ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, Vol 8(19), DOI: 10.17485/ijst/2015/v8i19/76217, August 2015 ISSN (Online) : 0974-5645

Performance Analysis of SOFM based Reduced Complexity Feature Extraction Methods with back Propagation Neural Network for Multilingual Digit Recognition

John Sahaya Rani Alex*, **Ajinkya Sunil Mukhedkar and Nithya Venkatesan**

School of Electronics Engineering Department, VIT University, Chennai - 600 127, Tamil Nadu, India,
jsranialex@vit.ac.in, ajinkya.sunil2013@vit.ac.in, nithya.v@vit.ac.in

Abstract

Background: Speech recognition is an active area of research, used to transliterate words vocalized by individuals in order to make them machine recognizable. For real time speech recognition applications the response time, size of training data and recognition accuracy are the important aspects. **Methods:** A Hybrid speech recognition system is proposed on the basis on Artificial Neural Network (ANN) in this research. The Self Organising Feature Map (SOFM) is used to reduce the feature vector dimensions which are extracted using the Mel-Frequency Cepstrum Coefficients (MFCC), Perceptual Linear Predictive (PLP) and Discrete Wavelet Transform (DWT) methods. The Back Propagation Network (BPN) algorithm is used for training the Artificial Neural Network for pattern classification. **Findings:** The proposed method is tested with TIDIGITS data. Results indicate that despite of the large reduction in the feature vector dimensions the recognition accuracy obtained using SOFM technique is same as that of the recognition accuracy of the conventional methods. The response time is also fast and the data size of the input data is reduced considerably. The proposed hybrid system is further tested using multilingual isolated digit data.

Keywords: Artificial Neural Network, Discrete Wavelet Transform, Feature Extraction, Mel Frequency Cepstrum Coefficients, Perceptual Linear Predictive, Self-organising Feature Map, Speech Recognition

1. Introduction

A speech-recognition system is categorized either as an isolated speech or continuous speech. The isolated word recognition involves a spoken word which is preceded and succeeded by silence, whereas continuous speech recognition does not. The main challenge involved in designing of any speech recognition system is that of modelling the dissimilarities of the identical word as vocalized by diverse accent, gender, regional dialects, voice pattern, etc. In the near future speech driven interfaces are going to replace nowadays conventional interfaces thus we have to address some very important issues related to that. Most of the speech driven interfaces that are available nowadays expect the user to speak in a particular global language. Needless to say, they are monolingual. But due to globalization as the spread of application increases steadily, quite often we have to come across the type of users whose utterance contains words from multiple languages. A good example might be a user who has a weak English background. This problem is very common in many countries where either English is not that widely spoken example China or many people are less English educated for example India. It is always going to happen that he might mix up his own native language when tries to use the interface in English. This brings the need for automatic speech recognition system which has the multilingual functions.

Researchers throughout the globe are focussed to get better performance of speech recognition system, reduce the feature vector dimensions and provide noise robust

*Author for correspondence

features[1]. Speech is a non-stationary signal thus requiring automated analysis. Therefore, need for machine learning algorithms and techniques in this area of research are stronger than ever. One such model is the Self-Organized Feature Map (SOFM) technique which has the capability to cluster the data and reduce its dimensions. For recognition purpose various models are implemented such as Hidden Markov Models (HMM), Neural Networks (NNs), Support Vector Machine (SVM) that are ideally suited for speech recognition domain which is characterized by large amount of data and variable pattern sequences.

The paper is prearranged according to sections as follows: Section 2 gives a brief idea about the feature extraction techniques. Section 3 gives a brief account of SOFM method. The following Section 4 gives an insight about the Artificial Neural Networks (ANN's). Experimental setup and methodology is given in Section 5. The results are displayed in Section 6. To end, the conclusion and the references are presented.

2. Feature Extraction Methods

It is process in which the input data is altered into the machine recognizable set of features. The most important task here is to choose carefully the extracted features using the methods mentioned such that the features set will remove the related information from the input speech data, such that the desired task is performed via this compact depiction instead of the full size input speech data. Literature illustrates various techniques for feature extraction[2] such as Mel Scale frequency Cepstrum Coefficient (MFCC), Perceptual Linear Predictive (PLP), Linear Predictive Coding (LPC), RASTA- PLP and Discrete Wavelet Transform (DWT).

2.1 Mel Frequency Cepstral Coefficients

The mel-cepstrum, presented by Davies and Mermelstein, exploits auditory values, as well as the decorrelating property of the Cepstrum[3]. MFCCs the most prosperous feature extraction method in speech recognition tasks is considered for this research. The MFCC parameterization is realized by the bank of symmetric overlapping triangular filters spaced linearly in a Mel-frequency axis, according to auditory perceptual considerations[4]. Figure 1 shows the block diagram of extraction process of speech parameters using MFCC technique.

The pre-emphasis task is performed to equalise the higher frequency constituent of the speech signal.

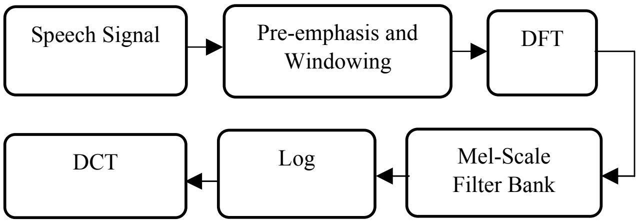


Figure 1. Block diagram of extraction process of speech parameters using MFCC technique.

$$H_{preem}(z)=1-a_{preem}z^{-1} \tag{1}$$

The following process of frame blocking; the input speech signal is divided in form of frames measuring 25 ms with an overlay of 10 ms. Then each frame is then multiplied by a Hamming window for smoothening transitions at the initial and end points.

$$w(n)=0.54-0.46*cos\left(\frac{2\pi(n-1)}{N-1}\right) \tag{2}$$

To obtain the magnitude spectrum for each frame DFT is applied. A filter bank, with triangular filters placed according to an empirical log-linear scale referred to as the mel-scale, is built to correspond the human sensory system with additional channels focused at low frequencies and fewer channels at high frequencies.

$$f_{cm(Mel)}=f_{L(Mel)}+m*\left(\frac{f_{H(Mel)}-f_{L(Mel)}}{M+1}\right)1\le m\le M \tag{3}$$

The natural logarithmic function compresses the Mel filter bank output modelling the apparent loudness at a given signal strength.

$$X_{m(\ln)}=\ln\left(X_m\right)1\le m\le M \tag{4}$$

Consider p as the order of the Mel scale Cepstrum. By allowing the initial p DCT coefficients, the feature vector is acquired. From (5) given below, k^{th} MFCC coefficient c_k is stated.

$$c_k=\sqrt{\frac{2}{M}}\sum_{m=1}^{M}X_{m(\ln)}*\cos\left(\frac{\pi(m-0.5)}{M}\right) \tag{5}$$

The static feature vector obtained from (5) is appended by a log energy component. If DCT give a set of twelve coefficient one log energy coefficient is appended to give a set of thirteen vectors representing each frame.

The dynamic features obtained from static one by taking the first order time derivatives are delta coefficients and taking the second order time derivatives are known as acceleration coefficients.

The delta features are calculated as follows,

$$\Delta \mathrm{c}(\mathrm{n}) = \frac{1}{\sum_{k=1}^{D} k^2} \sum_{k=0}^{D} k * \left[c(n+k) - c(n-k) \right] \qquad (6)$$

2.2 Perceptual Linear Predictive

Perceptual Linear Predictive (PLP) is in short the autoregressive all-pole model of the short term power spectrum of speech. Bark spaced filter bank is implemented. The block diagram of extraction process of speech parameters using PLP technique is shown in Figure 4.

The segments of the speech signals are weighted using the Hamming window equation

$$W(n) = 0.54q - 0.46\cos\left[2rn/(N-1)\right] \qquad (7)$$

N denotes the interval of the hamming window. The discrete time domain signal obtained from the above process of windowing is subjected to N-point DFT. DFT is applied for the conversion of windowed speech divisions in frequency domain. The frequency domain signal is then passed through Bark spaced auditory filter bank. The equal loudness and pre-emphasis block is associated with the degree estimation to the imbalanced sensitivity of human auditory system at entirely dissimilar frequencies and mimics the sensitivity of auditory perception. The intensity and loudness compression operation is estimation to the power law of auditory perception and mimics the nonlinear relation between the strength of sound and its apparent loudness. Then IDFT is applied to get the autocorrelation functions. The obtained autocorrelation values are implied for the calculation of the Yule-Walker equations in order to find the autoregressive coefficients for the all-pole model. The obtained autoregressive coefficients are further transformed into some other set of vectors such as cepstral coefficients of the all-pole model. Further the dynamic coefficients using (6) are obtained.

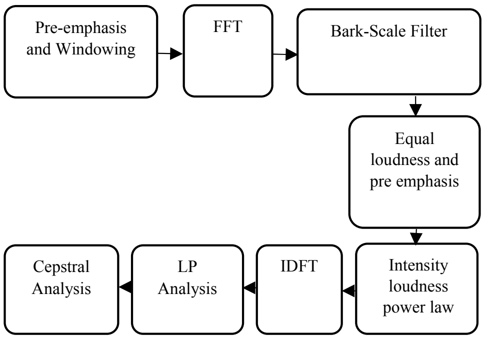

Pre-emphasis and Windowing
FFT
Bark-Scale Filter
Equal loudness and pre emphasis
Intensity loudness power law
IDFT
LP Analysis
Cepstral Analysis

Figure 2. Block diagram of extraction process of speech parameters using PLP technique.

2.3 Discrete Wavelet Transform

This transform is a new carefully worked-out tool for local depiction of non-stationary signals. Wavelets have energy focuses in time and are useful for the analysis of short-lived signals such as speech signals. DWT is the most favourable mathematical transformation which provides both the time frequency info of the signal and is computed by sequential low pass filtering and high pass filtering to construct a multi resolution time-frequency plane. In DWT a discrete signal x[k] is filtered by using a high pass filter and a low pass filter, which will separate the signals to high frequency and low frequency components. To reduce the number of samples in the subsequent output we apply a down sampling factor of ↓2. The Discrete Wavelet Transform is defined by the following Equation.

$$W(j,k) = \sum_j \sum_k \mathrm{X}(\mathrm{k}) * 2^{-\frac{j}{2}} \varphi\left(-2^{-j} n - k\right) \qquad (8)$$

Where Ψ (t) is the basic analysing function called the mother wavelet. The high frequency analysis is done using narrow windows and low frequency analysis is done using wide window. When we use a wide window, time resolution is poor, but frequency resolution is good and low frequencies are resolved in frequency domain. When narrow window is used, the time resolution is better and the high frequencies are resolved in time domain. The digital filtering technique can be expressed by the following Equations

$$\mathrm{Y}_{\mathrm{high}}[\mathrm{k}] = \sum_{\mathrm{n}} \mathrm{X}[\mathrm{n}]\mathrm{g}[2\mathrm{k}-1] \qquad (9)$$

$$\mathrm{Y}_{\mathrm{low}}[\mathrm{k}] = \sum_{\mathrm{n}} \mathrm{X}[\mathrm{n}]\mathrm{h}[2\mathrm{k}-1] \qquad (10)$$

Where $\mathrm{Y}_{\mathrm{high}}$ and $\mathrm{Y}_{\mathrm{low}}$ are the outputs of the high pass and low pass filters. Wavelet packet decomposition is used as shown in Figure 3 in this research. The wavelet packet method is a generalization of wavelet decomposition that offers a richer signal analysis. A six level packet decomposition with dB 32 i.e. 64 set of vectors will represent each utterance is implemented. The entropy type used here is log energy given by Equation below.

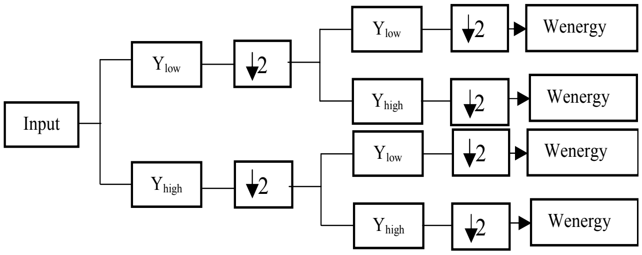


Figure 3. Two level wavelet packet decomposition.

$$Fi=\frac{\sum_k S_k^2}{\sum_i\sum_k S_k^2}\quad For\ I=1\ to\ N \tag{11}$$

3. Overview and Implementation of Self-Organising Feature Map

Self-Organizing Feature Maps (SOFM) popularly known Kohonen feature maps are neural networks which are used for the purpose of clustering. The main objective of the clustering task is to shrink the dimensions of input data by classifying or grouping analogous data items together in clusters or groups[6].

The training mode of SOFM is based on competitive learning. SOFM operates in dual mode namely the training mode and the clustering or grouping mode. The training mode operation includes the network discovering an output node in a way such that the Euclidean distance among the present input vector and the weight set linking the input elements to the discovered output element is as least as possible. This node is declared as the winner node. The weights of the winner and the corresponding weights of the neighbouring output elements to that of the winner are reorganized so that the updated weight set is nearby to the present input vector. The consequence of update for each element is comparative to a neighbourhood function, which in turn rely on the neurons distance to the winner neuron. The procedure is implemented for a repeated number of times for all input patterns in order for the weights to be stable. The parameters such as the neighbourhood function, the learning rate (η) and the termination conditions dependents upon the application. The following mode is clustering which is simple once the primary training mode is completed effectively. In this mode, after applying the input vector, only the winner unit is computed.

3.1 SOFM Training Algorithm

- The first task is assignment of small random values ranging from -1 to 1 to the weights.
- A vector u^k is selected from the whole training dataset which is supposed to be the input for present step.
- Finding the winning output node denoted by n_{win} is the next step which is done by

$$n_{win}=argmin\,\|u-w^j\| \tag{12}$$

In (12) symbol $\|.\|$ denotes the Euclidean norm and w_j is the weight vector associated to the related input nodes to the output node j.

- The modification of the weight vectors is done rendering to the formula:

$$w_{ij}(t+1)=w_{ij}(t)+\eta(t)(u_i-w_{ij})*N(j,t) \tag{13}$$

In (13) w_{ij} is the ith element of the weight vector w_j, symbol η(t) denotes the learning rate and N(j,t) denotes neighbourhood function.

- Steps 2 through 4 are performed again and again until convergence in the weights occurs.

4. Overview and Implementation of Back Propagation Network

An Artificial Neural Network (ANN) is a data processing model or scheme that tries to put on the adaptive biological learning capabilities of the human brain[7]. The design of a neural network is composed of a huge number of extremely interrelated processing elements commonly called as neurons or nodes.

The shaping of the neural networks appears to be layers. These layers construct a finite number of interrelated nodes comprising of an activation function. Patterns are feed to the system via the input layer, which transfers it to *n* number hidden layers where the concrete processing is completed via weighted connections. The hidden layers are then connected to an 'output layer' as shown in Figure 4. This output layer gives the answer.

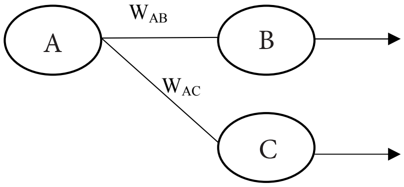


Figure 4. A mono connection learning in a BPN.

The information processing in ANN is thru parallel arrangement of input data with a huge number of processing neurons. During the learning phase the system factors fluctuate over time. Multi-Layer Perceptron (MLP) architecture is used for classification of patterns[8]. A pattern is transferred in the single direction i.e. from the input to the output and hence this architecture is popularly known as feed forward. The MLP networks are learned with using the Backward Propagation (BP) algorithm which is broadly used in machine learning implementations[9].

The connection shown in Figure 4 among neuron, A represents the hidden layer neuron and neuron B represents the output layer neuron and takes the weight denoted by W_{AB}. The Figure 4 shows alternative linking, among neuron A and C. The algorithm works in step wise manner as follows:

- At first the inputs are applied to the network in column wise fashion and the output is computed but this initial output is not reliable since the primary weights are random.
- The following step is computation of neuron B error. It is computed using the following formula

$$\mathrm{Error}_B = \mathrm{Output}_B\left(1-\mathrm{Output}_B\right)\left(\mathrm{Target}_B-\mathrm{Output}_B\right) \quad (14)$$

  The term “Output (1-Output)” is essential in since the activation function is Sigmoid.
- Changing weights is done as follows. Consider W^{+}_{AB} is the modified weight and W_{AB} be the primary weight.

$$W^{+}{}_{AB} = W_{AB} + \left(Error_B * Output_A\right) \quad (15)$$

  The output of the linking neuron A is taken into consideration and not neuron B. Modification of every weights in the output layer is completed in this fashion.
- The next step of calculating the errors related to the hidden layer of neurons. For the calculation of these error no direct method is available since the target file is not there, hence the method of Back Propagation comes into picture in which the value is back propagated from the output layer. Errors from the output neurons are back propagated through the weights in order to find the errors of hidden layer. In Figure 4 neuron A is linked to B and C, propagating the errors from B and C for generation of an error for A.

$$Error_A = Output_A\left(1-Output_A\right)\left(Error_B W_{AB} + Error_C W_{AC}\right) \quad (16)$$

- After obtaining the error for the hidden layer neurons follow step 3 to alter the weights of the hidden layer. Repeating this process a network with n number of layers can be trained.

5. Simulation Setup

This section evaluates the performance of the proposed technique on isolated word recognition system.

5.1 TIDIGIT Standard Database

The isolated word recognition experiments were conducted using TI-Digits database[10] in clean environments. There are 455 samples each of ten English digits ‘zero’ through ‘nine’ which are used for training purpose. Different samples of low pitch male voice and high pitch female voice are taken.

5.2 Multilingual Database

For the multilingual application Hindi, Marathi and Tamil languages digits from ‘zero’ to ‘nine’ were considered. A database of 22 speakers for each language was created for training and testing purpose. The recording was done in a silent room using WavSurfer[11] tool.

5.3 Methodology

At first the speech database is arranged in a hierarchical manner according to the different speakers, man and women. The proposed SOFM for back propagation neural network based multilingual digit recognition is shown in Figure 5. The speech database considered in this approach is the isolated TIDIGITS English language database. Feature extraction is done using the different methods explained in Section 2. SOFM training algorithm explained in Section 3 is implemented using MATLAB. The input to the SOFM program is a column wise arranged speech feature vectors. The Target file is created

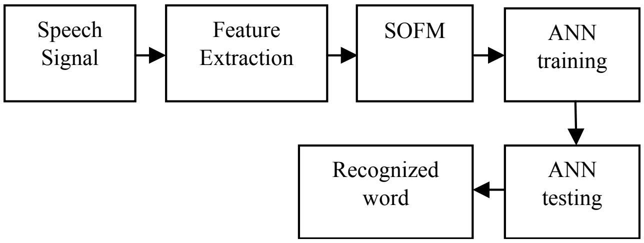


Figure 5. Proposed SOFM for back propagation neural network based multilingual digit recognition.

according to the input pattern. Input parameter for ANN was selected according to the criteria. The dimensions to be reduced were specified. The ANN training and testing is implemented in MATLAB. Among the whole input database 75 percent of the data was given to training, 15 percent to Testing and 15 percent to validation.

For example, MFCC_E method of feature extraction is implemented with 13 vectors each frame and there are average 80 frames for one utterance. Thus there are 1040 vectors for one utterance. If arranged column wise there are 1040 vectors in one column which represent a pattern for one speech utterance. Such 4558 utterances for digits 'zero' to 'nine' are considered for this example. Thus input database of dimension 1040 X 4558 is created. Target file is created in which a fixed pattern represents a digit, such as target pattern '1000000000' represents the digit 'zero' while target pattern '0100000000' represents digit 'one' and so on. This input file is feed to SOFM method of reducing the feature vector dimension. For instance the dimension is reduced to 100 X 4558 through SOFM technique. Further this reduced dimension database serves as input to ANN. These parameters are feed to the program and the confusion matrix is obtained which gives the overall accuracy of the system. The evaluation time is also recorded. The system configurations are Intel Core i5-2450M CPU @ 2.4 GHz, 4.0 GB RAM.

6. Result and Analysis

Table 1 gives the recognition accuracy of different feature extraction methods for TIDIGIT Standard database using back propagation neural network. While Table 2 shows the recognition accuracy of different feature extraction methods for TIDIGIT standard database using proposed SOFM for back propagation neural network method. From these tables we observe that despite of the reduction in dimensions of feature vectors there is no considerate amount of degradation in recognition accuracy. The evaluation time is reduced approximately four times after reducing the dimensions. Besides the size of the speech database is reduced twenty times.

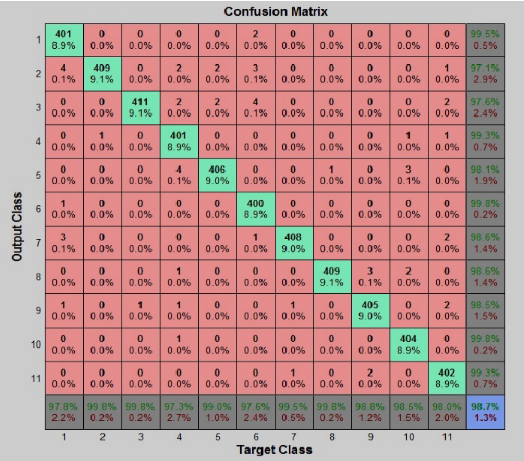
Table 1. Recognition accuracy of different feature extraction methods for TIDIGIT Standard database using back propagation neural network

Method	Dimension	Accuracy in percentage	Evaluation Time in seconds
MFCC_E	1000 X 4558	98.7 %	67.099070
MFCC_E_D	1500 X 4558	98.2 %	70.126348
MFCC_E_D_A	2000 X 4558	98.0 %	74.226358
PLP_D_A	2000 X 4558	93.2 %	80.493666
DWT	64 X 4558	91.8 %	26.210033

Table 2. Recognition accuracy of different feature extraction methods for TIDIGIT Standard database using proposed SOFM for back propagation neural network method

Method	Dimension	Accuracy in percentage	Evaluation Time in seconds
MFCC_E	100 X 4558	98.2 %	18.961330
MFCC_E_D	100 X 4558	98.0 %	19.661030
MFCC_E_D_A	100 X 4558	97.8 %	20.190133
PLP_D_A	100 X 4558	92.7 %	21.641950
DWT	16 X 4558	84.5 %	12.890122

Figure 6 shows the confusion matrix for the MFCC_E method. The research area of machine learning algorithms, a confusion matrix popularly known as a possibility table or in some areas an error matrix, is a detailed table outline that allows imagining of the performance of an algorithm. Figure 7 is of the performance graph which plots the training, validation, and test performances given the training data. The confusion matrix after using proposed hybrid method is shown in Figure 8 while its performance plots in Figure 9. Likewise confusion matrix and performance plot were plotted for other methods included in the tables.



Confusion Matrix

Output Class \ Target Class	1	2	3	4	5	6	7	8	9	10	11	
1	401 8.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	99.5% 0.5%
2	4 0.1%	409 9.1%	0 0.0%	2 0.0%	2 0.0%	3 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	97.1% 2.9%
3	0 0.0%	0 0.0%	411 9.1%	2 0.0%	2 0.0%	4 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	97.6% 2.4%
4	0 0.0%	1 0.0%	0 0.0%	401 8.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	1 0.0%	99.3% 0.7%
5	0 0.0%	0 0.0%	0 0.0%	4 0.1%	406 9.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	3 0.1%	0 0.0%	98.1% 1.9%
6	1 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	400 8.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	99.8% 0.2%
7	3 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	408 9.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	98.6% 1.4%
8	0 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	409 9.1%	3 0.1%	2 0.0%	0 0.0%	98.6% 1.4%
9	1 0.0%	0 0.0%	1 0.0%	1 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	405 9.0%	0 0.0%	2 0.0%	98.5% 1.5%
10	0 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	404 8.9%	0 0.0%	99.8% 0.2%
11	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	2 0.0%	0 0.0%	402 8.9%	99.3% 0.7%
	97.8% 2.2%	99.8% 0.2%	99.8% 0.2%	97.3% 2.7%	99.0% 1.0%	97.6% 2.4%	99.5% 0.5%	99.8% 0.2%	98.8% 1.2%	98.5% 1.5%	98.0% 2.0%	98.7% 1.3%

Figure 6. Confusion matrix showing the recognition accuracy of MFCC_E method of feature extraction with dimension 1000X4558.

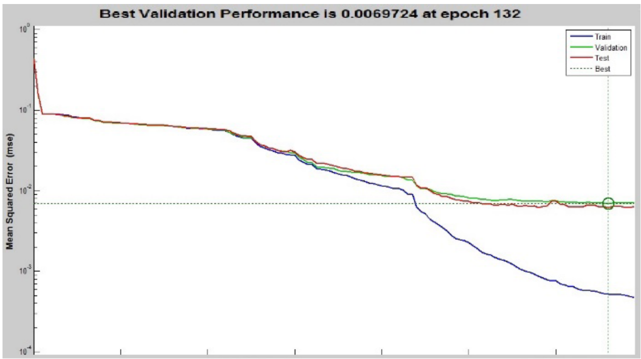
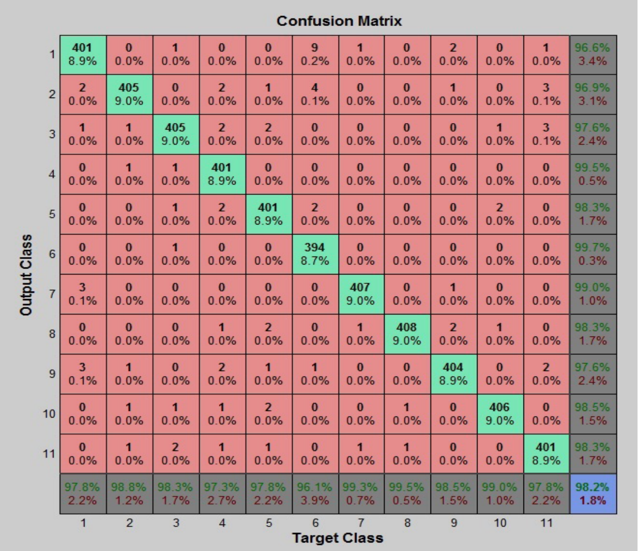


Figure 7. Performance graph of MFCC_E method feature extraction with dimension 1000X4558.



	1	2	3	4	5	6	7	8	9	10	11	
1	401 8.9%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	9 0.2%	1 0.0%	0 0.0%	2 0.0%	0 0.0%	1 0.0%	96.6% 3.4%
2	2 0.0%	405 9.0%	0 0.0%	2 0.0%	1 0.0%	4 0.1%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	3 0.1%	96.9% 3.1%
3	1 0.0%	1 0.0%	405 9.0%	2 0.0%	2 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	3 0.1%	97.6% 2.4%
4	0 0.0%	1 0.0%	1 0.0%	401 8.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	99.5% 0.5%
5	0 0.0%	0 0.0%	1 0.0%	2 0.0%	401 8.9%	2 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	98.3% 1.7%
6	0 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	394 8.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	99.7% 0.3%
7	3 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	407 9.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	99.0% 1.0%
8	0 0.0%	0 0.0%	0 0.0%	1 0.0%	2 0.0%	0 0.0%	1 0.0%	408 9.0%	2 0.0%	1 0.0%	0 0.0%	98.3% 1.7%
9	3 0.1%	1 0.0%	0 0.0%	2 0.0%	1 0.0%	1 0.0%	0 0.0%	0 0.0%	404 8.9%	0 0.0%	2 0.0%	97.6% 2.4%
10	0 0.0%	1 0.0%	1 0.0%	1 0.0%	2 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	406 9.0%	0 0.0%	98.5% 1.5%
11	0 0.0%	1 0.0%	2 0.0%	1 0.0%	1 0.0%	0 0.0%	1 0.0%	1 0.0%	0 0.0%	0 0.0%	401 8.9%	98.3% 1.7%
	97.8% 2.2%	98.8% 1.2%	98.3% 1.7%	97.3% 2.7%	97.8% 2.2%	96.1% 3.9%	99.3% 0.7%	99.5% 0.5%	98.5% 1.5%	99.0% 1.0%	97.8% 2.2%	98.2% 1.8%

Figure 8. Confusion matrix showing the recognition accuracy of dimension reduced by SOFM method for MFCC_E method of feature extraction with dimension 100X4558.

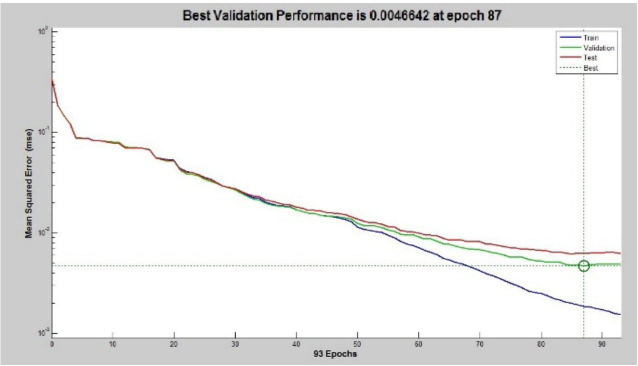


Figure 9. Performance graph of dimension reduced by SOFM method for MFCC_E method of feature extraction with dimension 100X4558.

Table 3. Recognition accuracy of different feature extraction methods for proposed SOFM for back propagation neural network based multilingual digit recognition

Method	Dimension	Accuracy in percentage
PLP	100 X 880	80.1 %
DWT	16 X 880	66.2 %

Table 3 shows two method of feature extraction namely DWT and PLP considered for multilingual application. The system performed effectively faster using DWT method of feature extraction (from Table 2). Thus it was considered for multilingual application. While PLP method of feature extraction has proved efficient by previous experiments and hence considered for this application. The recognition accuracy of PLP method is greater than DWT method.

7. Conclusion

The SOFM method of reducing the feature vector dimension proved efficient without degrading the recognition accuracy of the system. There is considerable reduction in the evaluation time of the system due to reduced dimensions. Thus the response time of the system is quicker. For a real time application like the multilingual speech recognition the PLP method outperformed the other methods of feature extraction. Since, PLP implements bark scale auditory filter bank, the intensity-loudness power-law relation, the equal-loudness curve, critical-band resolution curves and autoregressive modelling for smoothing out the detail from the obtained auditory spectrum, making PLP approximates more closely to human hearing. The recognition accuracy of wavelet packet decomposition method was lower compared to PLP since pre-emphasis and windowing was not carried out. Hence the proposed method performs effectively in real time applications since the size of data is reduced and the system response time is quicker.

8. References

1. Mishra AN, Chandra M, Biswas A, Sharan SN. Robust features for connected Hindi digits recognition. International Journal of Signal Processing, Image Processing and Pattern Recognition. 2011; 4(2):79–90.
2. Iosif M, Todor G. Comparison of speech features on the speech recognition task. Journal of Computer Science. 2007; 3(8):608–16.

3. John SRA, Nithya V. Modified MFCC methods based on KL-transform and power law for robust speech recognition. Journal of Theoretical and Applied Information Technology. 2014 Sep; 67(2):527-32.
4. Ajinkya SM, John SRA. Robust feature extraction methods for speech recognition in noisy environments. IEEE First International Conference on Networks and Soft Computing; 2014.
5. Mallat SA. A theory for multi resolution signal decomposition the wavelet representation.
6. Christos F, Andreas S. Self-organizing hidden Markova model map. Neural Networks. 2013; 48:133–47.
7. Vapnik VN. Statistical Learning Theory. New York: Wiley-Inter science; 1998.
8. Fausett L. Fundamentals of Neural Networks. Prentice Hall.
9. Haykin S. Neural Networks: A Comprehensive Foundation. Englewood Cliffs, NJ: Prentice-Hall, New York; 1999.
10. Available from: http://ecs.utdallas.edu
11. Available from: http://www.speech.kth.se/wavesurfer
12. Available from: http://www.cs.hmc.edu/~kpang/nn/som.html
13. Available from: http://websom.hut.fi/websom