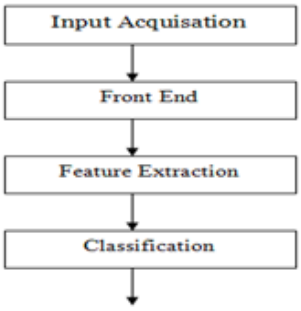

Fig 2: Process of Speech Recognition

3.1 Learning Method

Learning is necessary when the information about inputs/outputs is unknown or incomplete. Learning is the method of setting the appropriate weight values. There are two types of training namely supervised and unsupervised. Supervised learning requires the network to have an external teacher. The algorithm adjusts weights using input-output data to match the input-output characteristics of a network to the desired characteristics. In the learning without supervision, the desired response is not known and in supervised learning at each instant of time when the input is applied, the desired response of the system provided by the teacher is assumed. The distance between the actual and desired response serves as an error measure and is used to correct network parameters externally.

4. RECOGNITION METHODOLOGY

In multi-class mode such as the present case, each classifier tries to identify whether the set of input feature vectors, derived from the current signal, belongs to a specific class of numbers or not, and to which class exactly. For samples that can not be realized as a specific class a random class is selected.

5. CLASSIFIERS

Several classifiers are tested for mentioned dataset. The structures of successful classifiers in recognition are described in following subsections.

5.1. Multi-Layer Perceptron

This is perhaps the most popular network architecture in use today, due originally to Rumelhart and McClelland (1986). The units each performed a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output, and the units are arranged in a layered feedforward topology. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) the free parameters of the model. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. Important issues in MLP design include specification of the number of hidden layers and the number of units in these layers.

The number of input and output units is defined by the problem (there may be some uncertainty about precisely which inputs to use, a point to which we will return later. However, for the moment we will assume that the input variables are intuitively selected and are all meaningful). The number of hidden units to use is far from clear. As good a starting point as any is to use

one hidden layer, with the number of units equal to half the sum of the number of input and output units. Again, we will discuss how to choose a sensible number later.

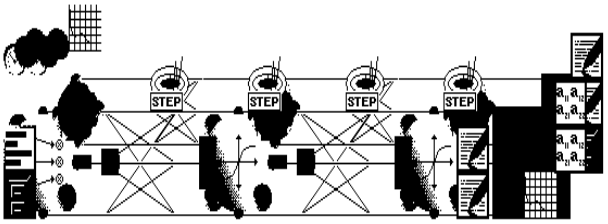

Fig 3: MLP Network architecture with step learning rule.

Table 1. Training Parameters for MLP Classifier

Max.No.of Epochs	100
Learning Rate	0.01
Initialization method	Random Gaussian

This network has an input layer (on the left) with three neurons, one hidden layer (in the middle) with three neurons and an output layer (on the right) with three neurons.

There is one neuron in the input layer for each predictor variable. In the case of categorical variables, N-1 neurons are used to represent the N categories of the variable.

Input Layer — A vector of predictor variable values (x1...xp) is presented to the input layer. The input layer (or processing before the input layer) standardizes these values so that the range of each variable is -1 to 1. The input layer distributes the values to each of the neurons in the hidden layer. In addition to the predictor variables, there is a constant input of 1.0, called the bias that is fed to each of the hidden layers; the bias is multiplied by a weight and added to the sum going into the neuron.

Hidden Layer — Arriving at a neuron in the hidden layer, the value from each input neuron is multiplied by a weight (wji), and the resulting weighted values are added together producing a combined value uj. The weighted sum (uj) is fed into a transfer function, σ, which outputs a value hj. The outputs from the hidden layer are distributed to the output layer.

Output Layer — Arriving at a neuron in the output layer, the value from each hidden layer neuron is multiplied by a weight (wkj), and the resulting weighted values are added together producing a combined value vj. The weighted sum (vj) is fed into a transfer function, σ, which outputs a value yk. The y values are the outputs of the network.

If a regression analysis is being performed with a continuous target variable, then there is a single neuron in the output layer, and it generates a single y value. For classification problems with categorical target variables, there are N neurons in the output layer producing N values, one for each of the N categories of the target variable.

5.2 Radial Basis Neural Networks

The core of a speech recognition system is the recognition engine. The one chosen in the paper is the Radial Basis Function Neural Network (RBF).

This is a static two neuron layers feed forward network with the first layer L1, called the hidden layer and the second layer, L2, called the output layer. L1 consists of kernel nodes that compute a localized and radially symmetric basis functions.

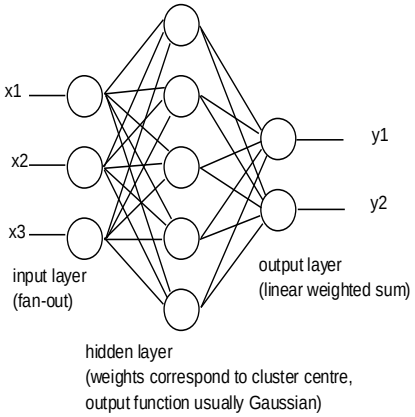

Fig. : 4 Radial Basis Function Neural Network

Table 2. Training Parameters for RBF Classifier

Radial Assignment	Sample Training Cases
Transfer Function	TanhAxon
Learning Rule	LevenbergMarquarade
Learning Rate Start	0.01
Learning Rate Decay	0.001
Threshold	0.01

The pattern recognition approach avoids explicit segmentation and labeling of speech. Instead, the recognizer used the patterns directly. It is based on comparing a given speech pattern with previously stored ones. The way speech patterns are formulated in the reference database affects the performance of the recognizer. In general, there are two common representations,

The output y of an input vector x to a (RBF) neural network with H nodes in the hidden layer is governed by:

$$y = \sum_{h=0}^{H-1} w_h \, \phi_h(x)$$

Where w_h are linear weights ϕ_h are the radial symmetric basis functions. Each one of these functions is characterized by its center c_h and by its spread or width σ_h. The range of each of these functions is [0, 1].

5.3 Modular Neural Networks

When modularity is applied for the creation of a modular neural network (MNN) based controller, three general steps are commonly observed. Those are task decomposition, training and multi-module decision-making (Auda and Kamel, 1999). Task decomposition is about dividing the required controller into several sub-controllers, and assigning each sub-controller to one neural module. Then the modules should be trained either in parallel or in different processes following a sequence indicated by the modular design. Finally, when the modules have been prepared, a multi-module decision making strategy is implemented which indicates how all those modules should interact in order to generate the global controller response. This modularization approach can be seen as a modularization at the level of the task.

A modular neural network system consists of several modules that can be arbitrarily connected to each other.

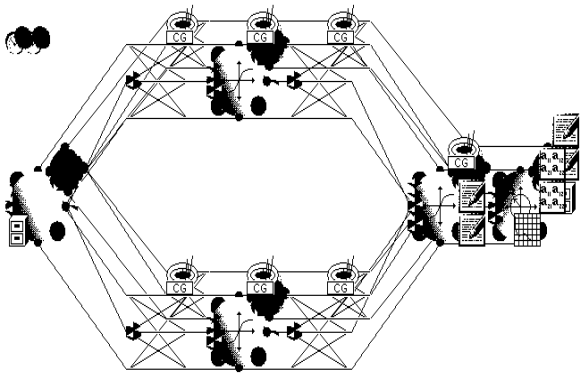

Fig 5: Architecture of Modular Neural Network Model.

Table 3. Training Parameters for MNN Classifier

Upper Transfer Function	TanhAxon
Upper Transfer Function	TanhAxon
Lower Transfer Function	LevenbergMarquarade
Learning Rule	100
Threshold	0.01

The modules are not limited to neural networks. They can be data filters, multiplexers, and so on. The dataflow between the modules is triggered by the data itself according to predefined logics. There are three main reasons why the modular neural network approach is more effective than the single neural network approach:

1) And, after dividing, it often becomes easier to select and train a neural network to solve a specific subtask. A modular system that consists of neural networks trained in this way is a hierarchical mixture of domain experts.
2) It is usually possible to separate independent or less correlated input data sources. The separated data input sources can be processed more efficiently by different neural networks that are trained specifically for them. Moreover, by doing that, much less data examples will be needed for training the whole system.
3) Compared to single neural network systems, it is usually easier to improve the desired behavior of modular systems by changing the architectures. For example, a modular system can be made insensitive to the momentary unavailability of certain data input sources, or the distribution of classification errors can be controlled within desired classes.

5.3 SOM Network Architecture

The hidden layer of an ANN is one of the most complex parts to design in an artificial neural network. This section proposes a Supervised SOM Based Architecture, which consists of Input layer, Competitive layer and Output layer.

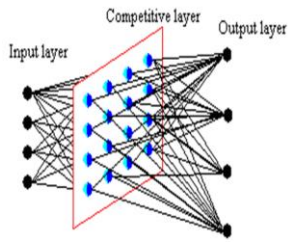

Fig 6. SOM Network Architecture

Table 4. Training Parameters for SOM Classifier

Max.no.of Epochs	100
Learning Rate Start	0.01
Learning Rate End	0.001
Transfer Function	TahnAxon
Learning Rule	LevenbergMarquarade
Neighbourhood Start	3
Neighbourhood End	1

Input Layer

Accepts multidimensional input pattern from the environment. An input pattern is represented by a vector. Each neurode in the input layer represents one dimension of the input pattern. An input neurode distributes its assigned element of the input vector to the competitive layer.

Competitive layer

Each neurode in the competitive layer receives a sum of weighted inputs from the input layer. Every neurode in the competitive layer is associated with a collection of other neurodes which make up its 'neighbourhood'. We can organize Competitive layer on any dimension. Upon receipt of a given input, some of the neurodes will be sufficiently excited to fire. This event can have either an inhibitory, or an excitatory effect on its neighborhood. The model has been copied from biological systems, and is known as 'on-center, off-surround' architecture, also known as lateral feedback / inhibition.

Output layer

Organization of the output layer is application-dependent. Strictly speaking, not necessary for proper functioning of a Kohonen network. The "output" of the network is the way we choose to view the interconnections between nodes in the competitive layer. If nodes are arranged along a single dimension, output can be seen as a continuum:
Self-organizing feature maps (SOFMs) transform the input of arbitrary dimension into a one or two dimensional discrete map subject to a topological (neighborhood preserving) constraint. The feature maps are computed using Kohonen unsupervised learning. The output of the SOFM can be used as input to a supervised classification neural network such as the MLP. This network's key advantage is the clustering produced by the SOFM which reduces the input space into representative features using a self-organizing process. Hence the underlying structure of the input space is kept, while the dimensionality of the space is reduced.

6. PERFORMANCE EVALUATION

There are 5 speakers and 4 features LPCC, MFCC, Pitch and Intensity. The recognition rate of speakers for the classifiers SOM, RBF, MNN, TLNN and MLP are estimated and presented in tables 5-9.

Table 5 . SOM Recognition Rate

	LPCC	MFCC	PITCH	INTENSITY
Speaker number	Recognitio n Rate(%)	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognition Rate (%)
Speaker1	88.05	89.58	88.56	97.50
Speaker2	84.90	89.19	82.70	97.40
Speaker3	86.04	89.27	98.22	98.54
Speaker4	96.18	89.56	95.34	98.72
Speaker5	93.92	86.58	80.23	98.69

Table 6. Multi Layer Perceptron

	LPCC	MFCC	PITCH	INTENSITY
Speaker number	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognition Rate (%)
Speaker1	84	80	82	90
Speaker2	80	85	80	94
Speaker3	85.21	87.00	94	96.21
Speaker4	95.21	88.52	94.21	96
Speaker5	90	84.20	81021	95.21

Table 7. Radial Basis Function Recognition Rate

	LPCC	MFCC	PITCH	INTENSITY
Speaker number	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognition Rate (%)
Speaker1	87.04	85.20	88.12	95.20
Speaker2	82.10	87.60	82.00	96.00
Speaker3	85	88	96	98.26
Speaker4	95.36	89	95.34	98.11
Speaker5	92	86.21	80.12	96.00

Table 8 . Modular Neural Network

	LPCC	MFCC	PITCH	INTENSITY
Speaker number	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognition Rate (%)
Speaker1	86	81	85.21	92
Speaker2	85	86	82	95
Speaker3	85	88	95	97.12
Speaker4	96	90	96.21	97.5
Speaker5	91	85.20	82.68	96.8

Table 9. Time Lagged Neural Network

	LPCC	MFCC	PITCH	INTENSITY
Speaker number	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognitio n Rate (%)	Recognition Rate (%)
Speaker1	85	85.21	86	94
Speaker2	80	86.21	81	94
Speaker3	85	86.21	95	97.26
Speaker4	94	87	95	97.12
Speaker5	91	85	82	95.80

7. NEAREST NEIGHBOUR CLASSIFICATION

7.1 Nearest neighbour classification with Euclidean distance

To understand nearest neighbour classification using numeric attribute values, let us suppose the following.

- $A_1, A_2, ..., A_n$ are the $n \geq 1$ numeric attributes of the patterns in a training set.
- $A_1^{'}, A_2^{'}, ..., A_n^{'}$ are the values of the attributes for some training pattern, such that the values of $A_i \ is \ A_i^{'}, \ for \ 1 \leq i \leq n.$
- $A_1^{''}, A_2^{''}, ..., A_n^{''}$ are the values of the attributes for some all recall pattern, such that the value of $A_i \ is \ A_i^{''}, \ for \ 1 \leq i \leq n.$

The above training pattern can be represented as a point in the n-dimensional $A_1 - A_2 ... - A_n$ coordinate space, such that the coordinate of $A_i \ is \ A_i^{'}, \ for \ 1 \leq i \leq n.$ Similarly, the above recall pattern can be represented as a point in this space, with the coordinate of $A_i \ being \ A_i^{''}, \ for \ 1 \leq i \leq n.$ Then, the Euclidean distance between the training pattern and the recall pattern is defined to be

$$\sqrt{\sum_{i=1}^{n} (A_i^{''} - A_i^{'})^2} \qquad (1)$$

Each training pattern is viewed to be a neighbour of the recall pattern. The smaller the Euclidean distance between the recall pattern and a neighbour, the near the recall pattern is to that neighbour. To classify a given recall pattern by the nearest neighbour classifier, do the following.

1. Calculate the Euclidean distance between the recall pattern and each of its neighbours.
2. Assign the recall pattern to the class of its nearest neighbour.

If the nearest neighbour happens to be an outlier, then the recall pattern is likely to be misclassified. A modification to the

procedure is to assign the recall pattern to the class most frequent among its $k \geq 1$ nearest neighbours, where the value of k is a design choice that one can make. In practice, one may try out different values of k and select the value that gives the best results. To obtain a value of k heuristically, often found useful in practice, start trying with an integer approximation of the square root of the number of training patterns in the class that has the fewest patterns.

The SOM based speech recognition system is compared with other classifiers by using Euclidean distance classification and the results are placed in table 10.

Table 10. Nearest Neighbour Classifier with Euclidean Distance

SOM Vs. Other classifiers	MLP Based speaker 1	RBF Based speaker 1	MNN Based speakers 1	TDNN Based speaker 1
SOM Based speaker 1	10.9	5.068	10.98	6.86
SOM Vs. Other classifiers	MLP Based speaker 2	RBF Based speaker 2	MNN Based speakers 2	TDNN Based speaker 2
SOM Based speaker 2	8.44	9.198	8.45	11.71
SOM Vs. Other classifiers	MLP Based speaker 3	RBF Based speaker 3	MNN Based speakers 3	TDNN Based speaker 3
SOM Based speaker 3	7.31	8.22	7.30	7.59
SOM Vs. Other classifiers	MLP Based speaker 4	RBF Based speaker 4	MNN Based speakers 4	TDNN Based speaker 4
SOM Based speaker 4	10.90	10.00	11.04	8.91
SOM Vs. Other classifiers	MLP Based speaker 5	RBF Based speaker 5	MNN Based speakers 5	TDNN Based speaker 5
SOM Based speaker 5	8.55	10.02	7.93	8.69

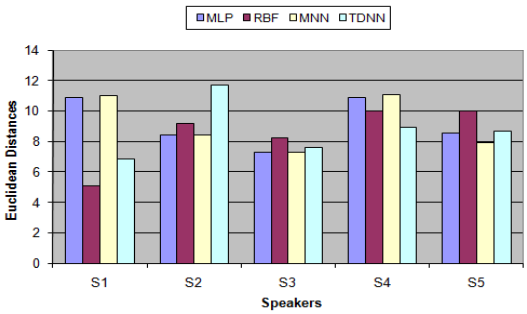

Fig 7. Bar graphs obtained for Nearest Neighbour Classification with Euclidean Distance

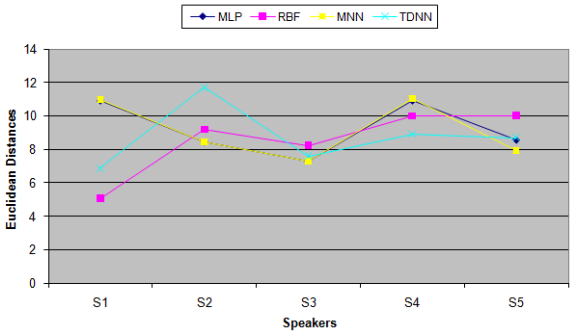

Fig 8. Line graphs obtained for Nearest Neighbour Classification with Euclidean Distance

7.2 Nearest neighbour classification with weighted Euclidean distance

The weighted Euclidean distance can be obtained from the equation.

$$\sqrt{\sum_{i=1}^{n} w_i (A_i'' - A_i')^2} \qquad (2)$$

where $w_i = \dfrac{h_i^{-2}}{\sum_{j=1}^{n} h_j^{-2}}$ and $h_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$

is the distance between scatter point to the interpolation point.

The SOM based speech recognition system is compared with other classifiers by using Weighted Euclidean distance classification and the results are placed in table 11.

Table 11. Nearest Neighbour Classifier with Weighted Euclidean Distances

SOM Vs. Other classifiers	MLP Based speaker 1	RBF Based speaker 1	MNN Based speakers 1	TDNN Based speaker 1
SOM Based speaker 1	2.02	3.96	2.22	1.79
SOM Vs. Other classifiers	MLP Based speaker 2	RBF Based speaker 2	MNN Based speakers 2	TDNN Based speaker 2
SOM Based speaker 2	2.60	0.64	1.72	2.68
SOM Vs. Other classifiers	MLP Based speaker 3	RBF Based speaker 3	MNN Based speakers 3	TDNN Based speaker 3
SOM Based speaker 3	1.02	1.26	0.90	0.80
SOM Vs. Other classifiers	MLP Based speaker 4	RBF Based speaker 4	MNN Based speakers 4	TDNN Based speaker 4
SOM Based speaker 4	1.98	0.95	3.62	3.21
SOM Vs. Other classifiers	MLP Based speaker 5	RBF Based speaker 5	MNN Based speakers 5	TDNN Based speaker 5
SOM Based speaker 5	1.27	2.14	2.24	2.68

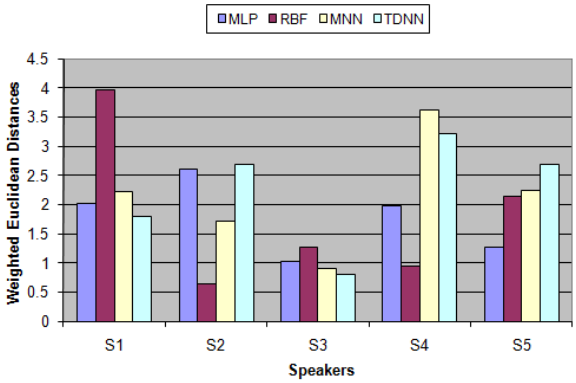

Fig 9. Line grphs obtained for Nearest Neighbour Classification with Weighted Euclidean Distance

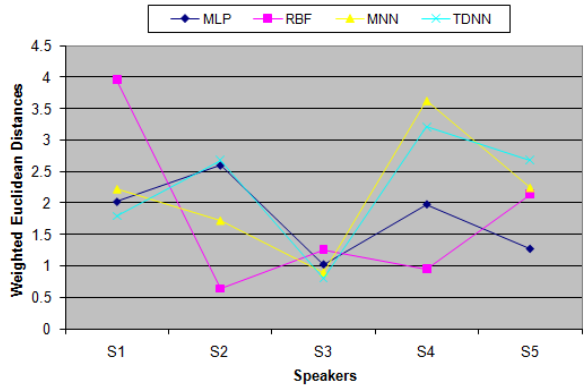

Fig 10. Line grphs obtained for Nearest Neighbour Classification with Weighted Euclidean Distance

8. CONCLUSION

The features extracted in this paper are LPCC, MFCC, Pitch and Intensity. The classifiers used are Self Organized Maps, Radial Basis Function, Modular Neural network, MultiLayer Perceptron, Time Lagged Neural network respectively. The speakers recognition rate is maximum for Self Organized Map. Both Euclidean distance and Euclidean weighted distances are used to find the nearest neighbour classification of speakers .

In the Euclidean distance classification phase for SOM speech recognition system, it is found that for speakers (1-5) the nearest neighbour Classifiers are Radial Basis Function, Multi Layer Perceptron, Modular Neural Network Multi Layer Perceptron and Time Lagged Neural Network respectively. Where as in the Euclidean weighted distance phase for SOM based speech recognition system, it is found that speakers 1-5 the nearest neighbours are Time Lagged Neural network, Multi Layer perceptron Neural network, Modular Neural Network respectively. As the distance decreases, accuracy increases. Hence the Weighted Euclidean distance classification is the best classification to classify the speakers. The outlayers of SOM based system, for the speakers 1 – 5 are respectively MNN, TDNN, RBF, MNN and RBF classifiers.

9. REFERENCES

1. T. Deselaers, G. Heigold, and H. Ney. Speech Recognition with State-based Nearest Neighbour Classifiers. In Interspeech, pages 2093-2096, Antwerp, Belgium, August 2007.
2. CHEN, O. ABDULLA, W.H., SALCIC, Z. 'Performance Evaluation of Different Front-End Processing for Speech Recognition Systems', Technical Report, SoE-621, 2005, pp1-36
3. C.J.Wellekens,"**Introduction to Speech Recognition Using Neural Networks**", Proceedings of ESANN-1998, pp.227-236, Bruges
4. Speech recognition Applications by Steve Green Dec 05/Jan 06.
5. Jihene El Malik, (1998). "Kohonen Clustering Networks For Use In Arabic Word Recognition System." Sciences Faculty of Monastir, Route de Kairouan, 14-16 December.
6. C.M.Bishop, Neural Networks for pattern recognition, oxford university press, 1995.
7. K. Haese, "Self-organizing feature maps with self-adjusting learning parameters", IEEE Transactions on Neural Networks, vol. 9, pp. 1270–1278, 1998.
8. Ben Gold and Nelson Morgan (2007) Speech and Audio Signal Processing, Wiley India Edition, New Delhi.
9. Yegnanarayana, B, (2006), Artificial neural networks Prentice-Hall of India, New Delhi.