LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification

Eslam Mansour mohammed[1], Mohammed Sharaf Sayed[2],
Abdallaa Mohammed Moselhy[1] and Abdelaziz Alsayed Abdelnaiem[2]

[1]*Department of Electrical and Computer Engineering, Higher Technological Institute, 10th of Ramadan city, Egypt*
[2]*Department of Electronics and Communications Engineering, Zagazig University, Zagazig, Egypt*

eng_emm4g@yahoo.com, msayed@zu.edu.eg, abdallamselhy@yahoo.com, AAabdelnaiem@hotmail.com

Abstract

Automatic language identification plays an essential role in wide range of multi-lingual services. Automatic translators to certain language or routing an incoming telephone call to a human switchboard operator fluent in the corresponding language are examples of these applications that require automatic language identification. This paper investigates the usage of Linear Predictive Coding (LPC) and/or Mel Frequency Cepstral Coefficients (MFCC) with Artificial Neural Network (ANN) for automatic language identification. Different orders for the LPC and MFCC have been tested. In addition, different hidden layers, different neurons in every hidden layers and different transfer functions have been tested in the ANN. Three languages; Arabic, English and French have been used in this paper to evaluate the performance of the automatic language identification systems.

Keywords: *Automatic language identification, Linear Predictive Coding, Mel Frequency Cepstral Coefficients, Artificial Neural Network*

1. Introduction

Nowadays, language identification and recognition system is used to replace hand switch in conference hall to select the appropriate speaker. It has also applications in automatic language translation and routing incoming telephone calls to a human switchboard operator fluent in the corresponding language. In human-based language identification (LID) systems, several people are needed and they have to be trained to properly recognize a set of languages. While automated LID systems have several advantages such as reduced costs and shorter training periods.

Several methods have been used for spoken language identification. These methods depend on domination of vowel and consonant sound, energy of speaker, feature extraction method or model training method. A framework to address the quantization issues which arise in fixed-point isolated word recognition was introduced in [1]. Comparison of recognition rate on the basis of domination of vowel and consonant sound in spoken Hindi Hybrid Paired Words (SHHPW) was presented in [2]. Study of gender identification system based on the energy of speaker utterances, feature extraction using Wavelet Packet Transform (WPT) and gender identification using Artificial Neural Network (ANN) [3]. A technique for the vowel classification using linear prediction coefficient (LPC) with combination of statistical approach and Artificial Neural Network (ANN) was proposed in [4]. Thirty different words are selected

for database and further these Words are categorized into three different groups G1, G2 and G3 on the basis of vowels and consonants using Broad Acoustic classification [5]. A speech recognition system using fuzzy matching method was presented which was implemented on PC [6]. Speech processing and recognition is intensive field of research due to broad variety of applications. Speech recognition is involved in our daily life activities like mobile applications, heather forecasting, agriculture, healthcare, speech assisted computer games, biometric recognition, *etc.* [7]. Gap between paired word act like a speech code and play significant role in recognition process [8, 9].

LPC and MFCC with ANN are among the most usable techniques for spoken language identification. LPC and MFCC methods are used for extracting features of a speech signal and ANN is used as the recognition and identification method. A back propagation method is used to train the ANN. speech signals are sampled directly from the microphone and then they are processed using LPC and\or MFCC methods for extracting the features of speech signal. For each word signal, LPC and MFCC method produces many data. Then, these data become the input of the ANN. The ANN was trained by using huge data training. This data training includes 20 words for every language; each of these words was repeated 10 times. Results show that the highest language identification rate that can be achieved by this system is 100% when MFCC method used for extracting the features of speech signal with order 10 used and ANN with two hidden layers, 30 neurons and linear transfer function for 1st hidden layer, 40 neurons and tan-sigmoid transfer function for 2nd hidden layer and tan-sigmoid transfer function for output layer.

This paper investigates the usage of Linear Predictive Coding (LPC) and/or Mel Frequency Cepstral Coefficients (MFCC) with Artificial Neural Network (ANN) for automatic language identification. The rest of the paper is organized as follows; Section II feature extraction techniques. Section III presents recognition and identification using artificial neural network. Section IV demonstrates and discusses the steps of the recognition and identification system. Section V presents experimental and numerical results. Section VI concludes this paper.

2. Feature Extraction Techniques

A. *Linear Predictive coding (LPC) Analysis*

Speech signal sampled directly from microphone, is processed for extracting the features. Method used for feature extraction process is Linear Predictive Coding using LPC Processor.

The basic steps of LPC processor include the following:

1. Preemphasis: The digitized speech signal, $s(n)$, is put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing [10,11].

The output of the preemphasizer network, is related to the input to the network, $s(n)$, by difference equation:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \quad (1)$$

2. Frame Blocking: The output of preemphasis step, $s\tilde{}(n)$ is blocked into frames of N samples, with adjacent frames being separated by M samples. If $xl(n)$ is the l *th* frame of speech, and there are L frames within entire speech signal,then

$$x_l(n) = \tilde{s}(Ml + n) \quad (2)$$

where $n = 0,1,\ldots,N-1$ and $l = 0,1,\ldots,L-1$

3. Windowing: After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. If we define the window as $w(n)$, $0 \le n \le N-1$, then the result of windowing is the signal:

$$\tilde{x}_l(n) = x_l(n)w(n) \quad (3)$$

Where $0 \le n \le N-1$

Typical window is the Hamming window, which has the form

$$w(n) = 0.54 - 0.46\cos\left\{\frac{2\pi n}{N-1}\right\} \qquad 0 \le n \le N-1 \quad (4)$$

4. Autocorrelation Analysis: The next step is to auto correlate each frame of windowed signal in order to give

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \quad m = 0,1,\ldots\ldots,p \quad (5)$$

Where the highest autocorrelation value, p, is the order of the LPC analysis.

5. LPC Analysis: The next processing step is the LPC analysis, which converts each frame of $p + 1$ autocorrelations into LPC parameter set by using Durbin's method. This can formally be given as the following algorithm:

$$E^{(0)} = r(0) \quad (6)$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|i-j|)}{E^{i-1}} \qquad 1 \le i \le p \quad (7)$$

$$\alpha_i^{(i)} = k_i \quad (8)$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i\, \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \quad (9)$$

$$E^{(i)} = \left(1 - k_i^2\right)E^{i-1} \quad (10)$$

By solving (6) to (10) recursively for $i = 1, 2,\ldots,p$, the LPC coefficient, $m\ a$, is given as

$$a_m = \alpha_m^{(p)} \quad (11)$$

6. LPC Parameter Conversion to Cepstral Coefficients:

LPC cepstral coefficients, is a very important LPC parameter set, which can be derived directly from the LPC coefficient set. The recursion used is

$$c_m = a_m + \sum_{k=1}^{m-1} \left\{\frac{k}{m}\right\} . c_k . a_{m-k} \qquad 1 \le m \le p \quad (12)$$

$$c_m = \sum_{k=m-p}^{m-1} \left\{\frac{k}{m}\right\} . c_k . a_{m-k} \qquad m > p \quad (13)$$

B. Mel Cepstrum analysis

This analysis technique uses cepstrum with a nonlinear frequency axis following *mel* scale [12]. For obtaining *mel* cepstrum the speech waveform *s(n)* is first windowed with analysis window *w(n)* and then its DFT *S(k)* is computed. The magnitude of *S(k)* is then weighted by a series of *mel* filter frequency responses whose center frequencies and bandwidth roughly match those of auditory critical band filters.

The next step in determining the *mel* cepstrum is to compute the energy in this weighted sequence. If *Vl(k)* is the frequency response of *lth mel* scale filter. The resulting energies are given for each speech frame at a time n and for the *l*th *mel* scale filter are

$$E_{mel}(n,l) = \left(\frac{1}{A_l}\right)\sum_{K=L1}^{U1} |V_l(k)S_K|^2 \qquad (1)$$

Where *Ul* and *Ll* are upper and lower frequency indices over which each filter is nonzero and *Al* is the energy of filter which normalizes the filter according to their varying bandwidths so as to give equal energy for flat spectrum.

The real cepstrum associated with *Emel(n,l)* is referred as the *mel*-cepstrum and is computed for the speech frame at time *n* as

$$C_{mel}(n,m) = (\tfrac{1}{N})\sum_{l=0}^{N-1} \log\{E_{mel}(n,l)\}\cos[2\Pi(l+\tfrac{1}{2})/N] \qquad (2)$$

Such *mel* cepstral coefficients *Cmel* provide alternative representation for speech spectra which exploits auditory Principles as well as decorrelating property of cepstrum.

3. Recognition and Identification using Artificial Neural Network

A Neural Networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.

An Artificial Neural Network is used as recognition and identification method. The network has varying neurons input n, which receive input of LPC or MFCC or Both. Number of hidden layer varies from 1 to 4 layers and number of neurons in each hidden layer varies from 10 to 50 neurons. The output of the network is a code of recognized language.

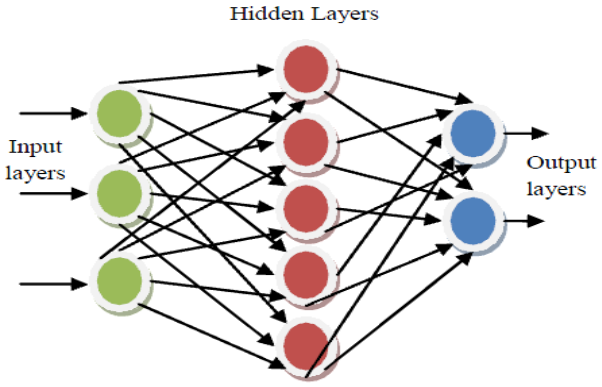

Figure 1. Structure of Multilayered ANN

Figure 1 shows a back propagation algorithm used as the training method of the designed artificial neural network. The basic architecture of a general neural network will be divided into three types of layers- input, hidden and output. The signal will flow stringently in a feed forward direction from input to output. Non linear separable classes are recognized by the extra

layers. In this research work Multi Layer Perceptron network is used which consists, input, one or more hidden and output layers [13].

The field of neural networks has a history of some five decades but has found solid application only in the past fifteen years, and the field is still developing rapidly.

Neural networks have been trained to perform complex functions in various fields of application including pattern recognition, identification, classification, speech, vision, and control systems. Today neural networks can be trained to solve problems that are difficult for conventional computers or human beings.

A backpropagation algorithm includes the following steps:

1. Initialize weights and biases to small random numbers.

2. Present a training data to neural network and calculate the output by propagating the input forward.

3.changing in numbers of hidden layers and transfer function for every hidden layer and for output layer and also changing in number of neurons in every hidden layer until reach to maximum recognition and language identification rate or to minimum error.

4. Steps of Identification and Recognition System

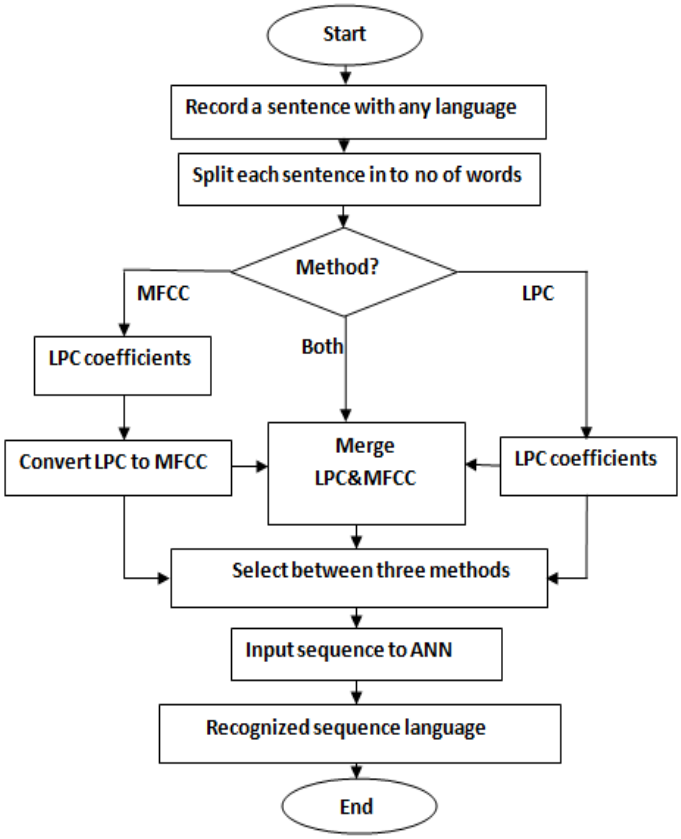

Figure 2. Steps of Language Identification and Recognition System

Figure 2 shows the steps representation of the designed system used for language identification and recognition. First record sentence with any chosen language, and then split

this sentence to number of words. The splitting step depends on two threshold 1st threshold is noise threshold and 2nd is sample threshold .these two threshold determined according to speed of recording and the noise of environment of recording. After that feature extraction process was done for every word using LPC or MFCC or Both (LPC+MFCC). These extracted features feeds into ANN. Finally ANN is used to compare these features with old features stored previously to yield sequence of codes each code refer to specific language.

5. Experimental and Numerical Results

Table I.

LPC order	Training Error %	Testing Error %
3	25.1667	85
5	13.5	55
7	14.66	60
10	**10**	**10**
15	8.1667	50
20	11.5	70

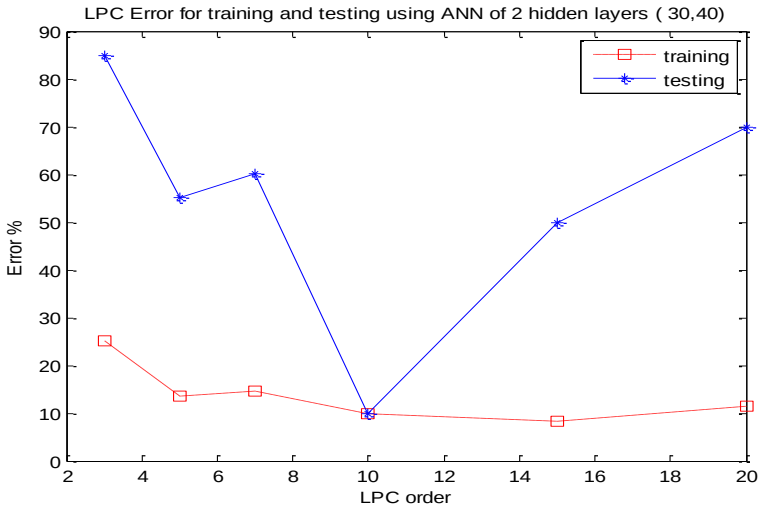

Figure 3. Shows the Performance of the Neural Network using LPC

These experimental and numerical results shown in Table 1 and Figure 3 are taken when LPC used as feature extraction method and using ANN with two hidden layers, 30 neurons and linear transfer function for 1st hidden layer , 40 neurons and tan-sigmoid transfer function for 2nd hidden layer and tan-sigmoid transfer function for output layer.

Table 2.

MFCC order	Training Error %	Testing Error %
3	20	20
5	15	15
7	10	10
10	**0**	**0**
15	5	5
20	4	4

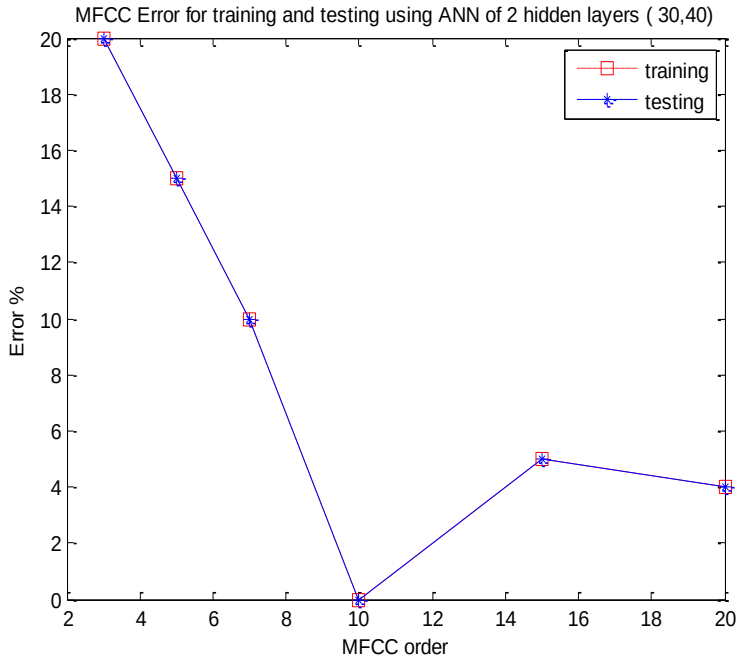

Figure 4. Shows the Performance of the Neural Network using MFCC

These experimental and numerical results shown in Table 2 and Figure 4 are taken when MFCC used as feature extraction method and using ANN with two hidden layers, 30 neurons and linear transfer function for 1st hidden layer , 40 neurons and tan-sigmoid transfer function for 2nd hidden layer and tan-sigmoid transfer function for output layer.

Table 3.

LPC+MFCC orders	Training Error %	Testing Error %
3	10	10
5	15	15
7	10	10
10	**5**	**5**
15	20	20
20	15	15

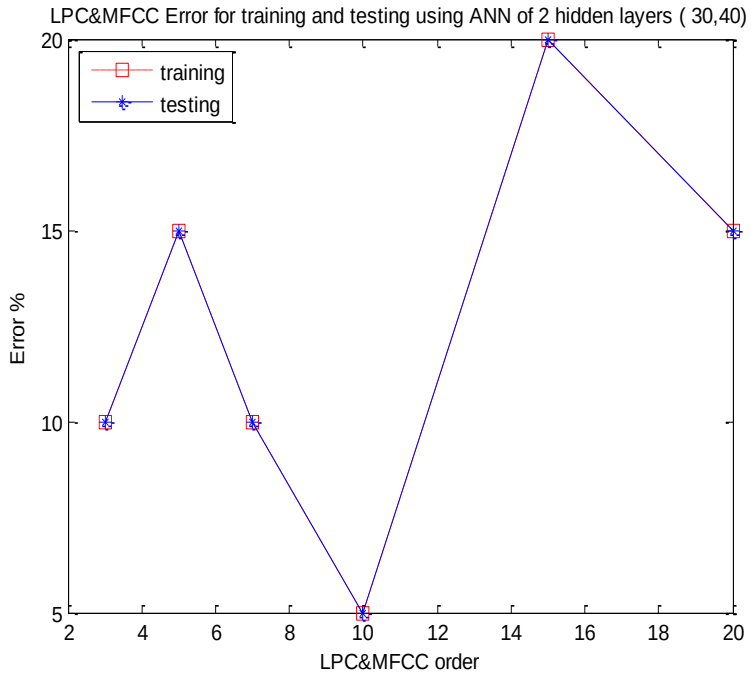

Figure 5. Shows the Performance of the Neural Network using LPC and MFCC

These experimental and numerical results shown in Table 3 and Figure 5 are taken when LPC and MFCC used as feature extraction method and using ANN with two hidden layer, 30 neurons and linear transfer function for 1st hidden layer, 40 neurons and tan-sigmoid transfer function for 2nd hidden layer and tan-sigmoid transfer function for output layer.

Table 4.

LPC order	Training Error %	Testing Error %
3	21.6667	60
5	14.3333	55
7	12.5	70
10	7.6667	60
15	6.5	35
20	4	30

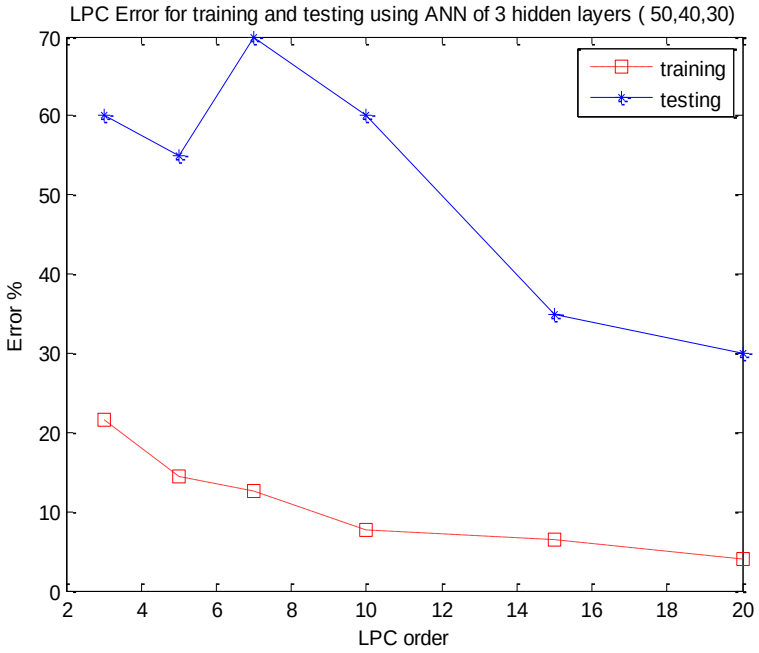

Figure 6. Shows the Performance of the Neural Network using LPC

These experimental and numerical results shown in Table 4 and Figure 6 are taken when LPC used as feature extraction method and using ANN with three hidden layers, 50 neurons and linear transfer function for 1st hidden layer, 40 neurons and tan-sigmoid transfer function for 2nd hidden layer, 30 neurons and tan-sigmoid transfer function for 3rd hidden layer and tan-sigmoid transfer function for output layer.

Table 5.

MFCC order	Training Error %	Testing Error %
3	20	20
5	15	15
7	10	10
10	10	10
15	10	10
20	5	5

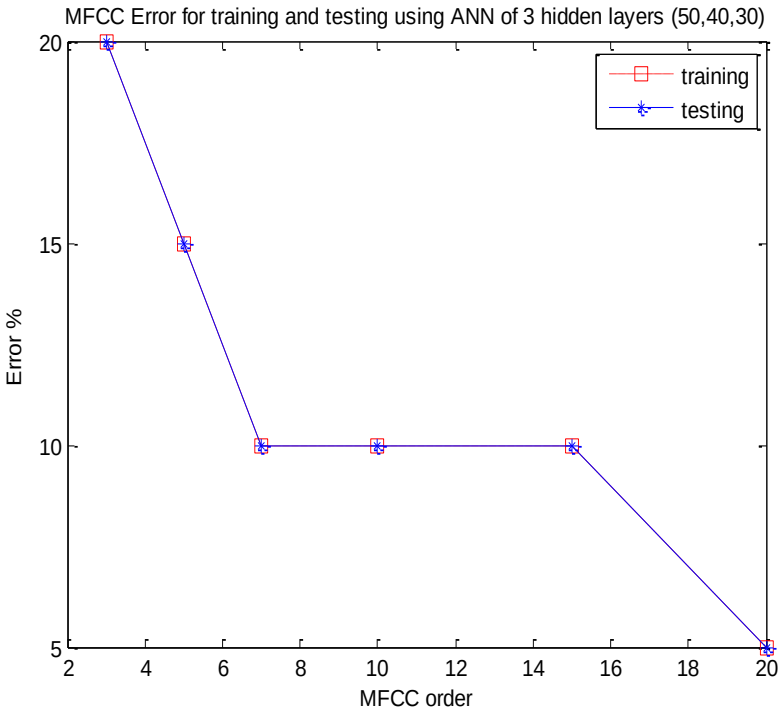

Figure 7. Shows the Performance of the Neural Network using MFCC

These experimental and numerical results shown in Table 5 and Figure 7 are taken when MFCC used as feature extraction method and using ANN with three hidden layers, 50 neurons and linear transfer function for 1st hidden layer, 40 neurons and tan-sigmoid transfer function for 2nd hidden layer, 30 neurons and tan-sigmoid transfer function for 3rd hidden layer and tan-sigmoid transfer function for output layer.

Table 6.

LPC&MFCC orders	Training Error %	Testing Error %
3	25.1667	85
5	10.833	55
7	7.1667	50
10	10	10
15	5	5
20	5	5

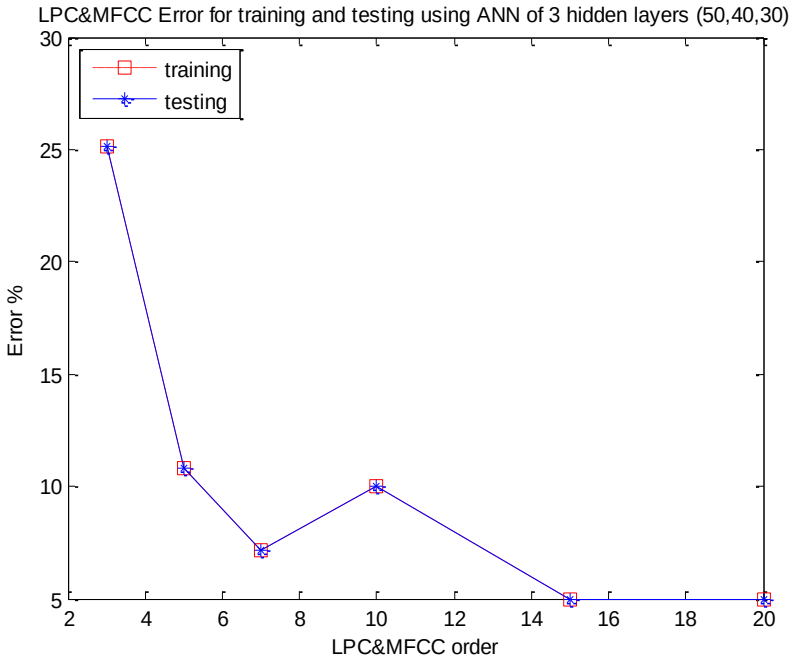

Figure 8. Shows the Performance of the Neural Network using LPC and MFCC

These experimental and numerical results shown in Table 6 and Figure 8 when LPC and MFCC used as feature extraction method and using ANN with three hidden layer, 50 neurons and linear transfer function for 1st hidden layer, 40 neurons and tan-sigmoid transfer function for 2nd hidden layer, 30 neurons and tan-sigmoid transfer function for 3rd hidden layer and tan-sigmoid transfer function for output layer.

6. Conclusion and Discussion

From experimental results, it can be concluded that Mel Frequency Cepstral Coefficient (MFCC) and Artificial Neural Network (ANN) can identify and recognize the speech signal better than using Linear Predictive Coding (LPC). The highest identification and recognition rate that can be achieved is 100%. This result is achieved by using MFCC with order 10 and ANN with two hidden layers, 30 neurons and linear transfer function for 1st hidden layer and 40 neurons and tan-sigmoid transfer function for 2nd hidden layer and tan-sigmoid transfer function for output layer. An ANN is trained using 200 samples per language and tested using 20 samples from different language. Compare with previous work, ANN has been proven that it give better recognition and identification rate.

References

[1] Y. M. Lam, M. W. Mak and P. H. W. Leong, "Fixed point implementations of Speech Recognition Systems", Proceedings of the International Signal Processing Conference. Dallas, **(2003)**.
[2] A. Singh, D. Kumar and V. Singh, "Broad Acoustic Classification of Spoken Hindi Hybrid Paired Words using Artificial Neural Networks", International journal of computer application, (0975-8887), vol. 52, **(2012)** November.
[3] E. Khalaf, K Daqrouq and M. Sherif, "Wavelet Packet and Percent of Energy Distribution with Neural Network Based Gender Identification System", Journal of applied science, ISSN 1812-5654 ,Asian Network for scientific information, vol. 11, no. 16, **(2011)**, pp. 2940-2946.
[4] R. B.Shinde and V. P. Pawar, "Vowel classification based on LPC and ANN", International journal of computer applications (0975-8887), vol. 50, no. 6, **(2012)**.
[5] S. V. Chourasia, K. Samudravijaya and M. Chandwani, "Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database", Proceedings of International Conference on Speech Databases and Assessment, Jakarta, Indonesia, **(2005)**.
[6] L. H. Thiang, "Limited Word Recognition Using Fuzzy Matching", Proceedings of International Conference on Opto-Electronics and Laser Applications. Jakarta, **(2002)**.
[7] L. Rabiner, B. H. Juang and B. Yegnanarayana, "Fundamentals of speech Recognition", 1st edition, Pearson education in south Asia, **(2009)**.
[8] D. K. Rajoriya, R. S. Anand and R. P. Maheshwari, "Enhanced recognition rate of spoken Hindi paired word using probabilistic neural network approach", International Journal of Information and Communication Technology, Inderscience Publishers, Geneva, Switzerland, vol. 3, **(2011)**.
[9] R. Hariharan, J. Hakkinan and K. Laurila, "Robust end of utterance detection for real time speech recognition applications", IEEE International conference on Acoustics, Speech and Signal Processing, **(2001)**.
[10] L. Rabiner and B. H. Jung, "Fundamentals of Speech Recognition", Prentice Hall, New Jersey, **(1993)**.
[11] D. O. Shaughnessy, "Speech Communication: Human and Machine", India, University Press, **(2001)**.
[12] B. Petek and J. Tebelskis, "Context- Dependent Hidden Control Neural Network Architecture for Continuous Speech Recognition", Proceeding IEEE International Conference on Acoustics, Speech and Signal Processing, **(1992)**.