International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.1 (2014), pp.109-120
http://dx.doi.org/10.14257/ijsip.2014.7.1.11

Neural Network based Classification for Speaker Identification

V. Srinivas[1], Dr. Ch. Santhi rani[2] and Dr. T. Madhu[3]

[1, 3]*Swarnandhra Institute of Engineering and Technology, Narsapur*
[2] *D.M.S&S.V.H Engineering College, Machilipatnam*
[1]*srinivas.siet@gmail.com*, [2]*santhirani.ece@gmail.com*, [3]*principal.siet@gmail.com*

Abstract

Speaker Recognition is a challenging task and is widely used in many speech aided applications. This study proposes a new Neural Network (NN) model for identifying the speaker, based on the acoustic features of a given speech sample extracted by applying wavelet transform on raw signals. Wrapper based feature selection applies dimensionality reduction by kernel PCA and ranking by Info gain. Only top ranked features are selected and used for neural network classifier. The proposed neural network classifier is trained to assign a speaker name as label to the test voice data. Multi-Layer Perceptron (MLP) is implemented for classification, and the performance is compared with the proposed NN model.

Keywords*: Speaker Recognition, Speech Processing, Principle Component Analysis (PCA), Kernel PCA, Wrapper based extraction, Multi-Layer Perceptron (MLP)*

1. Introduction

Speaker recognition is one of the processes of speech processing, wherein the speaker is identified by recognizing the spoken phrase. Speech verification is used to decide if a speaker is whom he claims to be. Many speech aided applications require automatic speaker recognition technologies. A speech conveys linguistic and speaker information. Linguistic information contains the message and language in a speech. Speaker information contains emotional, regional and physiological view of the speaker [1]. Humans have the ability to decode the speech signals and understand the information in speech and recognize the speaker. This perception and understanding abilities of humans are needed in many applications such as voice command control, audio archive indexing and audio retrieval etc. The tasks used for speaker recognition are speaker identification, speaker verification or detection and Segmentation and Classification [2].

1) **Speaker Identification**: Identifies a speaker out of a collection of known speakers using a given voice sample.

2) **Speaker Verification or Detection**: Authentication of the speaker is verified by designing a binary decision problem.

3) **Speaker segmentation and classification**: Ether speech of an individual or when speech of individual is intermixed with other's speech is given as a sample; the desired speech segment must be separated before recognition. This task is useful in multi-speaker recognition problems. Here, the given audio is segmented into homogeneous audio segments and labels are assigned to identify the speaker.

Automatic speaker recognition may be text dependent or text independent. In text dependent category, user gives the text of speech to the system and the knowledge of the phrase is useful for better recognition. In text independent task, the system does not know the phrase spoken in the speech. This will increase the flexibility of recognition but

ISSN: 2005-4254 IJSIP

reduces the accuracy proportional to the amount of speech. The basic structure of speaker recognition system [3] is given in the Figure 1.

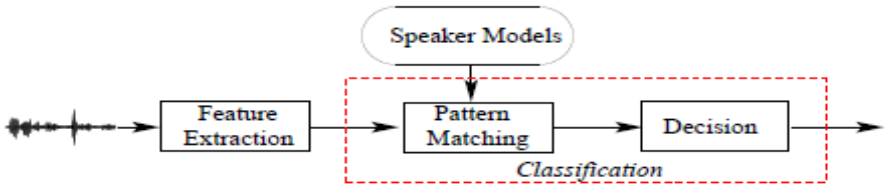

Figure 1. Steps of General Speaker Recognition Problem

Feature Extraction gives the speaker specific information from the given speech signal by performing complex transformations. Different levels of transformations are performed by using semantic, acoustic, phonologic and acoustic features. Pattern matching module compares or matches the extracted features with speaker models such as hidden Markov model, dynamic time wrapping model and vector quantization model. Decision model finds similarity score between the given test sample and the claimed speaker to recognize the speaker.

The acoustic features contain the characteristic information of the speech signal and are useful for recognizing the speaker. Widely used acoustic features are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Prediction Cepstral (PLPC). MFCC features [4] are derived from Fast Fourier Transform (FFT) power spectrum. The centre and the bandwidth of the filter bank are selected based on the Mel-frequency scale. Therefore, this feature gives more details on the low frequencies. Then the filter bank details are transformed by Discrete Cosine Transformation (DCT) to get the features of the given voice signal. MFCC is widely used in many speaker recognition problems.

LPCC has the adaptive details and uses the all-pole model that represents a smoothed spectrum [5]. All-pole representation is transformed by a recursive formula based on the prediction and cepstral coefficient of the poles. PLPC features are hybrid of the filter bank and all-pole model of spectral representation. Therefore, the details of the features are determined by both the filter bank and all-pole representation.

In this proposed study, a neural network model is used for speaker recognition. From the input speech signals, the features are extracted by Wavelet Transform (WT) as it decomposes non stationary complex signals such as music, speech and images into elementary forms with good precision [6]. After extracting features, feature reduction will reduce the run time and selection of best features will improve the efficiency of recognition/classification. Principle Component Analysis (PCA) is a dimensionality reduction method used for many classification applications. Recently, PCA and Linear Discriminate Analysis (LDA) are used for optimizing the transformation and reduce the dimensions of the feature space for acoustic modelling when multi class scenarios are used [7]. In this research, a wrapper based feature selection is used by combining feature reduction by kernel based PCA and ranking by Info Gain. Then the classifiers such as MLP and proposed NN are used to assign a label to the input speech.

2. Related Works

Traditional speaker Recognition models used MFCC features. Nakagawa, *et al.*, [8] combined the phase information along with MFCC features to increase the recognition rate. Experiments were conducted with 35 Japanese speakers, and the classifier was trained in 5 sessions over 10 months. Results showed the reduced error rate when comparing the traditional methods that used only the MFCC feature. Chelali, *et al.*, [9]

developed a MLP classifier based on text dependent PLP features extraction by performing Fourier Transform of the input voice signal. 14 Arabic phonemes and 4 native Arabic speakers were used for training the MLP.

Kral [10] applied WT for the tasks in signal processing. Experiments were conducted with 2 Czech preachers and 10 to 50 native speakers. The Gaussian Mixed Model (GMM) and MLP were used as classifiers. Features were extracted by using FT and WT. Results revealed that high precision was reached when using WT than feature extraction by Fourier transform.

Speaker identification system should be robust and efficient to extract features from noisy input signals. As WT had time frequency and multi-resolution properties, Mohmoud, *et al.*, [11] applied WT in speaker identification systems. Mel-Frequency cepstral coefficients were extracted from the input voice signals. Test pattern was taken with white Gaussian noise of 20 dB S/N ratio and recognition rate of 97.3 % was achieved. Applying WT for authentication in expert systems was proposed by Wu and Lin [12].

Hilal, *et al.*, [13] investigated Discrete Wavelet Transform (DWT) and Power Spectral Density based feature extraction and correlation of coefficients for speaker recognition in mobile systems for enhanced security by checking the passwords and PIN numbers. When the correlation value was above than a predetermined threshold then sample was assigned a predetermined label. Experimental results gave 95 % recognition rate. Ziolko, *et al.*, [14] combined the benefits of FFT and WT for extracting features from speech signals. FFT was capable of expressing time frequency changes, but analyzing windows might create artefacts. Performing WT on input speech signals then applying FFT on WT coefficients made finding coefficients in the same frequency and in the same domain, easier to locate.

A novel semi-supervised speaker identification method to identify the non-stationary influences such as session dependent variation, recording environment change and physical conditions or emotions were proposed by Yamada, *et al.*, [15]. Voice quality variants were expected to follow the covariate shift model, where voice feature distribution alone changes in training and test phases. The proposed method used kernel logistic regression and cross validation weighted versions, and it was capable of mitigating covariate shift influence. Experiments showed through the text-independent/dependent speaker identification simulations that the proposed method promises much with regard to voice quality variations.

Kekre and Kulkarni[17] presented a Vector Quantization (VQ) method for Speaker Identification consisting of training and testing phases, and VQ was used for feature extraction in both. Two variations were used. In method A, codebooks generated from speech samples were converted into 16 dimensional vectors with an overlap of 4. In method B, speech samples generated codebooks were converted into 16 dimensional vectors without overlap. Test sample codebook was generated and compared with database stored reference samples codebooks for speaker identification. Results from both schemes showed that method B provided slightly better results than method A.

Zhao proposed local spatiotemporal descriptors for visual based speaker recognition and representation. Spatiotemporal dynamic texture features of local binary patterns were extracted from localized mouth regions to describe motion information in utterances, which captured spatial/temporal transition characteristics. Structural edge map features were extracted from image frames to represent appearance characteristics. Combining dynamic texture and structural features had motion and appearance together, providing description ability for speech's spatiotemporal development. The proposed method got promising recognition results on experiments on BANCA and XM2VTS databases, compared to the other features.

Zhang developed a text independent classifier for recognizing a speaker. The sample spaces were spanned by PCA. PCA Combined with classifier was applied for speeches of

YOHO CORPUS. Promising results were achieved in this experiment. Zhou and Shan designed a Probabilistic Neural Network (PNN) to recognize the speaker. When the samples are more for training the classifier, the redundancy was so high. Therefore, PCA was used to reduce the characteristic parameters and optimize NNs.

Takiguchi and Ariki used Kernel PCA for extracting features from the distorted speech signals. Traditional methods used DCT for extracting MFCC features. MFCC feature limits the details on low order features. Therefore, Kernel PCA was used to project the noise in high ordered spectrum. Khan and Farooq proposed Principle Component Analysis- Linear Discriminate Analysis Feature Extraction for Pattern Recognition algorithms. Nisha and Jayasheela analyzed various classifiers that can be used for speaker recognition. Authors used filter banked cepstral parameters by applying Fourier transforms on short timing windows. For designing the classifier, Gaussian Mixture Model, Hidden Markov model, Stochastical model and Artificial Neural Network were implemented for text independent speaker identification. The classifier results were analyzed by false rejection and false acceptance.

3. Materials and Methods

In this study, from the given speech signals, the features are extracted by Wavelet Transformation. Best features are extracted by dimensionality reduction by PCA and kernel PCA. To select a best feature, the reduced features are ranked, and top ranked features are used for training the proposed classifier. The flowchart of the proposed classifier is given in the following Figure 2.

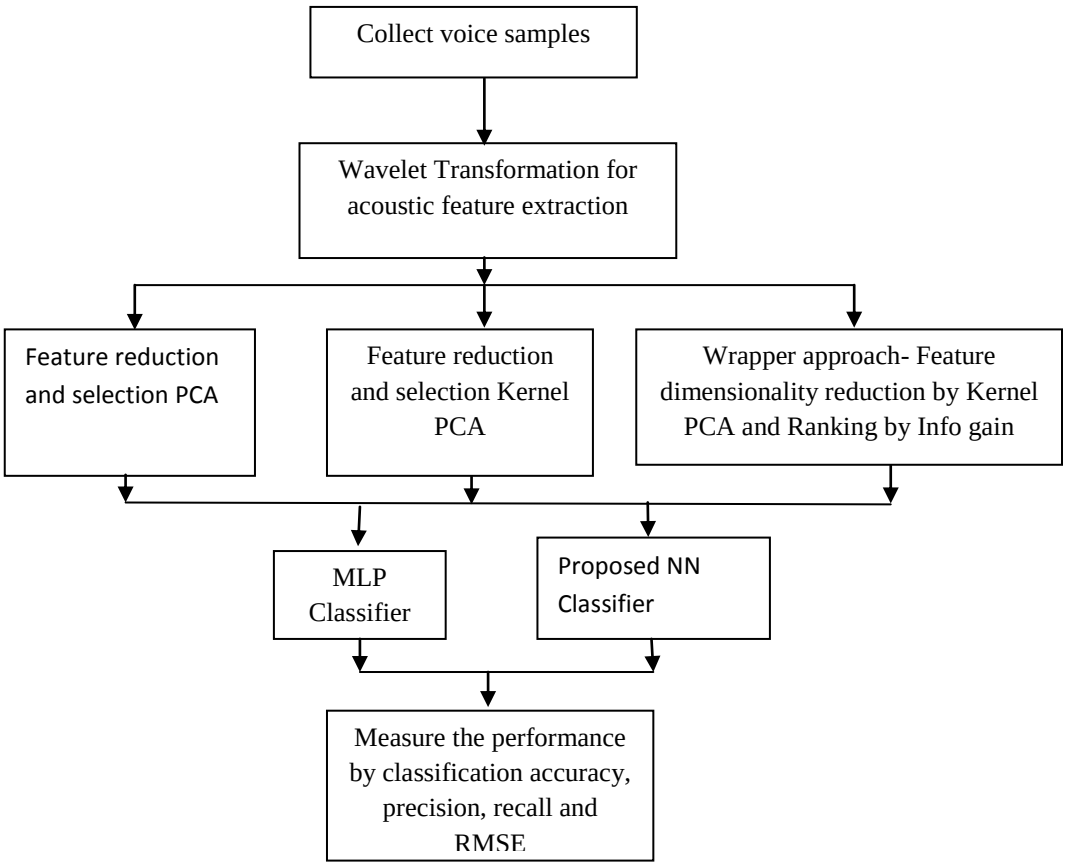

Figure 2. Flow Chart of the Methodology

Wavelet Transform

In 1982, Wavelet transform was introduced by Jean Model for the analysis of seismic wave analysis. Wavelets are a family of functions which are constructed from the translations or dilations of a single wavelet called the mother wavelet

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right), \quad a,b \in R,\ a \neq 0$$

Where 'a' is a scaling parameter, and it denotes the degree of compression and 'b' has the time locations of the wavelet.

Continuous Wavelet Transform (CWT) is known by the wavelet function ψ adding signal times multiplied by scaled and shifted versions. Mathematically the continuous wavelet is defined by the coefficients given by

$$C(scale, position) = \int_{-\infty}^{\infty} f(t)\psi(scale, position, t)dt$$

Original signals constituent wavelets are got by multiplying coefficients by applicable scaled and shifted wavelets. Daubechies proposed symlets-symmetrical wavelets are obtained by modifying indications of the db family. db wavelets have maximal phase while symlets have minimal phase.

Kernel Principal Component Analysis

Kernel PCA is used for reducing the dimensionality of the feature space . The given non-empty data set 'X' is represented as $x_1, x_2, \ldots x_m$; k is a definite positive kernel and used for nonlinear similarity measure. The principal components are the data set X is retrieved by mapping $\phi(x_1),...,\phi(x_m)$

Then covariance matrix is calculated by the following formula,

$$C := \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\phi(x_i)^T$$

C can be diagonalized by calculating the m Eigen values λ, and solution v.

$$Cv = \lambda v$$

The solution v is expanded by,

$$v = \sum_{i=1}^{m}\alpha_i\phi(x_i)$$

and

$$m\lambda\alpha = K\alpha$$

Where $\alpha = (\alpha_1,...,\alpha_m)^T$ and $K_{ij} = k(x_i, x_j)$

Then p numbers of features are extracted by

$$\langle v^p, \phi(x)\rangle = \frac{1}{\sqrt{\lambda^p}}\sum_{i=1}^{m}\alpha_i^p k(x_i, x)$$

Feature Ranking by Info Gain

Info gain is a measure used to find the in formativeness or importance of a feature to resulting classes. Info Gain (IG) of feature 'A' is calculated by the difference between

expected information for classification of the data set D and actual information needed for classification by selecting a feature 'A'. The IG formula is,

$$\text{GAIN}(\text{A}) = \text{Info}(\text{D}) - \text{Info}_{\text{A}}(\text{D})$$

Expected information needed for classifying documents in D is calculated by,

$$nfo(A) = -\sum_{i=1}^{m} P_i log_2(P_i)$$

Where p_i is the probability of a document in D belongs to Class C_i. If feature 'A' has v number of distinct values (A belongs to $\{a_1,a_2,\ldots,a_v\}$), then actual information needed for classification by selecting a feature A is,

$$Info_A(D) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} * Info(D_j)$$

Where Dj is the set of documents in D that have value a_j for the feature 'A'.

IG is calculated for all the features. Feature with high IG minimizes the information needed for classification.

Multilayer Perceptron (MLP) Network

MLP is a feed forward artificial neural network, suitable for supervised learning. It uses multiple layers of nodes connected by a directed graph, and each link has some initial weight. Nodes in each layer are fully connected with nodes in the next layer. The three different layers are input layer, hidden layer and output layer. All the nodes except the nodes in the input layer use some activation function. Supervised learning is achieved by back propagation for training the nodes in the network. When the input is given to all the nodes of input layer it is propagated to the output layer. At the output layer, each node calculates the difference between the actual result and the expected result as error value. Then error values are propagated back to the hidden layer and link weights are adjusted based on some learning factor.

Proposed Neural Network

The proposed MLPNN is made up of two sub- neural networks. The sub-networks are termed upper and lower network. The upper network is made up of two hidden layers, with five neurons in each layer. The lower network is made up of one hidden layer with five neurons. The sigmoid transfer function is used in upper network and Tanh function in lower network. The different functions help to minimize the mutual interference during simultaneous processing and execution of task. The block diagram of the proposed neural network is shown in Figure 3. Table 1 gives the parameters of the proposed neural network.

Performance Metrics

The performance of the proposed classifiers is compared by the parameters such as classification accuracy, Precision, Recall and RMSE and are calculated as

$$\text{Pr}ecision = \frac{True \text{ positives}}{True \text{ positives} + false \text{ positives}}$$

$$\text{Re}call = \frac{True \text{ positives}}{True \text{ positives} + false \text{ negatives}}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{d}(y_i - y_i^{'})^2}{d}}$$

Table 1. Parameters for the Proposed Neural Network

Parameter	Values
Input Neuron	20
Output Neuron	5
Number of Hidden Layer - upper	2
Number of Hidden Layer - lower	1
Number of processing elements upper	5
Number of processing elements lower	5
Transfer function of hidden layer upper	Sigmoid
Transfer function of hidden layer lower	Tanh
Learning Rule of hidden layer	Momentum
Step size	0.1
Momentum	0.7
Transfer function of output layer	Tanh
Learning Rule of output layer	Momentum
Step size	0.1
Momentum	0.7

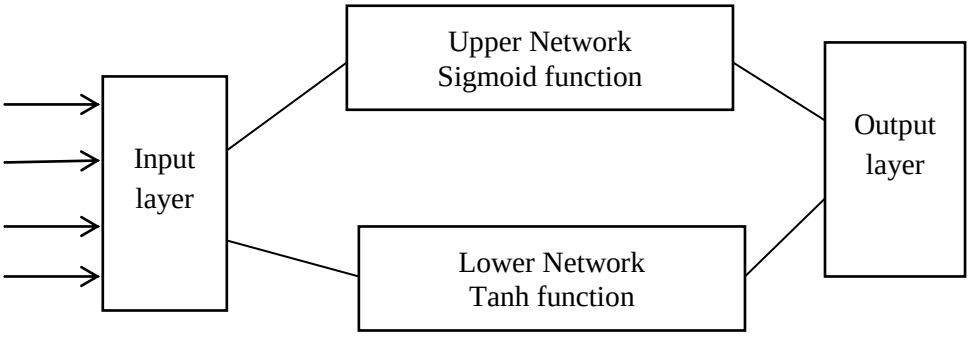

Figure 3. Block Diagram of Proposed Neural Network

4. Experiments and Results

Dataset

The audio dataset was collected for speaker identification to develop country contexts. It includes 83 unique voices, 35 female and 48 male. It provides audio for performing limited vocabulary speaker identification through digit utterances. Data was collected in partnership with Microsoft Research, India. Data was collected over telephone using an Interactive Voice Response (IVR) system in March, 2011. Participants are Indian nationals from differing backgrounds, each being given a few lines of digits, and asked to read numbers after being prompted in the system. Each participant read five lines of digits, one digit at a time. The numbers were read in English. There are various background noise levels, ranging from faint hisses to audible conversations/ songs. Totally, about 30% of the audio has some background noise. From these samples, a new neural network based classifier was trained. The performance of the proposed classifier is represented by the classification accuracy, Average Precision, Average Recall and Root Mean Squared Error (RMSE) and compared the performance with MLP classifier.

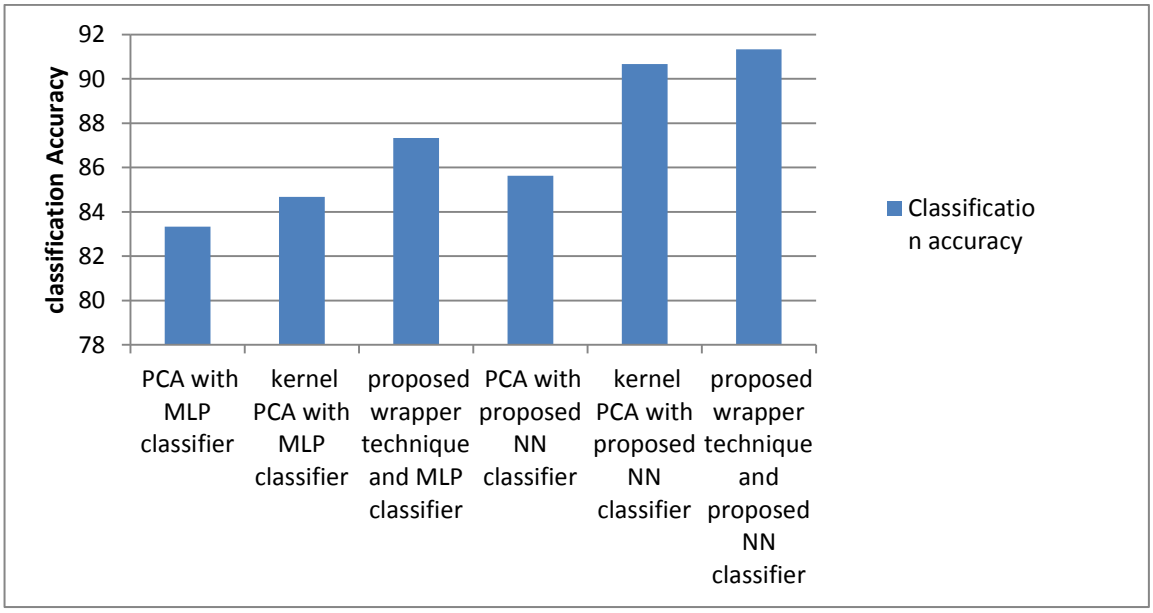

Figure 4. Classification Accuracy

Figure 4 shows the comparison classification accuracy of proposed NN with wrapper based feature extraction and selection with other MLP classifiers. It is observed that classification accuracy of proposed NN improves by 2.75 %, 7.09% and 4.58 % when comparing to feature selection by PCA with MLP, Kernel PCA with MLP and wrapper based selection with MLP respectively.

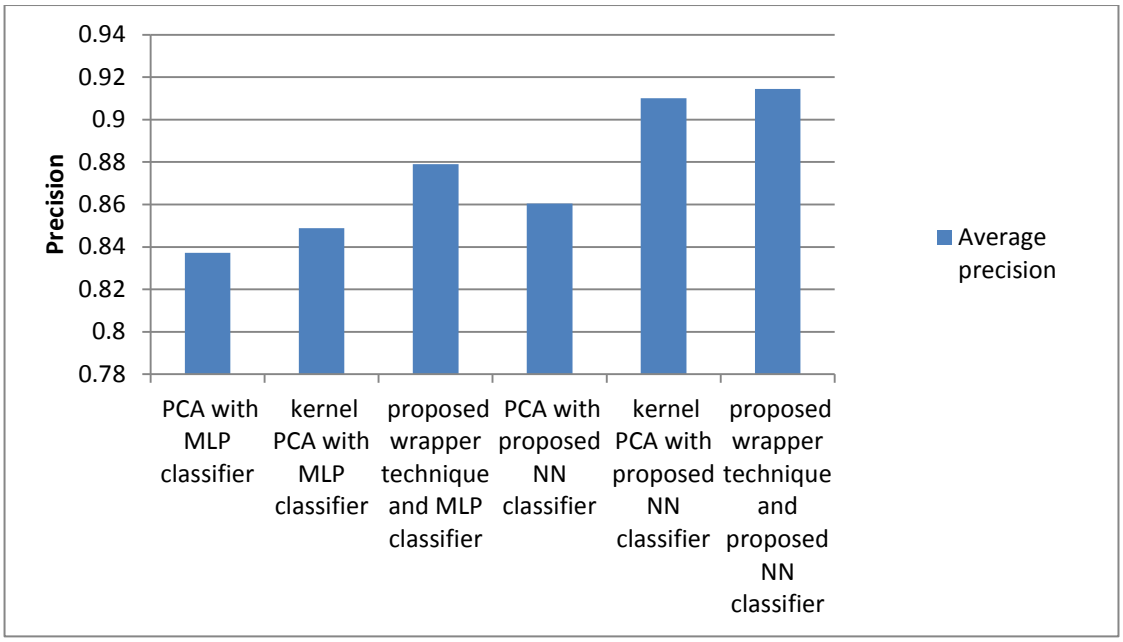

Figure 5. Average Precision

Figure 5 shows the comparison Average precision of proposed NN with wrapper based feature extraction and selection with other MLP classifiers. It is observed that Average precision of proposed NN improves by 2.78 %, 7.58% and 4.03 % when comparing to feature selection by PCA with MLP, Kernel PCA with MLP and wrapper based selection with MLP respectively.

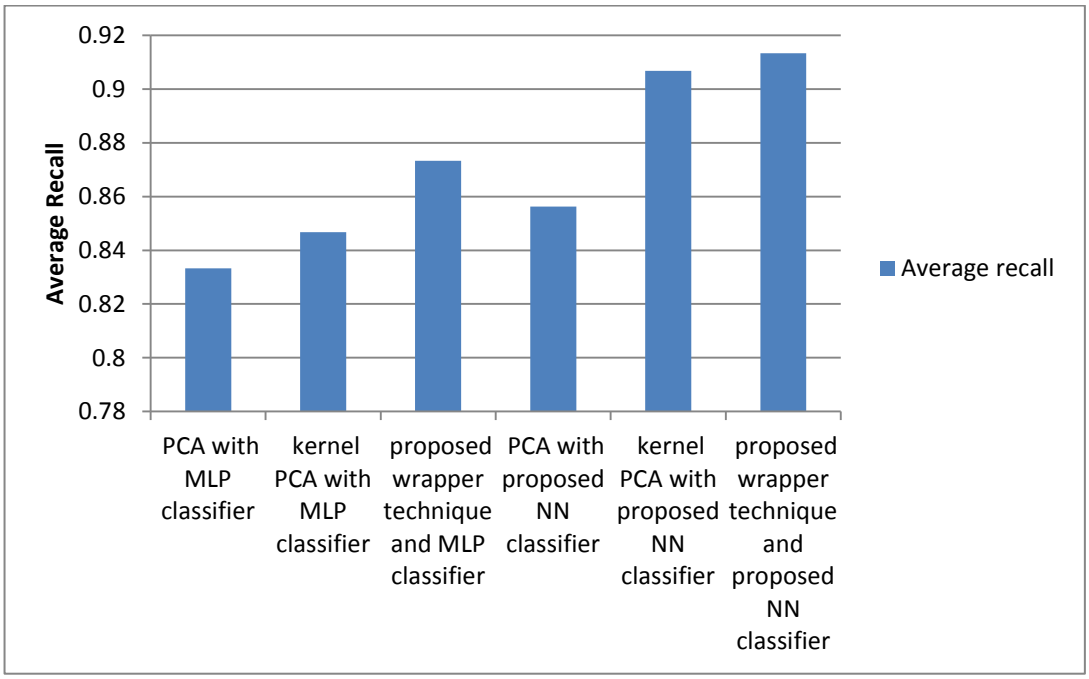

Figure 6. Average Recall

Figure 6 shows the comparison Average recall of proposed NN with wrapper based feature extraction and selection with other MLP classifiers. It is observed that Average recall of proposed NN improves by 2.76%, 7.09% and 4.58 % when comparing to feature

selection by PCA with MLP, Kernel PCA with MLP and wrapper based selection with MLP respectively.

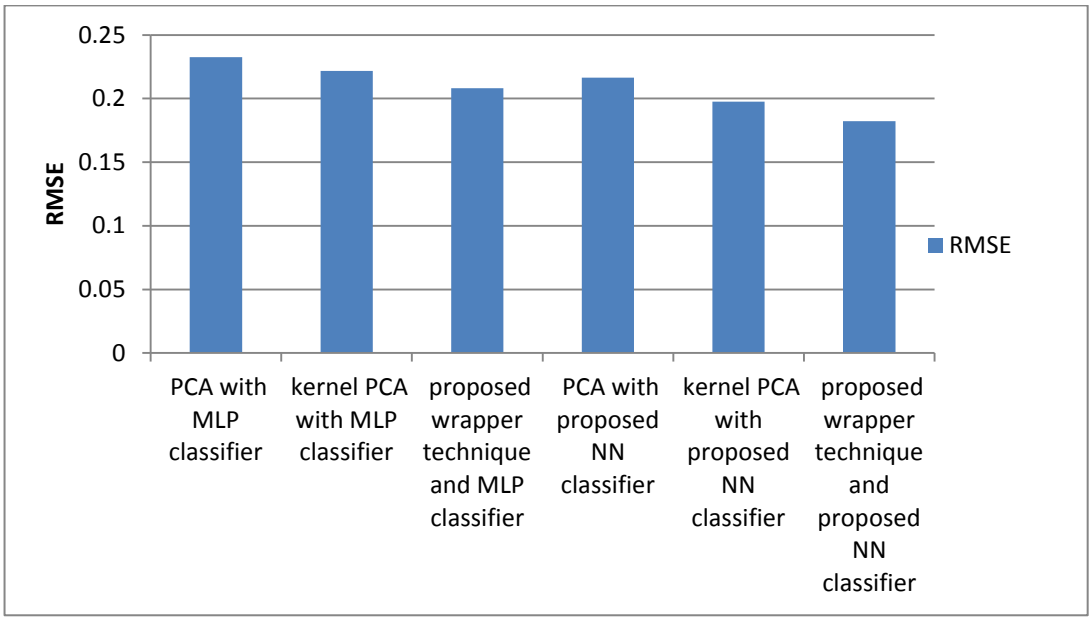

Figure 7. Root Mean Square Error

Figure 7 shows the comparison RMSE of proposed NN with wrapper based feature extraction and selection with other MLP classifiers. It is observed that RMSE of proposed NN decreases by 6.96%, 10.91% and 12.49 % when comparing to feature selection by PCA with MLP, Kernel PCA with MLP and wrapper based selection with MLP respectively.

5. Conclusion

In speech processing, the speakers are identified by recognizing the spoken phrases. From the input speech signals, features are extracted by WT and selected by wrapper based approach and a new NN is proposed for classification of sample into one of the known speakers list. Classification accuracy, Average Precision, Average Recall and RMSE are used to compare the performance with MLP classifier. This proposed approach improved classification accuracy up to 7.09 %, average recall up to 7.09 %, and average precision up to 7.58 % and reduced RMSE up to 12.49 %.

References

[1] S. Furui, "An Overview of Speaker Recognition Technology", Proceeding of Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, **(1994)**.
[2] S. Furui, "Recent Advances in Speaker Recognition", Pattern Recognition Letters, vol. 18, pp. 859-872, **(1997)**.
[3] B. S. Atal, "An Automatic Recognition of Speaker from their Voices", Proceedings of the IEEE, vol. 64, **(1976)**, pp. 460-475.
[4] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Mono-syllabic words Recognition in continuously spoken Utterances", IEEE transactions on speech and Audio Processing, vol. 28, **(1980)**, pp. 357-366.
[5] J. Mokhoul, "Linear Prediction - A Tutorial Review", Proceedings of IEEE, vol. 63, **(1975)**, pp. 561-580.
[6] M. Sifuzzaman, M. R. Islam and M. Z. Ali, "Application of Wavelet Transform and its Advantages Compared to Fourier Transform", Journal of Physical Sciences, vol. 13, **(2009)**, pp. 121-134.
[7] O.-W. Kwon, K. Chan and T.-W. Lee, "Speech Feature Analysis Using Variation Bayesian PCA", IEEE Signal Processing Letters, **(2002)**.

[8] S. Nakagawa, L. Wang and S. Ohtsuka, "Speaker Identification and Verification by combining MFCC and Phrase Information", IEEE transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, **(2012)** May.
[9] F. zohra Chelali, A. Djeradi and R. Djeradi, "Speaker Identification System based on PLP Coefficients and Artificial Neural Network", Proceedings of the World Congress on Engineering 2011, London, U.K., vol. II WCE **(2011)** July 6-8.
[10] P. Kral, "Discrete Wavelet Transform for Automatic Speaker Recognition", 3rd International Conference on Image and Signal Processing, vol. 7, **(2010)** October.
[11] M. I. Abdalla and H. S. Ali, "Wavelet Based Mel-Frequency Cepstral Coefficients for Speaker Identification using Hidden Markov Models", Journal of Telecommunications, vol. 1, no. 2, **(2010)** March.
[12] J. Wu -D. and B.-F. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition Expert Systems with Applications", vol. 36, **(2009)**, pp. 3136-3143.
[13] D. Avci, "An expert system for speaker identification using adaptive wavelet sure entropy", Expert Systems with Applications, vol. 36, **(2009)**, pp. 6295-6300.
[14] T. Abu Hilal, H. Abu Hilal, R. El Shalabi and K. Daqrouq, "Speaker Verification System Using Discrete Wavelet Transform And Formants Extraction Based On The Correlation Coefficient", Proceedings of the International Multi-Conference of Engineers and Computer Scientists, **(2011)**.
[15] M. Ziółko, R. Samborski, J. Gałka and B. Ziółko, "Wavelet-Fourier Analysis for Speaker Recognition", Zakopane–Ko´scielisko, **(2011)** September 1-6.
[16] M. Yamada, M. Sugiyama and T. Matsui, "Semi-supervised speaker identification under covariate shift", Signal Processing, vol. 90, no. 8, **(2010)**, pp. 2353-2361.
[17] H. B. Kekre and V. Kulkarni, "Speaker identification by using vector quantization", International Journal of Engineering Science and Technology, vol. 2, no. 5, **(2010)**, pp. 1325-1331.

Authors

V. Srinivas received the B.Tech & M.Tech from jntuk; Kakinada. He is having 15 years experience in teaching. His Research interests include speech processing, Speaker recognition.

Dr.Ch.Santhirani received PhD from JNTUH. She is having 20 years of experience in Teaching. Presently she is working as a professor in ECE department. She is actively involved in R&D activities in Wireless networks. Her area of interest is wireless communication.

Dr.T.Madhu received PhD from Osmania University. He is having 20 years of experience in teaching and administration. Presently he is working as a principal in SIET in Narsapur. His area of interest is navigational electronics and global position systems. He is actively involved in R&D activities in developing Global positions systems.

