Multi Lingual Speaker Identification on Foreign Languages using Artificial Neural Network

Prateek Agrawal, Harjeet Kaur, Gurpreet Kaur
Lovely Professional University, Phagwara

ABSTRACT
Based on the Back Propagation Algorithm, this paper portrait a method for speaker identification in multiple foreign languages. In order to identify speaker, the complete process goes through recording of the speech utterances of different speakers in multiple foreign languages, features extraction, data clustering and system training. In order to realize the purpose, a database has been prepared which contains one sentence in 8 different international languages i.e. Catalan, French, Finnish, Italian, Portuguese, Indonesian, Hindi, English spoken by 19 distinct speakers, both male and female, in each language. With total size of 760 speech utterances, the average performance of the system is 95.657%. Application of developed system is mainly used in speaker authentication in telephony security oriented applications where the normal conversations are of short durations and the tendency of the spokesperson is to switch language from one to another

Keywords
Artificial Neural Network (ANN), Back Propagation Algorithm (BPA), Cepstral Analysis, Multilingual Speaker Recognition, Power Spectral Density (PSD).

1. INTRODUCTION
Different people can be differentiated on the basis of their speech because they all have different unique speech characteristics which can identify as well as distinguish one person from the other. These features are extracted from the speech utterances using different algorithms for these diverse features. From the signal processing point of view, speech can be characterized in terms of the signal carrying message information. The speech can be representated in the waveforms [1]. In order to build a truly multilingual acoustic model, a strong practical approach to multilingual speech recognition for multiple countries , where more than 200 internationally recognized languages are spoken across the countries, has to be followed. This multilingual model should then be adapted to the target language with the help of a language identification system. Based on the information extracted from the speech signal, it can have three different recognition systems itself: Speaker Recognition, Language Recognition and Speech Text Recognition [11, 12, 14]. Speaker recognition can be separated into two different phases - Speaker Verification and Speaker Identification. The purpose of a speaker verification system is to confirm whether an unidentified voice matches the voice of a speaker whose identity is being claimed. Speaker identification process includes the detection of an unidentified voice from a set of known voices. Speaker verification systems are primarily used in security access control whereas speaker identification systems are mainly used in criminal investigation [3]. Speech recognition encompasses a large number of complex applications such as speech driven consumer applications, speech commissioning in logistics, checking & recording in quality assurance work etc**.** Speaker identification and speaker verification are the most economical and accepted systems for solving the problems of unauthorized use of computer and communications systems and multilevel access control [2]. This paper converges particularly on the problem of identification of spokesperson with simple isolated words spoken in eight different international languages as indicated in Figure 1.

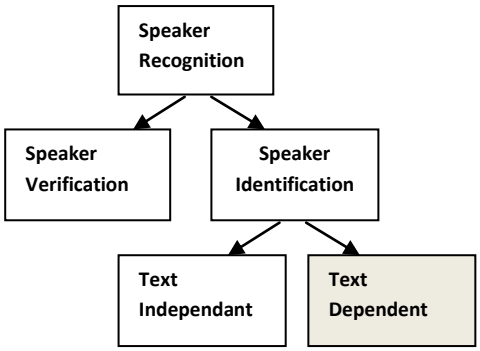


Figure 1: Taxonomy of Speaker Recognition

1.1 Back Propagation Training
One of the potential moves towards the solutions to the problem of speech recognition is ANN [6, 7]. Because of their ability to execute a eminent level of parallel computation, their exalted capability level of robustness and fault tolerance features, ANNs have advantages to handle both speech and speaker recognition. They can learn complex features from the data, due to the non-linear structure of artificial neuron [6, 8]. Different algorithms are being used for training purposes such as BPA, Radial Basis Function, Recurrent networks etc. In present piece of work, the BPA has been used to train the artificial neural network. The Recognition Component in present system is offered by Back Propagation (BP). The motivation behind using BPA is that when the data is increasing at high rates, the traditional approaches never compare results with the growing input size. The ANN performs well on all the available databases and is known to be an excellent tool for such problems. The back-propagation training algorithm is an iterative gradient algorithm designed to minimize the mean square error between the actual output of a multilayer feed-forward perceptron and the desired output. It requires continuous differentiable non-linearity [16].

2. Previous Work
There are number of practical works have been presented where most existing Automatic Speaker Recognition (ASR) systems are used. Authors in [15] proposed the use of a multilayer perceptron (MLP), which is trained using the back-propagation technique to be the engine of an automated digit recognition system using voice. Example for inputting credit card numbers, phone numbers etc. In view of only practically existing speaker recognition systems, most of the practical applications are of the small vocabulary or isolated word type.

Medium- and large vocabulary systems perform well in laboratories but not in real life [3]. Statistical methods like Hidden Markov Model (HMMs), Harmonic Product Spectrum (HPS) algorithm have been used for research multilingual speaker identification system [4]. ANNs have been used to identify the speakers in single language [3]. Text dependent Voice Recognition System [16] worked on eight regional languages containing 904 speech utterances to achieve efficiency of 95.354 % was presented. No research work has been carried out for speaker recognition with multiple foreign languages.

3. Features of Speech

Speech features play an important and useful role to differentiate one speaker from another [13]. In speaker independent speech recognition, premium is placed on extracting features that are somewhat invariant to changes in the speaker [5]. In this work the basic features of speech like Linear Prediction Coefficient (LPC), Linear Prediction Cepstral Coefficients (LPCC), Average Power Spectral Density, Cepstrum Coefficient, Line Spectral Frequency have been extracted using MATLAB.

3.1 Linear Prediction Coefficients

LPC determines the coefficients of a forward linear predictor by reducing the prediction error in the least squares sense. It has applications in filter design and speech coding.

[a,g] = lpc(x,p)

finds the coefficients of a pth-order linear predictor (FIR filter) that predicts the current value of the real-valued time series x based on past samples.

3.2 Linear Prediction Cepstral Coefficients (LPCC)

Linear frequency cepstral coefficient is found by using Discrete Fourier Transformation (DFT) by formula

$$LFCC_i = \sum_{k=0}^{n} Y_k * \cos(\frac{\prod * i * k}{N})$$

where i represents number of coefficients and k is then number of DFT.

3.3 Average power spectral density

This is another important feature of speech signal. PSD is intended for continuous spectra. The integral of the PSD over a given frequency band computes the average power in the signal over that frequency band. In contrast to the mean-squared spectrum, the peaks in the spectra do not reflect the power at a given frequency. Average PSD of any signal can be calculated as the ratio of total power to the frequency of the signal. [17]

3.4 Cepstrum Coefficients

The cepstral coefficients are the coefficients of the Fourier transform representation of the logarithm magnitude spectrum of the most distinctive feature that helps to differentiate the speakers.. Consider a sequence, x(n), having a Fourier transform $X(\omega)$. The cepstrum, $c_x(n)$, is defined by the inverse Fourier transform of $C_x(\omega)$, where $C_x(\omega) = \log_e X(\omega)$. [18]

3.5 Line Spectral Frequency

The LPC to LSF/LSP Conversion block takes a vector or matrix of linear prediction coefficients (LPCs) and converts it to a vector or matrix of line spectral pairs (LSPs) or line spectral frequencies (LSFs). When converting LPCs to LSFs, the block outputs match those of the *poly2lsf* function. [18]

4. Approach

Presented work has been done in two main parts, first multilingual speech analysis/synthesis and second creation of neural network model as a classifier for speaker identification. Multilingual Speech Analysis/Synthesis consists of collection of speech utterances, pre-processing of speech utterances, features extraction, clustering of entire featured data. This is the first sub-process of this work and it consists of sentences of different languages taken from a public domain. Sound wave files (.wav) are created by using microphone connected with Personal computer at sampling rate 44,100 KHz and 16 bit per sample with stereo channel. For recording of sentences Free Sound Recorder 9.3.1 has been used and for pre-processing software Cool Record Edit Pro. For developing the speech database one sentence (*Now this time your turn*) is recorded from 19 speakers (*11 male and 9 female*). The sentence is translated in 8 different languages English, Hindi, Catalan, Finnish, Portugese, French, Italian and Indonesian as shown in Table I. The sentence is taken in such a way that in each word a good combination of consonant and vowel appear. The reason behind this is, whenever we pronounce any letter a vowel sound is always generated .For the sentence considered all five words are used in all eight languages, collectively 760 words are recorded in eight different international languages.

Table I: Languages and Sentences

SNo	Language	Sentence
1.	English	Now this time your Turn
2.	Hindi	Abb is samay tumhari baari
3.	Catalan	Ara aquesta vegada el torn
4.	French	Maintenant cette fois votre tour
5.	Finnish	Nyt talla kertaa sinun vuorosi
6.	Portugese	Agora esta vez transformer oseu
7.	Italian	Questa volta il tuo tourno
8.	Indonesian	Sekarang kali ini giliran anda

Pre-processing includes resampling sound files from Stereo channel to Mono channel by clipping words from the sentence of particular language, reducing noise from the file (Figure 2 (a) and (b)). Silence and noise are removed from speech signal and then splitting each sentence into individual words in each language for every user, in this way speech database of 760 words is created.

Figure 2: (a) Data before resampling (Stereo Channel)

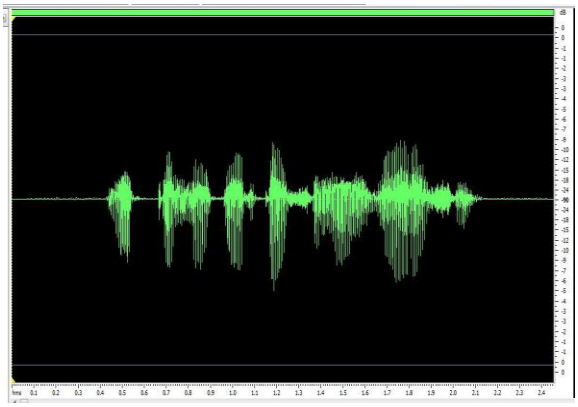


Figure 2: (b) Data after resampling (Mono Channel)

The *.wav* files containing complete sentences are split into individual *.wav* files containing one word each as shown in Figure 3.

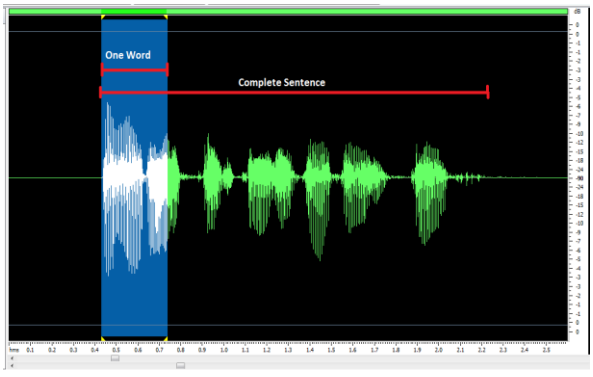


Figure 3: Splitting of sentences into individual word

Splitting is done so that features of individual words can be extracted and stored in the database so that when matching is done, each and every word can be matched and this can only be done if features of individual words are already stored in speech database. This improves accuracy of the system.

Next step is extracting the features from these speech signals with the help of MATLAB (A computational language). Five types of features are extracted – LPC, LPCC, Cepstrum coefficient, , average PSD and LSF. Out of these features, cepstrum coefficient is the feature with maximum knowledge. Figure 4 shows how the features extracted from speech samples have been stored in speech database in MATLAB. The columns indicate speech utterances and the rows indicate features. Eight languages per speaker and five words each language makes it forty utterances for each speaker. Therefore for nineteen speakers, total number of utterances would be 760. With fourteen extracted features, the matrix size is 14x760. After obtaining these exclusive attributes from every utterance, the system is trained by using BPA with single hidden layer. Numbers of hidden neurons are being kept less than the target numbers. The training parameters for momentum, maximum epochs, non linear function, umber of hidden layers, number of neurons per layer, number of targets, training parameter are illustrated in Table II.

Table II: Training Parameters

SNo	Parameter	Value
1.	Momentum	0.001
2.	Train Function	trainlm
3.	Training Parameter	0.30
4.	Maximum Epochs	1000
5.	Non-Linear Function	log-sigmoid
6.	Number of Hidden Layers	1
7.	Number of neurons in Hidden Layers	40
8.	Number of Target	19

Subsequent to the completion of the training, the next move is to simulate trained network and to check whether the class of actual output and target output is same. For simulation, a sample data is taken from the input set for testing and it is imitated with the trained network. A target matrix shown in Figure 5, mapped with the dimensions of input matrix is created by assigning each column with a unique value ranging between 0 and 1. Figure 6 shows Levenberg-Marquardt (trainlm) neural network training with 1000 iterations and it took 52 minutes and 51 seconds to get trained

5. Results

When the proposed system is trained with Artificial Neural Network using Back Propagation Algorithm, keeping the number of Layers fixed, the system was able reach the performance of 95.657 % giving a total of 33 errors on a total number of 760 utterances (Input data) with 14 features. Table III shows the final training results in terms of Speakers, Number of input utterances, Number of errors and the percentage efficiency for nineteen speakers.

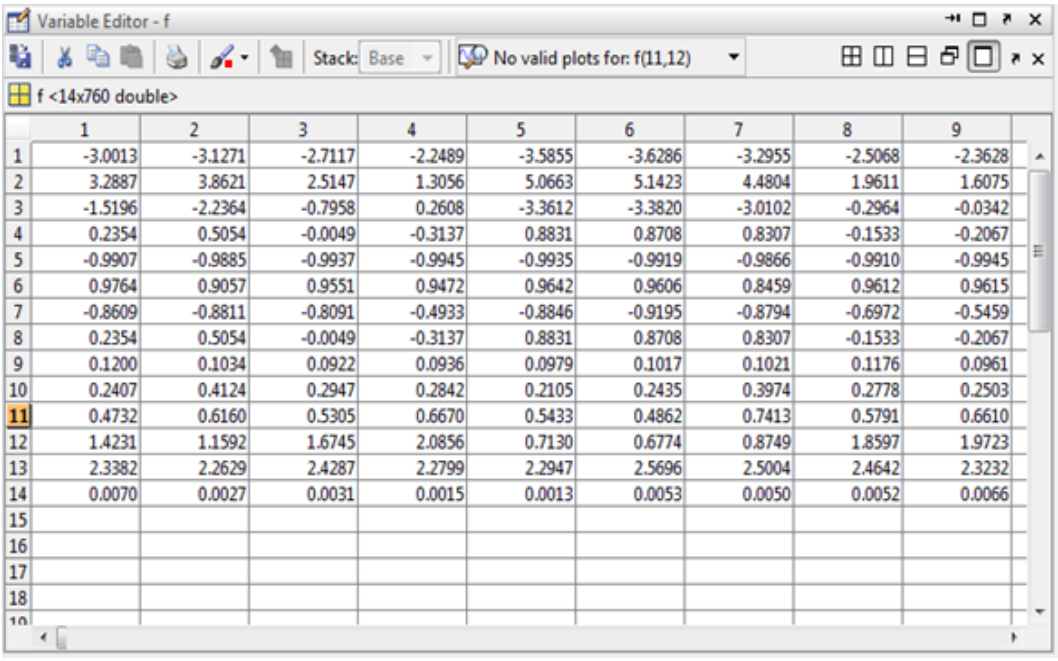
Table III: Training Result Data

Speaker Number	Number of Input Utterances	Number of Errors	Efficiency (%)
1	40	0	100
2	40	1	97.5
3	40	2	95
4	40	1	97.5
5	40	1	97.5
6	40	0	100
7	40	3	92.5
8	40	0	100
9	40	4	90
10	40	2	95
11	40	3	92.5
12	40	0	100
13	40	2	95
14	40	3	92.5
15	40	1	97.5
16	40	3	92.5
17	40	1	97.5
18	40	2	95
19	40	4	90
	Σ=760	Σ=33	Σ=95.657

The regression graph is shown below in Figure 7. Y is the output matrix obtained after training the input matrix under the defined conditions and parameters. T is the target matrix. Figure 8 gives the graphical representation of Mean square error w.r.t number of iterations. When the system is trained with ANN BPA as used by researchers, [16] for almost same problem with limited Indian Languages only, the performance was 95.354% with 42 errors. With the introduction of foreign languages and more words, the average performance of our system is of **95.657%** giving a total of **33 errors**.

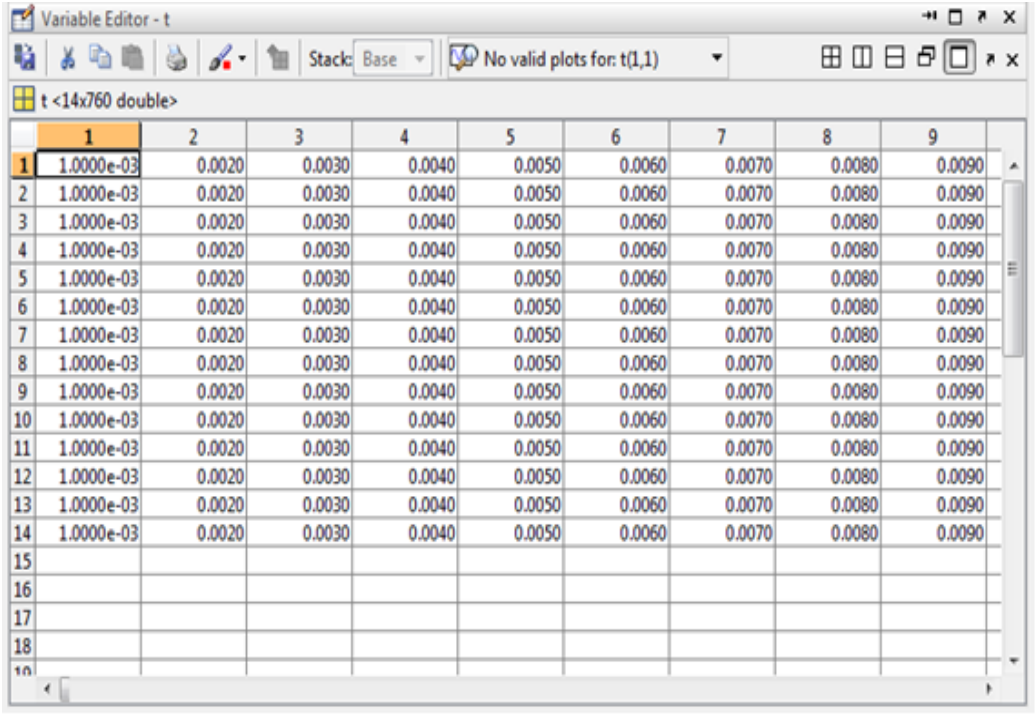
6. Conclusion

When using Back Propagation Algorithm, keeping the number of Layers fixed, the system was able reach the performance of 95.657 % giving a total of 33 errors on a total number of 760 utterances (Input data) with 14 features. Due to PC configuration problem number of neurons and number of layers could not be increased beyond a certain limit. Mean square error decreases with number of iterations and becomes almost stable after about 1000 iterations. Adding more features to system improves the user identification rate for the system. The result shows that BPA can be used for multi language system. This research focuses on text dependent speaker recognition. The desired goal of a system which can understand a text independent expression uttered by different speakers using various languages in different environments can be the further enhancement of this research.



	1	2	3	4	5	6	7	8	9
1	-3.0013	-3.1271	-2.7117	-2.2489	-3.5855	-3.6286	-3.2955	-2.5068	-2.3628
2	3.2887	3.8621	2.5147	1.3056	5.0663	5.1423	4.4804	1.9611	1.6075
3	-1.5196	-2.2364	-0.7958	0.2608	-3.3612	-3.3820	-3.0102	-0.2964	-0.0342
4	0.2354	0.5054	-0.0049	-0.3137	0.8831	0.8708	0.8307	-0.1533	-0.2067
5	-0.9907	-0.9885	-0.9937	-0.9945	-0.9935	-0.9919	-0.9866	-0.9910	-0.9945
6	0.9764	0.9057	0.9551	0.9472	0.9642	0.9606	0.8459	0.9612	0.9615
7	-0.8609	-0.8811	-0.8091	-0.4933	-0.8846	-0.9195	-0.8794	-0.6972	-0.5459
8	0.2354	0.5054	-0.0049	-0.3137	0.8831	0.8708	0.8307	-0.1533	-0.2067
9	0.1200	0.1034	0.0922	0.0936	0.0979	0.1017	0.1021	0.1176	0.0961
10	0.2407	0.4124	0.2947	0.2842	0.2105	0.2435	0.3974	0.2778	0.2503
11	0.4732	0.6160	0.5305	0.6670	0.5433	0.4862	0.7413	0.5791	0.6610
12	1.4231	1.1592	1.6745	2.0856	0.7130	0.6774	0.8749	1.8597	1.9723
13	2.3382	2.2629	2.4287	2.2799	2.2947	2.5696	2.5004	2.4642	2.3232
14	0.0070	0.0027	0.0031	0.0015	0.0013	0.0053	0.0050	0.0052	0.0066
15									
16									
17									
18									

Figure 4: Matrix showing extracted features from speech signals



	1	2	3	4	5	6	7	8	9
1	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
2	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
3	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
4	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
5	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
6	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
7	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
8	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
9	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
10	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
11	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
12	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
13	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
14	1.0000e-03	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
15									
16									
17									
18									

Figure 5: Target Matrix

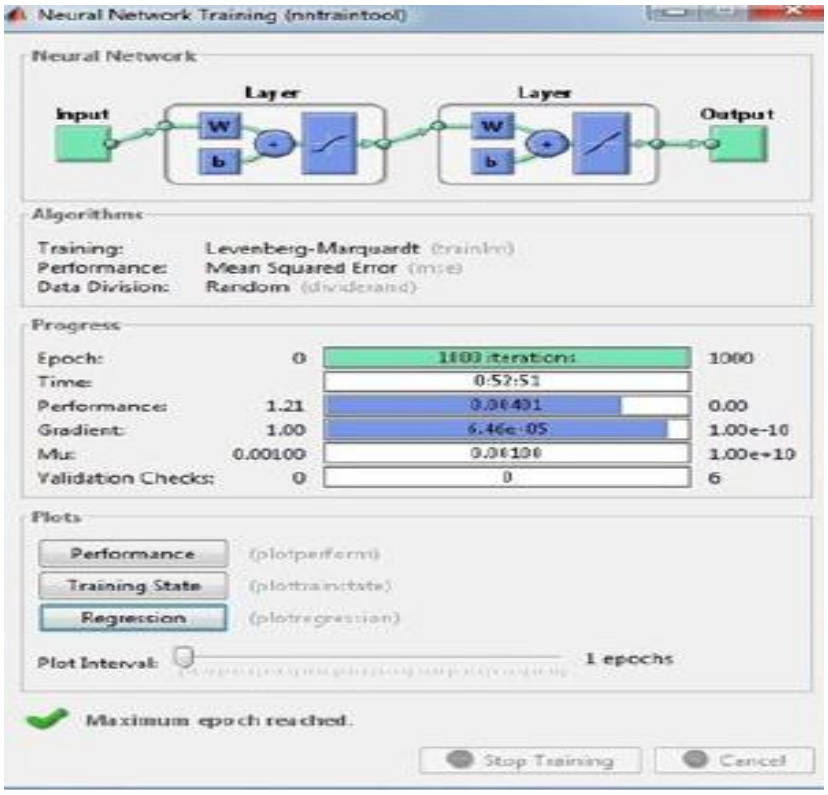


Figure 6: Levenberg-Marquardt (trainlm) neural network training

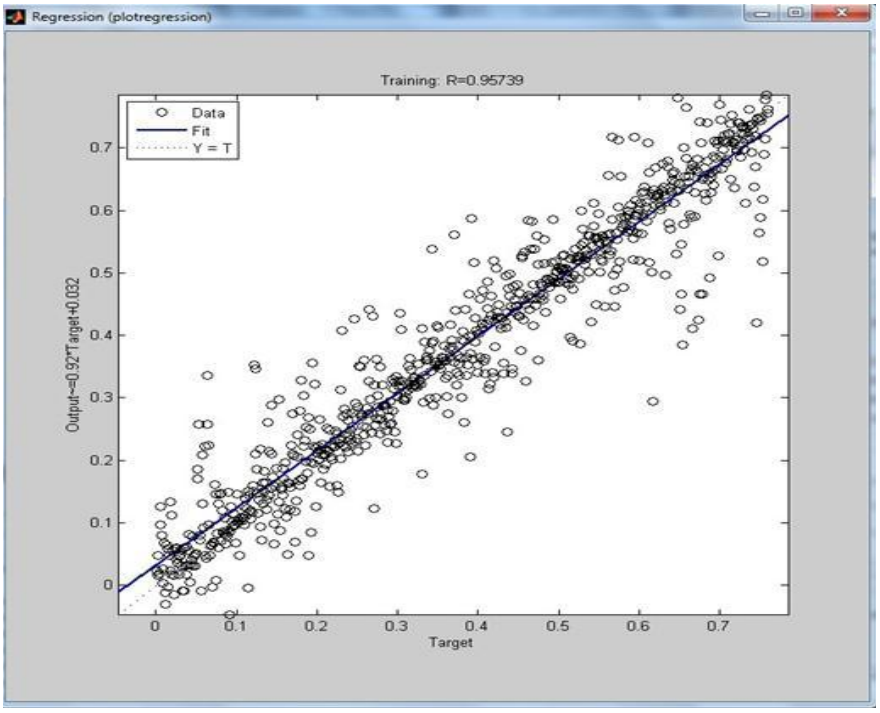


Figure 7: Regression Graph

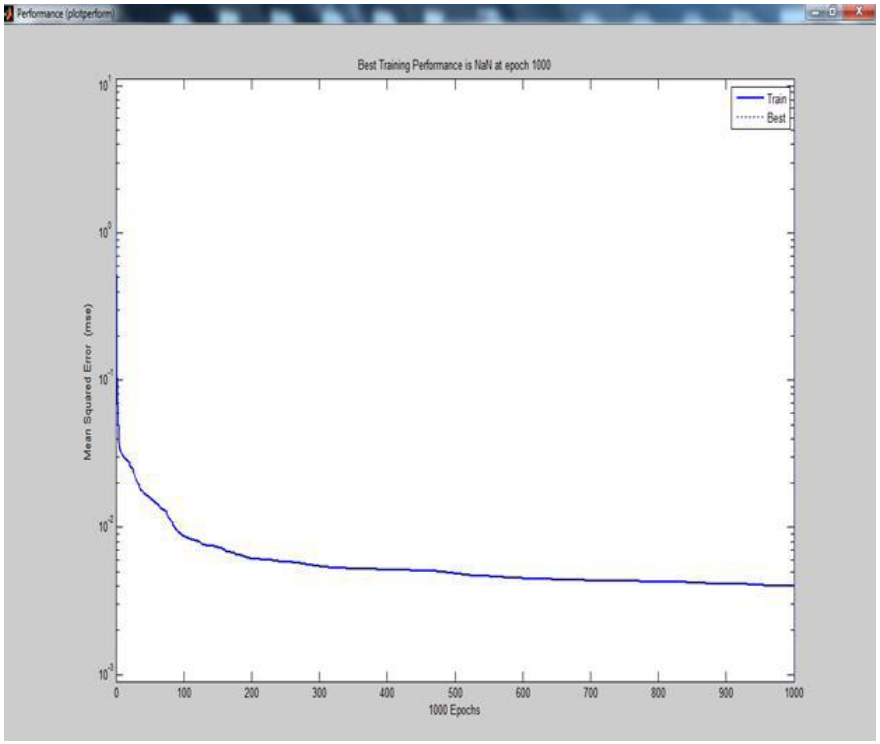


Figure 8: Mean square error w.r.t number of iterations

7. REFERENCES

[1] Love,C. and Kinsne,W. , A Speech Recognition System Using A Neural Network Model For Vocal Shaping, Department of Electrical and Computer Engineering, University of Manitoba Winnipeg, Manitoba, Canada R3T- 2N2, March 1991, 198 pp.

[2] Chougule,S. and Rege,P., Language Independent Speaker Identification, IEEE 3rd international conference, ieeexplore pp 364-368, May 2006.

[3] Mak,M.W.; Allen,W.G. and Sexton,G.G., Speaker Identification Using Radial Basis Functions University of Northumbria at Newcastle, U.K., IEEE 3rd International Conference, ieeexplore pp. 138-142, 1993

[4] Bum,D., Experiments On Neural Net Recognition Of spoken And Written, Text IEEE Transaction on Acoust.. Speech and Signal Proc., vol. ASSP-36. no. 7. pp. 1162-1168, 1988.

[5] Deiri,Z.and Botros ,N., LPC-Based Neural Network For Automatic Speech Recognition, Proc. IEEE Engineering in Medicine and Biology Soc., IEEE Cat. no. 90 CH2936-3. pp. 1429-1430, 1990

[6] Philip P. Wassermann,, Neural Computing: Theory and Practice , VNR, New York, 1989

[7] Richard P. Lipmann, Review of Neural Networks for Speech Recognition, Neural Computation 1, pp.1-38, Massachusets Institute of Technology, 1989.

[8] Zebdum, RS; Guedes, K; Vellasco, M.; Pacheco, M.A,. Short Term Load Forecasting Using Neural Nets , Proceedings of the International Workshop on Artificial Neural Networks, LNCS No. 930, Springer Verlag, Torremohos, Spain, JuneJ995

[9] Zurada ,J. M., Introduction to W c i a l Neural Systems , West Publishing Company, 1992.

[10] K. Rahul, S. Anupam, T Ritu, Fuzzy Neuro Systems for Machine Learning for Large Data Sets , Proceedings of the IEEE International Advance Computing Conference, ieeexplore, pp 223-227, March 2009, Patiala, India

[11] K. Santhosh, C. Mohandas ,V. P. and H. Li, Multilingual Speech Recognition: A Unified Approach , Proceedings of Interspeech 2005, (Eurospeech - 9th European Conference on Speech Communication and Technology), Lisboa, Sept. 2005.

[12] S. C. Kumar; L. Haizhou, Language identification System for Multilingual Speech Recognition Systems , Proceedings of the 9th International Conference Speech and Computer (SPECOM 2004), St. Petersburg, Russia, Sept. 2004.

[13] S. Hanwu, M.Bin, L. Haizhou, An Efficient Feature Selection Method for Speaker Recognition , International Symposium on Chinese Spoken Language Processing, December 16-19, 2008, China.

[14] M. Bin, G.Cuntai, L.Haizhou, L.Chin-Hu, Multilingual Speech Recognition with Language Identification , International Conference on Spoken Language Processing (ICSLP), DENVER-COLORADO, Sept. 16-20, 2002.

[15] N. Mohini, P.Manoj, P. Vijay, M.K. Prabhat, Vocal digit recognition using Artificial Neural Networks ,Dept. of Computer Engineering, VPCOE, Baramati in (2010 2nd International Conference on Computer Engineering and Technology) Page(s): V6-88 - V6-91

[16] A. Prateek, S. Anupam and T. Ritu, Multi Lingual Speaker Recognition Using Artificial Neural Networks , Advances in Computational Intelligence, Springer VerLlog,, 2009 , pages 1-9.

[17] http://www.sp4comm.org/webversion.html

[18] http://www.mathworks.in/help/toolbox/dsp/ref/lpctofromcepstralcoefficients.html