International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.6, No.6 (2013), pp.193-204
http://dx.doi.org/10.14257/ijsip.2013.6.6.18

Investigation of Decision Tree Induction, Probabilistic Technique and SVM For Speaker Identification

V. Srinivas, Ch. Santhi rani and T. Madhu

[1] *Swarnandhra Institute of Engineering and Technology, Narasapur*
[2] *D.M.S&S.V.H Engineering College, Machilipatnam*
[3]*Swarnandhra Institute of Engineering and Technology, Narasapur*

srinivas.siet@gmail.com, santhirani.ece@gmail.com, principal.siet@gmail.com

Abstract

Speaker recognition and speech recognition are both related. As against determining what was said, speaker recognition enables the automatic recognition of who is speaking based on the speaker's voice's unique characteristics. This paper presents a simple approach to text dependent speaker identification and is based on the Symlet wavelets for feature extraction. The extracted features are then classified using data mining algorithms. In this study, J48, Naïve Bayes and SVM are used for classifying the features.

Keywords: *Speaker Recognition, Text dependent, Symlet Wavelet, Naïve Bayes, J48 and SVM*

1. Introduction

The advent of digital computers in the 1950s spurred modern speech recognition. Along with speech analysing and capturing tools like analog-to-digital converters and sound spectrograms, computers enabled researchers to locate feature extraction methods from speech which ensured intra-word discrimination. Automatic speech segmentation advanced into linguistically relevant units (like phonemes, syllables, words) and also into new pattern-matching/classification algorithms. These techniques have improved to a level where very high recognition rates are assured with commercial systems being available at nominal prices. At present, speech recognition is used in manufacturing units which require voice data entry or commands when the operator's hands are occupied. Speech recognition is also applied in medicine, where voice input accelerates routine report writing. Speech recognition enables users to control personal workstations or for remote interaction with applications when they lack touch tone key pads. Speaker identification makes possible non-intrusive monitoring with high accuracy conforming to security requirements. It also provides greater freedom to the physically challenged [1].

Speaker recognition and speech recognition are both related. As against determining what was said, speaker recognition enables the automatic recognition of who is speaking based on the speaker's voice's unique characteristics [2]. Deciding whether a particular speaker uttered something is verification and locating a person's identity from well-known speaker's set is identification. The common form of speaker recognition (text-independent) is not very accurate for huge speaker populations, but if spoken words are user constrained (text-dependent) and prevents speech quality from varying much, then this too is possible on a workstation. When a person talking has to be identified, speech signals must be processed and speaker variability measures are to be extracted instead of being analysed by segments

ISSN: 2005-4254 IJSIP

corresponding to phonemes or text pieces. Only one classification is made for speaker recognition, based on input test utterance. Though studies reveal that certain acoustical features work better in speaker identity prediction, few recognizers examine specific sounds due to problems in phone segmentation and identification.

Both automatic speaker verification and identification use a stored reference patterns (templates) database for N known speakers. Both use analysis and decision techniques. Verification is easier as it compares test pattern against a reference pattern involving a binary decision: Is there a good match against the claimed speaker's template? Error rate for speaker identification is higher as it requires selecting which of system known N voices match the test voice or "no match" if test voice differs from reference templates. Comparing test and reference utterances for speaker identity is easier for identical underlying texts, as in text-dependent speaker recognition. Cooperative speakers allow application of speaker recognition directly through using same words to train and test the system. This is possible in verification, whereas speaker identification usually needs text-independent methods. Higher text-independent method error rates mean the requirement of more speech data for training and testing. Automatic computer speaker recognition is an active research area from early 1960s and spectrogram for personal identification was introduced. Text-independent speaker recognition is a popular research area, especially for applications like forensic science, intelligence gathering, and passive voice circuit's surveillance. Free-text recognition cannot control conditions influencing system performance, including speech signal variability, distortions and communication channel noise. Recognition has multiple problems including unconstrained input speech, uncooperative speakers, and uncontrolled environmental parameters which make it necessary to focus on an individual's features and his/her unique speech characteristics [3].

Various approaches are available in the literature for speaker identification based on the Gaussian mixture model (GMM) [4] or kernel methods such as the support vector machine (SVM)[5, 6], Non-negative matrix factorization [7]. In this paper, wavelet feature extraction speaker recognition, based on the Symlet wavelets, is investigated. The extracted features are then classified using data mining algorithms. In this study, J48 and Naïve Bayes are used for classifying the features. The literature survey is presented in Section 2, Section 3 deals with the materials and methods used in this investigation, Section 4 details the experimental details and Section 5 concludes the paper.

2. Related Works

Kekre *et al.*, [8] presented a simple text dependent speaker identification approach, combining spectrograms and Discrete Cosine Transform (DCT). This is based on DCT use to locate similarities between free sample spectrograms. The spectrogram set forms the database for experiments and not raw speech samples. Performance is compared for different number of DCT coefficients when applied on entire spectrogram, when DCT is applied to spectrogram divided into blocks and when DCT is applied to a spectrogram Row Mean. It revealed that the mathematical computations required for DCT on Row Mean of spectrogram is drastically less compared to the other two methods with almost equal identification rate.

Shafik *et al.*, [9] presented a robust speaker identification procedure from degraded speech signals based on the Mel-frequency cepstral coefficients (MFCCs) for feature extraction from degraded speech signals and wavelet transforms of such signals. It is a known fact that MFCCs based speaker identification procedure is not robust when noise and telephone degradation are present. So degraded signals wavelet transform feature extraction adds speech features from signal approximation and detail components which in turn help in achieving

high identification rates. The proposed method uses Neural Networks to match features. Comparison between the proposed method and traditional MFCCs based feature extraction from noisy speech signals/telephone degraded speech signals with additive white Gaussian noise (AWGN) and colour noise reveals that the proposed method has better recognition rates computed at different degradation cases.

Li *et al.*, [10] presented an ear-based feature extraction algorithm where feature is based on a recently published time-frequency transform and modules set to simulate signal processing in the cochlea. The feature is applied to speaker identification to offset acoustic mismatch problems in training/testing. Usually acoustic models performance drops when trained in clean speech and tested on noisy speech. The proposed feature shows strong mismatched situation robustness. As experiments show, both MFCC and the proposed feature have near perfect performances in speaker identification, in clean testing conditions, but when input signal SNR drops to 6 dB, MFCC feature's average accuracy is only 41.2%, when the proposed feature still continues with an average accuracy of 88.3%.

Yamada *et al.*, [11] suggested a novel semi-supervised speaker identification method to alleviate non-stationary influence like session dependent variation, recording environment change, and physical conditions/emotions. Voice quality variants are expected to follow the covariate shift model, where voice feature distribution alone changes in training and test phases. The proposed method includes kernel logistic regression and cross validation weighted versions and can in theory be capable of mitigating covariate shift influence. Experiments show that through text-independent/dependent speaker identification simulations that the proposed method promises much with regard to voice quality variations.

Kekre *et al.*, [12] presented Vector Quantization method for Speaker Identification consisting of training and testing phases. Vector quantization (VQ) is used for feature extraction by two methods. In method A, codebooks generated from speech samples are converted into 16 dimensional vectors with an overlap of 4. In method B, speech samples generated codebooks are converted into 16 dimensional vectors without overlap. Test sample codebook is generated and compared with database stored reference samples codebooks for speaker identification. Results from both schemes when compared shows that method 2 provides slightly better results than method 1.

Zhao *et al.*, [13] proposed local spatio-temporal descriptors for visual based speaker recognition and representation. Spatiotemporal dynamic texture features of local binary patterns extracted from localized mouth regions describe motion information in utterances, which capture spatial/temporal transition characteristics. Structural edge map features are extracted from image frames to represent appearance characteristics. Combining dynamic texture and structural features has 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.

3. Methodology

An Automatic speaker identification system has 2 stages; feature extraction and classification as seen in Figure 1 operating in training and recognition modes. Both include a feature extraction step, sometimes referred to as the system's front end. Feature extractor converts digital speech signal into a numerical descriptor sequence called feature vector [14]. Features in this paper use Symlet wavelet for extraction.

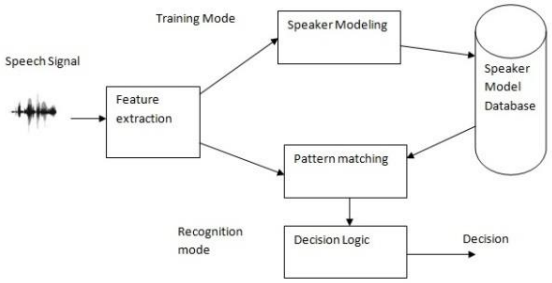

Figure 1. Automatic speaker identification system

For successful classification, every speaker is modelled using a data samples set in training mode, from where a feature vectors set is generated and saved in a database. Features are extracted from training data striping away unnecessary training speech samples information leaving only speaker characteristic information with which speaker models are constructed [14]. When a data sample from an unknown speaker arrives, pattern matching techniques map features from input speech sample to a model that corresponds to a known speaker.

Dataset

This audio data 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 [15]. Data was collected over telephone using an IVR (Interactive voice response) system in March, 2011, India. 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.

Feature Extraction

Symlet Wavelet

Feature extraction contributes to speaker identification based on low-level properties. Extraction produces enough information for speaker discrimination capturing this in a form/size that ensures efficient modelling. So feature extraction is defined as the process of reducing data present in a given speech sample while retaining speaker discriminative information at the same time. The Fourier transform (FT) includes fixed time-frequency resolution and a well-defined inverse transform. Fast algorithms exist for forward and inverse transforms which are simple and efficient computation algorithms, when applied to speech processing. Wavelets are time and frequency bound waveforms. Wavelet analysis splits mother wavelet signals into shifted and scaled versions. 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

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

Many wavelet coefficients C, a scale and position function are due to CWT. Original signals constituent wavelets are obtained by multiplying coefficients by applicable scaled and shifted wavelets. Daubechies Proposed Symlet-symmetrical wavelets - by modifying indications of the db family [16]. Both wavelet families are similar, with difference of db wavelets having maximal phase while Symlets have minimal phase. They are compactly supported wavelets with slight asymmetry with wavelet coefficient for it being any positive even number/highest number of vanishing moments for a support width.

Principal Component Analysis(PCA)

PCA is an established feature extraction technique for dimensionality reduction based on the assumption that most class information is in directions along which the variations are the largest. These directions are principal components. A common PCA derivation in terms of a standardized linear projection maximizing variance is the projected space. PCA is useful for data compression, reducing dimensions number without information loss.PCA is used to reduce the dimension of the feature vector extracted [17].

Mel frequency cepstral coefficients (MFCC)

Mel frequency cepstral coefficients (MFCC) are the most extensively used technique for both speech and speaker recognition. A Mel is a unit of measure which is based on the human ear's perceived frequency. The Mel scale consists of linear frequency spacing approximately below 1000 Hz. The approximation of Mel can be represented as shown below from frequency:

Mel (f) =2595*log (1+f/100)

Where f is the real frequency and Mel (f) is the perceived frequency.

The MFCCs can be obtained as follows:

- By taking the Fourier transform of a signal.
- Mapping the powers of the spectrum obtained onto the Mel scale, by using triangular overlapping windows.
- Taking the logs of powers at each of the Mel frequencies.
- Taking the discrete cosine transform for the list of Mel log powers.
- The MFCCs are the amplitudes of the Spectrum [18].

Classifiers

Classification in automatic speaker identification systems is a feature matching process between new speaker features and those saved in the database.

Naive Bayes

Given an objects set of known class and with a known variables vector, the aim is rule construction enabling assigning future objects to a class, if only variables vectors are given describing future objects. Problems of supervised classification are ubiquitous, and methods for such rule construction were developed. Naïve Bayes classifier is a commonly used classifier, easy to build not needing complicated iterative parameter estimation schemes to be applicable for large data sets. Also, Naive Bayes model appeals due to its simplicity,

elegance, and robustness. Although, an old classification algorithm, it is still effective in its simple form with modifications being introduced, by statistical, data mining, machine learning, and pattern recognition communities to ensure better flexibility.

Attribute conditional probabilities in the predicted training data set class is estimated by Naïve Bayes classifier, classification being on the parameter training data's mean and variance. Inputs are represented by feature vector and classified to a likely class. Naïve Bayes classifier assumes independent features thereby simplifying learning. When inputs are represented by feature vector X and classes by C, Naïve Bayes predict class as follows:

$$P(X|C) = \Pi_{i=1}^{n} P(X_i|C)$$

Where $X=(X_1,\ldots, X_n)$ is the feature vector and C is a class.

J48

Decision tree structures organize classification schemes by visualizing the steps taken to arrive at the classification. Every decision tree begins with a root node, considered the "parent" of other nodes. Each tree node evaluates a data attribute and determines the path to follow. The decision test compares a value against some constant. Decision tree classification is done through routing from root node until arrival at a leaf node. Decision trees represent information from a machine learning algorithm, offering a fast way to express structures in data. The J48 algorithm has many options related to tree pruning. Many algorithms try to "prune", their results. Pruning produces fewer, easily interpreted results and can also be a tool to correct overfitting. The algorithm described above recursively classifies until every leaf is pure, ensuring that data has been categorized as close to perfect as possible ensuring maximum accuracy on training data. It could create excessive rules that describe particular data idiosyncrasies alone. When tested on new data, rules may not be effective. Pruning reduces model accuracy on training data as pruning employs various means to relax decision tree specificity, hopefully improving its test data performance. The overall concept is gradual generalization of a decision tree until it attains a flexibility and accuracy balance.

Support Vector Machine (SVM)

SVM minimizes the structural risk while learning stage. It mainly aims at decreasing the generalization error instead of directly minimizing learning error. Hence, SVM is able to perform well when it is applied to data outside the training set. In recent years SVM learning has been widely used in real world applications as it offers superior performance than that of other competing methods [21]. SVM can also be used in application of pattern classification and nonlinear regression. The SVM becomes popular one due to its attractive features and its promising empirical performance. The initiative of support vector machine is to construct a hyper plane because the decision surface in such a way that the margin of separation between positive and negative example are maximized. The SVM with input vector $\vec{x}$, and normal vector $\vec{w}$ to hyper plane, the output u is given by:

$$u = \vec{w}.\vec{x} - b$$

The separating hyper plane is the plane $u = 0$. The margin is obtained by:

$$m = \frac{1}{\|w\|_2}$$

Maximizing the margin is same as solving the ensuing optimization problem:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

Subject to $y_i = \left(\vec{w}.\vec{x} - b\right) \geq 1$

Where b is a bias variable and N is the number of training example. It follows that the margin corresponds to the quantity $1/\|w\|$ and the maximization of margin is achieved by minimizing $\|w\|^2$

4. Results

The speech samples from the dataset were used for speaker identification. 50 samples were used for evaluating the classifiers. Examples of the speech input file given to a participant is as follows:

Line1: 265038971478190452362178963450013762589 48

Line2: 02154368

Line3: 6704352918719

Line4: 0635748219561047289

Line5: 7852934016275316948052843

The features from the samples are extracted using Symlet wavelets. The resulting features were reduced using PCA for efficient classification. The input sample and the output are shown in Figure 2 and 3 respectively.

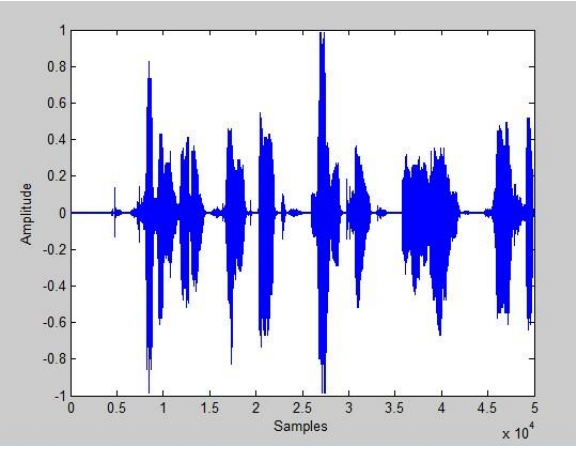

Figure 2. Input Speech

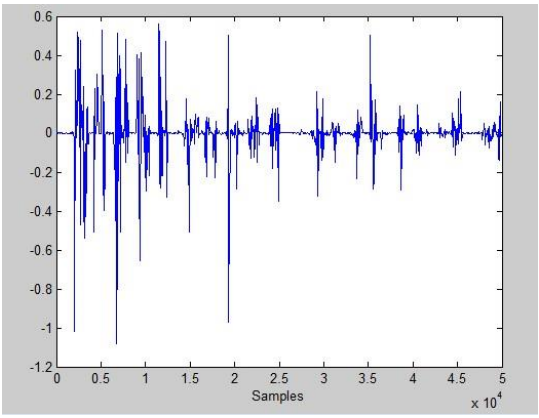

Figure 3. Output in Symlet Wavelet

The samples are classified using Naïve Bayes, j48 and SVM. The summary of results for classification accuracy is shown in Figure 4.

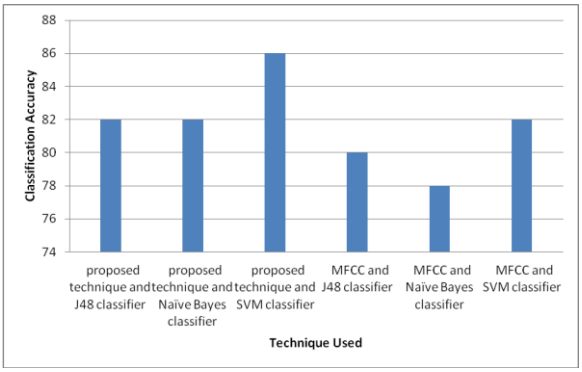

Figure 4. Classification accuracy of the proposed feature extraction technique

It is observed from Figure 4 that the classification accuracy achieved by both j48 and Naive Bayes is same at 82%. However, the classification accuracy of the SVM classifier with proposed feature extraction improves the classification accuracy and accuracy of 86% is achieved. The root mean squared error for j48 is slightly less than the Naive Bayes indicating better performance from j48 as shown in Figure5.

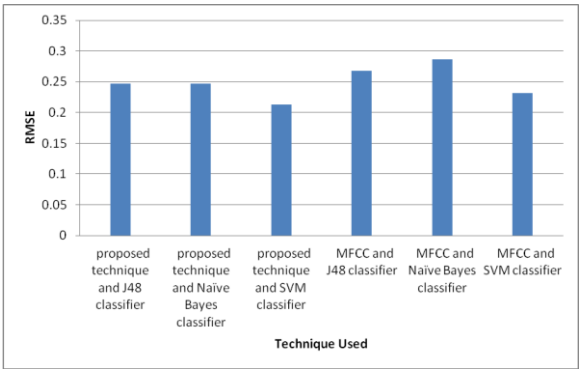

Figure 5. The Root Mean Squared Error

Table 1 gives the precision, recall and f-Measure by class for both the classifiers. Figure 4 and 5 shows the precision, recall and the f-Measure respectively

Table1. Precision and Recall

	Precision	Recall
Proposed technique and J48	0.832	0.82
Proposed technique and Naïve Bayes	0.836	0.82
Proposed technique and SVM	0.874	0.89
MFCC and J48	0.79	0.804
MFCC and Naïve Bayes	0.786	0.78
MFCC and SVM	0.82	0.82

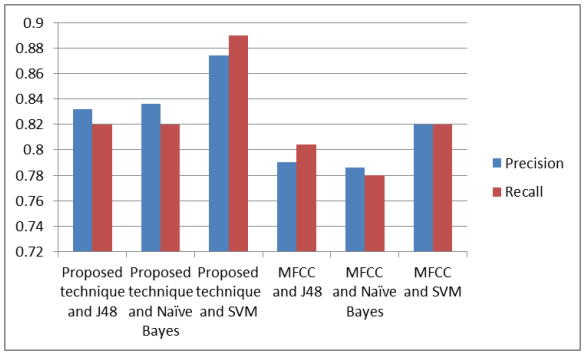

Figure 6: Precision and Recall

It is observed from the above graphs and table that though there is a minor variation of values of precision and recall for different classes; the weighted average of the precision and recall for both the classifiers are nearly same for the proposed method. Both the classifiers Naive Bayes and j48 perform equally well for classifying the speech samples and SVM achieves better performance. Further investigations are required to refine the feature extraction process and also to investigate the performance of soft computing methods for classification.

5. Conclusions

Speech Recognition research faces multiple problems such as unconstrained input speech, uncooperative speakers and uncontrolled environmental parameters which make it necessary to focus on an individual's features and his/her unique speech characteristics. In this paper, a wavelet feature extraction speaker recognition approach based on the Symlet wavelets is investigated. The extracted features are then classified using data mining algorithms, J48, Naïve Bayes and SVM. Experimental results show that classification accuracy of 86 % is achieved by the classifiers. Further investigations are required to improve the classifier efficiency.

References

[1] J. -S. Pan, Z. -M. Lu, and S. -H. Sun, "An Efficient Encoding Algorithm for Vector Quantization Based on SubvectorTechnique", IEEE Transactions on image processing, vol. 12, no. 3, **(2003)** March.
[2] T. Dutta, "Text dependent speaker identification based on spectrograms", Proceedings of Image and vision computing, New Zealand, **(2007)**, pp. 238-243.
[3] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing, vol. 10, no. 1, **(2000)**, pp. 19–41.
[4] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition", in Proceedings of the IEEE International Conference on Audio Speech and Signal Processing, Orland, Florida, USA, **(2002)**, pp. 161–164.
[5] J. Mariethoz and S. Bengio, "A kernel trick for sequences applied to text-independentspeaker verification systems", Pattern Recognition, vol. 40, no. 8, **(2007)**, pp. 2315–2324.
[6] B. Raj, T. Virtanen, S. Chaudhure and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition", In Proc. International Conference on Speech and Language Processing, **(2010)**.
[7] J. Barker, N. Ma, A. Coy and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition", Computer Speech & Language, vol. 24, no. 1, **(2010)**, pp. 94-111.
[8] H. B. Kekre, T. K. Sarode, S. J. Natu and P. J. Natu, "Performance Comparison Of 2-D DCT On Full/Block Spectrogram And 1-D DCT On Row Mean Of Spectrogram For Speaker Identification", International Journal of Biometrics and Bioinformatics (IJBB), vol. 4, no. 3, **(2010)**, pp. 100.
[9] A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam and F. A. El-samie, "A Wavelet Based Approach for Speaker Identification from Degraded Speech", International Journal of Communication Networks and Information Security (IJCNIS), vol. 1, no. 3, **(2011)**.
[10] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature", 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), IEEE, **(2010)** March, pp. 4514-4517.
[11] M. Yamada, M. Sugiyama and T. Matsui, "Semi-supervised speaker identification under covariate shift", Signal Processing, vol. 90, no. 8, **(2010)**, pp. 2353-2361.
[12] 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.
[13] G. Zhao, X. Huang, Y. Gizatdinova and M. Pietikäinen, "Combining dynamic texture and structural features for speaker identification", In Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence, **(2010)** October, pp. 93-98, ACM.
[14] D. Pullella, "Speaker Identification Using Higher Order Spectra", Dissertation of Bachelor of Electrical and Electronic Engineering, University of Western Australia, **(2006)**.
[15] A. Reda, S. Panjwani and E. Cutrell, "Hyke: A Low-cost Remote Attendance Tracking System for Developing Regions", The 5th ACM Workshop on Networked Systems for Developing Regions, NSDR, **(2011)**.
[16] V. U. Kale and N. N. Khalsa, "Performance Evaluation of Various Wavelets for Image Compression of Natural and Artificial Images", International Journal of Computer Science & Communication, vol. 1, no. 1, **(2010)** January-June, pp. 179-184.
[17] A. Malhi and R. X. Gao, "PCA-based feature selection scheme for machine defect classification", Instrumentation and Measurement, IEEE Transactions on, vol. 53, no. 6, **(2004)**, pp. 1517-1525.
[18] I. Rish, "An empirical study of the naive Bayes classifier", In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, **(2001)** August, pp. 41-46.

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 areas of interest are navigational electronics and global position system, Speech & Image Processing.