Multiclass SVM based Spoken Hindi Numerals Recognition

Teena Mittal[1] and Rajendra Kumar Sharma[2]
[1]Department of Electronics and Communication Engineering, Thapar University, India
[2]School of Mathematics and Computer Applications, Thapar University, India

Abstract: *This paper presents recognition of isolated Hindi numerals using multiclass Support Vector Machine (SVM). The acoustic features in terms of Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) and combination of LPC and MFCC have been considered as inputs to the recognition process. The extracted acoustic features are given as input to the SVM. The classification is performed in two steps. In first step, a one-versus-all SVM classifier is used to identify the Hindi language. Further, in second step ten one-versus-all classifiers are used to recognize numerals. The linear, polynomial and RBF kernels are used for the construction of SVM for recognition purpose. In the first phase, the best kernel strategy was explored for a fixed number of frames of the speech signal. The highest recognition rate has been achieved using linear kernel strategy. Next, the number of frames in order to calculate LPCs and MFCCs was varied and recognition accuracy was calculated. The highest recognition accuracy achieved in this study is 96.8%.*

Keywords: *LPC, MFCC, Hindi numerals, speech recognition, SVM.*

Received November 9, 2012; accepted March 9, 2014; published online August 16, 2015

1. Introduction

Automatic Speech Recognition (ASR) has been a very active research area from the past several decades. It is one of the fastest developing fields in the framework of speech science and engineering [21]. Speech recognition is a process to recognize words of some known language, captured by a microphone, or a telephone [19]. Many speech recognition systems have been developed for English and other European languages. However, for Hindi language, there have been certain attempts, with variable success rate, to create speech recognition systems [23, 34]. Chandrasekhar and Yegnanarayana [8] have proposed a model for recognition of Consonant-Vowel (CV) units of Hindi speech. Aggarwal and Dave [1] have developed a speech recognition system for ten Hindi digits using genetically optimized multilayer perception.

For ASR by computers, feature vectors are extracted from speech waveforms followed by a training phase. For the feature extraction phase, Linear Predictive Coding (LPC) and Mel-Frequency Cepstral Coefficient (MFCC) have extensively been used in literature. Paul *et al.* [18] have used LPC for Bangla speech recognition. Hai and Joo [14] have implemented a speech recognition system to recognize ten English digits by applying the improved LPC feature extraction method. Sanand and Umesh [24] have proposed a method to analytically obtain a linear-transformation on the conventional MFCC features. Aida-Zade *et al.* [2] have analyzed computing algorithms of speech features for the Azerbaijani language. They developed the determination algorithms of MFCC and LPC.

Support Vector Machine (SVM) is a promising machine learning technique that has generated a lot of interest in the pattern recognition community in recent years. This method has successfully been applied by several researchers in various fields like detection, verification, and recognition of faces, objects, handwritten characters, speech etc., [32]. Burges [5] has discussed the concepts of Vapnik Chervonenkis (VC) dimension and Structural Risk Minimization (SRM) of SVM. Gordan *et al.* [13] have worked on visual speech recognition network based on SVM and obtained a word recognition rate of 90.6%. Ganapathiraju *et al.* [11] have discussed the application of SVMs to large vocabulary speech recognition. Gangashetty *et al.* [12] have proposed a SVM based recognition system for CV utterances of speech. Campbell *et al.* [6] have applied SVM for speaker and language identification. Ramirez *et al.* [22] have presented the effectiveness of SVM learning concepts for robust speech end point detection. Solera-Urena *et al.* [28] have compared two SVM based approaches to speech recognition with the classical HMM system in noisy environments. Liu *et al.* [16] have proposed an improved hybrid SVM and duration distribution based hidden Markov model for robust continuous digital speech recognition. Sloin and Burshtein [27] have presented discriminative training algorithm that uses SVMs to improve the classification of discrete and continuous output probability HMMs. Manikandan and Venkataramani [17] have compared the performance of different decision-tree based SVM classifiers by applying optimum threshold pruning technique.

The approach described in this paper is for a speaker-dependent, isolated word recognizer for Hindi numerals. It uses LPC and MFCC features for front-end processing of speech signal and SVM for recognition.

This paper is organized as follows: Section 2 discusses the speech recognition system with data pre-processing, feature extraction and classification modules as subsections. Section 3 presents the experimental details of the proposed system. Section 4 contains the results and discussion and section 5 concludes the paper.

2. Speech Recognition System

Speech recognition systems have become one of the major applications for machine learning and pattern recognition. A speech recognition system consists of three modules: Data preprocessing, feature extraction and classification. A word is recorded through a microphone and stored in the form of wave files. These wave files are given as input to data pre-processing phase.

2.1. Data Pre-Processing

The block diagram of data pre-processing phase involves pre-emphasis, end point detection, framing and windowing as shown in Figure 1.

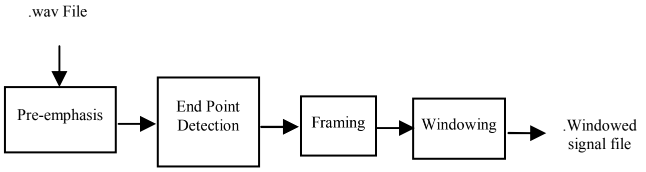

Figure 1. Block diagram of data pre-processing phase.

The process of spectrally flatten the digitized speech signal by passing it through a first order Finite Impulse Response (FIR) filter is known as pre-emphasis. In speech processing, it is very important to detect the voice region. So, end point detection is applied to remove the silence region before and after the voice region [26]. For this purpose, energy and zero crossing rates are calculated [10]. After removing the silence, the speech signal is divided into frames. Framing is the process of segmenting the speech samples into small frames of approximately 20 to 30ms. Framing enables the non-stationary speech signal to be segmented into quasi-stationary frames [30]. In addition, each frame overlaps its previous frame by a predefined size. After framing, windowing is done to each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Normally, the Hamming window is used for windowing as it introduces the least amount of distortion.

2.2 Feature Extraction

Feature extraction is one of the key dimensions of design in ASR [20]. For obtaining LPC, Linear prediction is used in which future values of a digital signal are estimated as a linear function of previous samples [21]. The basic idea behind LPC is that a given speech sample can be approximated as a linear combination of past speech samples [4] and a mean square prediction error is computed between the speech sample and linearly predicted samples. By minimizing this error, a unique set of predictor coefficients can be determined.

MFCC is a short-time analysis scheme, in which a signature of the acoustic signal spectrum is computed from a filter-bank with central frequencies projected uniformly on the Mel scale [20]. Mel frequencies are based on the known variation of the human ear's critical bandwidths with frequency filters spaced linearly at low frequencies (<1KHz) and logarithmically at high frequencies (> 1KHz) [32].

2.3 Classification Module

It is the most important component of the system. The feature vector extracted from the feature extraction module is here passed to the classification module for recognition. In this study, SVM has been used for recognition.

The SVM, developed by Vapnik [33], works on the principles of SRM [5]. In SVM, original input space is mapped into a higher dimensional feature space in which an optimal separating hyper plane is constructed on the basis of SRM to maximize the margin between two classes.

The key feature of SVM is the use of kernels [5, 9] that implicitly compute an inner product between the two data vectors in the high dimensional feature space. There are various kinds of kernels [25] available such as linear, polynomial and Radial Basis Function (RBF) given as [15]:

Linear Kernel:

$$K(x_i, x_j) = \left\langle x_i, x_j \right\rangle \tag{1}$$

Polynomial Kernel:

$$K(x_i, x_j) = \left(\gamma \left\langle x_i, x_j \right\rangle + r\right)^d, \gamma > 0 \tag{2}$$

RBF Kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0 \tag{3}$$

Where γ, r and d are kernel parameters.

Some of the kernel functions may not provide optimal SVM configuration. The choice of a kernel function often has a bearing on the results of analysis.

There are various methods to solve multi-class problem using SVM. One classifier is one-versus-one, which learns to classify one class from another class [28] and another is one-versus-all, which learns to classify one class from all other classes [17, 33]. If the number of classes is N then $N(N\text{-}1)/2$ one-versus-one classifiers are required as compared to N one-versus-all classifiers [3, 31, 35].

3. Experiments

In the present work, a speech recognition system is developed for Hindi language numerals using MATLAB. For this system, initially speech samples of numerals in Hindi and other three languages (Marathi, Punjabi and Bengali) are recorded and stored in .wav files. A database consisting of 20 utterances for each numeral in each language spoken by a single female speaker has been created in this experimentation. This database was recorded at 44.1KHz in room conditions. After that the data preprocessing is done as discussed in sub-section 2.1. In this work, dynamic size frames are used to overcome the disadvantage with fixed size frames which persists different length of speech signal [7, 30].

In the first phase of experiment, the speech signal is divided into a fixed number of 25 frames with 50% superposition. As such, each signal (numeral) is divided into 25 frames. For each of the 25 frames, LPC and MFCC features are extracted for all numerals and 12 coefficients per frame are extracted resulting into 300 coefficients.

In the second phase, to observe the influence of the choice of number of frames on the recognition rate, a simulation study is carried out with different values of frames while keeping other parameters fixed. The speech signal is divided into varying number of frames starting from 20 up to 30 in a step of 2. Here, 12 coefficients per frame are extracted, resulting into 240, 264, 288, 312, 336 and 360 LPC coefficients and MFCCs for each number of frames, respectively.

The extracted acoustic features are given as input to the SVM. The classification is performed in two steps. In first step, a one-versus-all SVM classifier is used to identify the Hindi language. Further, in second step ten one-versus-all classifiers are used to recognize numerals. Figure 2 shows the training strategy of SVM for Hindi numerals. Each SVM classifier is separately trained by using LPC coefficients, MFCCs and combination of LPC and MFCCs.

In each case, the SVMs are trained by using several values of k in k-fold cross validation approach [29]. In this approach, *k*-1 folds are used for training and the remaining patterns are used for testing of the model for each numeral. In this work, k has been selected as 3, 4, 5, 7 and 10.

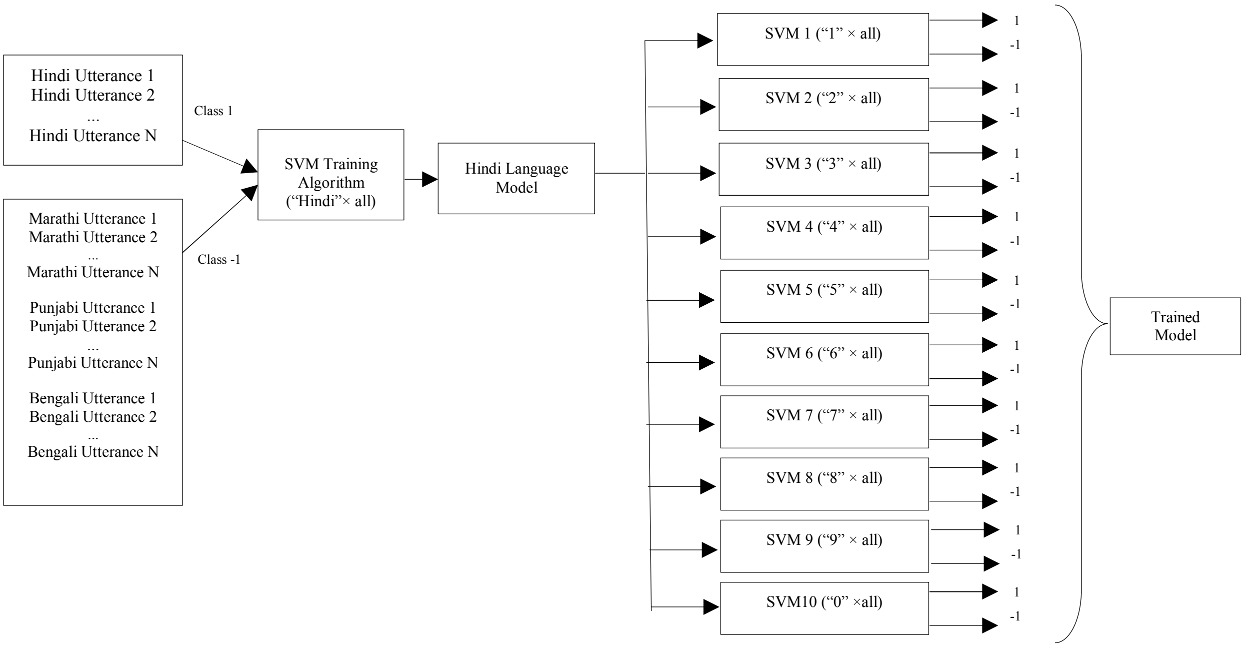

Figure 2. Training strategy of SVM for Hindi numerals.

4. Results and Discussion

Table 1 depicts the recognition rate for fixed number of frames by using linear, polynomial and RBF kernels. It can be observed that with LPC features as input and with linear kernel, a recognition rate of 74.0% is obtained with 10-fold cross validation. However, with MFCC features as input and SVM with linear kernel, a recognition rate of 92.5% is obtained with 10-fold cross validation. A recognition rate of 94.0% is obtained with 10-fold cross validation with combination of LPC and MFCC features as input and SVM with linear kernel. It is observed that for LPC and MFCC as well as combination of LPC and MFCC features, the best recognition rate is achieved with linear kernel. So, linear kernel has been selected for varying number of frames. The SVM performance is highly affected by kernel parameters. Normally, conventional gradient based search techniques are used to choose these parameters but sometimes solution converges to local optimum solution because of multimodal search area.

In the case of polynomial or RBF kernel, kernel-parameters (γ, r, d) need to be adjusted to get the reasonably good performance. The advantage of linear kernel is that no kernel parameter is required to set. So sometimes, linear kernel outperforms other kernel performances.

Table 1. Recognition rate obtained with different folds and kernels.

k-fold	Recognition Rate (%)								
	Linear Kernel			Polynomial Kernel			RBF Kernel		
	LPC	MFCC	Combination of LPC and MFCC	LPC	MFCC	Combination of LPC and MFCC	LPC	MFCC	Combination of LPC and MFCC
3	73.2	87.0	89.0	68.6	85.0	87.3	33.0	37.0	37.0
4	72.7	88.0	88.7	72.7	83.0	86.0	36.0	38.0	38.0
5	73.4	91.0	91.0	69.0	86.0	87.3	36.0	39.0	39.7
7	72.7	87.0	88.7	67.0	81.0	84.5	35.0	33.0	34.3
10	74.0	92.5	94.0	71.0	86.0	88.0	36.0	38.0	37.6

Figures 3 to 8 depict the effect of *k* in *k*-fold cross validation on the recognition rate when the number of frames is varied. It is observed that selection of appropriate value of frame is very important for better performance and it has been observed from Figure 5 that with LPC features as input, a recognition rate of 85.0% is obtained with 10-fold cross validation and 26 frames. However, with MFCC features as input, a recognition rate of 95.4% is obtained with 10-fold cross validation and 26 frames. The highest recognition rate of 96.8% is achieved, with combination of LPC and MFCC features used as input, while employing 10-fold cross validation and 26 frames.

It can also be observed from Figure 9 that the results obtained by MFCC are always better than LPC alone because MFCC are derived on the concept of logarithmically spaced filter bank, clubbed with the concept of human auditory system and hence, had the better response compare to LPC parameters. Another observation is that combination of MFCC and LPC features are better than those of results obtained with MFCC and LPC features on every execution that indicates the robustness of combined MFCC and LPC features.

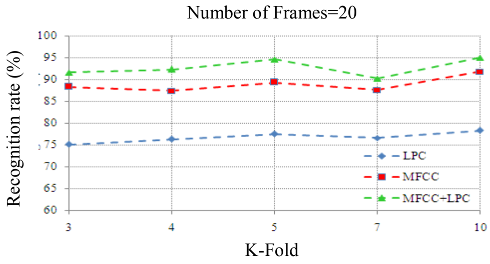

Figure 3. Recognition rate with 20 frames.

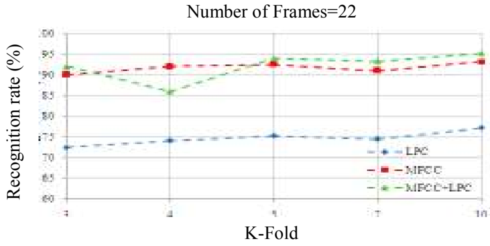

Figure 4. Recognition rate with 22 frames.

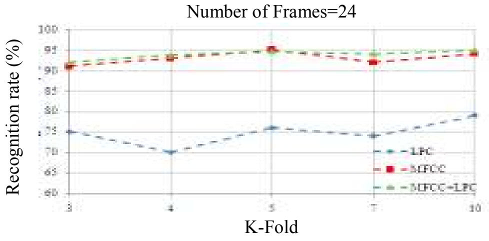

Figure 5. Recognition rate with 24 frames.

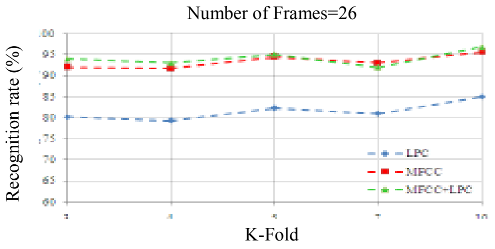

Figure 6. Recognition rate with 26 frames.

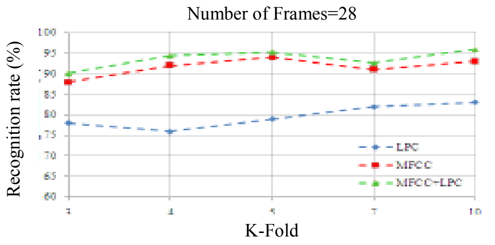

Figure 7. Recognition rate with 28 frames.

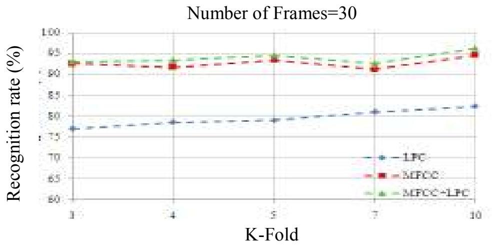

Figure 8. Recognition rate with 30 frames.

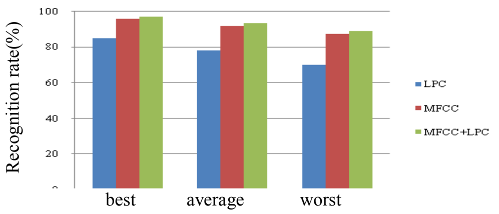

Figure 9. Comparison of best, average and worst results for three classes of features.

5. Conclusions

In this paper, a speaker-dependent, isolated word recognition system for Hindi numerals has been implemented. Features of speech in terms of LPC, MFCC and combination of LPC and MFCC are considered to recognize Hindi numerals. The SVM classification is performed in two steps. Initially, a one-versus-all SVM classifier is used to identify the Hindi language and then 10 one-versus-all classifiers are used to recognize Hindi numerals. The experiments have been carried out in two phases. In first phase, the number of frames has been fixed and different folds in *k*-fold cross validation have been applied for training and testing of SVM. To explore the best kernel strategy, linear, polynomial and RBF kernels have been used for the construction of SVM. The highest recognition rate of 94.0% has been achieved using linear kernel strategy with combination of LPC and MFCC features and 10-fold cross validation. The linear kernel strategy consistently dominated other kernel strategies in this phase of experiment. The linear kernel has advantage as compared to other kernels is that it does not require any kernel parameter to set.

In the second phase of experimentation, the number of frames has been varied to compute the recognition rate using SVM with linear kernel. The highest recognition rate of 96.8% could be achieved with combination of LPC and MFCC features, while employing 10-fold cross validation and 26 frames.

References

[1] Aggarwal K. and Dave M., "Application of Genetically Optimized Neural Networks for Hindi Speech Recognition System," *in Proceeding of World Congress on Information and Communication Technologies*, Mumbai, India, pp. 512-517, 2011.

[2] Aida-Zade K., Ardil C., and Rustamov S., "Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems," *International Journal of Signal Processing*, vol. 3, no. 1, pp. 105-111, 2006.

[3] Allwein E., Schapire R., and Singer Y., "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113-141, 2000.

[4] Atal B. and Rabiner L., "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 201-212, 1976.

[5] Burges C., "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[6] Campbell W., Campbell J., Reynolds D., Singer E., and Torres-Carrasquillo P., "Support Vector Machines for Speaker and Language Recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210-229, 2006.

[7] Cerf P. and Compernolle D., "A New Variable Frame Rate Analysis Method for Speech Recognition," *IEEE Signal Processing Letters*, vol. 1, no. 12, pp. 185-187, 1994.

[8] Chandrasekhar C. and Yegnanarayana B., "A Constraint Satisfaction Model for Recognition of Stop Consonant–Vowel (SCV) Utterances," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, pp. 472-480, 2002.

[9] Cristianini N. and Taylor J., *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.

[10] Faycal Y. and Messaoud B., "Comparative Performance Study of Several Features for Voiced/Non-Voiced Classification," *the International Arab Journal of Information and Technology*, vol. 11, no. 3, pp. 293-299, 2014.

[11] Ganapathiraju A., Hamaker J., and Picone J., "Applications of Support Vector Machines to Speech Recognition," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2348-2355, 2004.

[12] Gangashetty S., Sekhar C., and Yegnanarayana B., "Acoustic Model Combination for Recognition of Speech in Multiple Languages using Support Vector Machines," *in Proceeding of IEEE International Joint Conference on Neural Networks*, pp. 3065-3069, 2004.

[13] Gordan M., Kotropoulos C., and Pitas I., "Application of Support Vector Machines Classifiers to Visual Speech Recognition," *in Proceedings of IEEE International Conference on ICIP*, pp. 129-132, 2002.

[14] Hai J. and Joo M., "Improved Linear Predictive Coding Method for Speech Recognition," *in Proceedings of IEEE International Conference on Information, Communication and Signal Processing*, pp. 1614-1618, 2003.

[15] Hwang D. and Kim D., "Near-Boundary Data Selection for Fast Support Vector Machines," *Malaysian Journal of Computer Science*, vol. 25, no. 1, pp. 23-37, 2012.

[16] Liu J., Wang Z., and Xiao X., "A Hybrid SVM/DDBHMM Decision Fusion Modeling for Robust Continuous Digital Speech Recognition," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 912-920, 2007.

[17] Manikandan J. and Venkataramani B., "Evaluation of Multiclass Support Vector Machine Classifiers using Optimum Threshold-based Pruning Technique," *IET Signal Processing*, vol. 5, no. 5, pp. 506-513, 2011.

[18] Paul A., Das D., and Kamal M., "Bangla Speech Recognition System using LPC and ANN," *in Proceedings of the 7th International Conference on Advances in Pattern Recognition*, Kolkata, India, pp. 171-174, 2009.

[19] Peacocke R. and Graf D., "An Introduction to Speech and Speaker Recognition," *Computer*, vol. 23, no. 8, pp. 26-33, 1990.

[20] Quatieri T., *Discrete-Time Speech Signal Processing Principles and Practice*, Prentice Hall, 2002.

[21] Rabiner L. and Juang B., *Fundamentals of Speech Recognition*, Pearson Education, 1993.

[22] Ramirez J., Yelamos P., Gorriz J., and Segura J., "SVM Based Speech End Point Detection using Contextual Speech Features," *IET Electronics Letters*, vol. 42, no. 7, pp. 426-428, 2006.

[23] Samudravijaya K., “Hindi Speech Recognition,” *Acoustic Society of India*, vol. 29, no. 1, pp. 385-393, 2001.

[24] Sanand D. and Umesh S., “VTLN using Analytically Determined Linear-Transformation on Conventional MFCC,” *IEEE Transaction on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1573-1584, 2012.

[25] Scholkopf B., Burges C., and Smola A., *Advances in Kernel Methods: Support Vector Machines*, Cambridge, MA: MIT Press, 1998.

[26] Shin J., Chang J., and Kim N., “Voice Activity Detection based on Statistical Models and Machine Learning Approaches,” *Computer Speech and Language*, vol. 24, no. 3, pp. 515-530, 2010.

[27] Sloin A. and Burshtein D., “Support Vector Machine Training for Improved Hidden Markov Modeling,” *IEEE Transaction on Signal Processing*, vol. 56, no. 1, pp. 172-188, 2008.

[28] Solera-Urena R., Martin-Iglesias D., Gallardo-Antolin A., Pelaez-Moreno C., and Diaz-de-Maria F., “Robust ASR using Support Vector Machines,” *Speech Communication*, vol. 49, no. 4, pp. 253-267, 2007.

[29] Stone M., “Cross Validation Choice and Assessment of Statistical Predictions,” *Journal of the Royal Statistical Society B*, vol. 36, no. 2, pp. 111-147, 1974.

[30] Tan Z. and Lindberg B., “Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798-807, 2010.

[31] Urena R., Moral A., Moreno C., Ramon M., and Maria F., “Real-time Robust Automatic Speech Recognition using Compact Support Vector Machines,” *IEEE Transaction on Audio, Speech and Language Processing*, vol. 20, no. 4, pp.1347-1361, 2012.

[32] Vapnik V., *Statistical Learning Theory*, Wiley NewYork, 1998.

[33] Vapnik V., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 2000.

[34] Verma A., Kumar M., and Rajput N., “A Large Vocabulary Continuous Speech Recognition System for Hindi,” *IBM Journal of Research and Development*, vol. 48, no. 5-6, pp. 703-715, 2004.

[35] Weston J. and Watkins C., “Support Vector Machines for Multi-Class Pattern Recognition,” *in Proceedings of European Symposium Artificial Neural Networks*, Belgium, pp. 219-224, 1999.

Rajendra Kumar Sharma received his PhD degree in 1993 from Indian Institute of Technology, Roorkee, India. Currently, he is Professor in School of Mathematics and Computer Applications, Thapar University, Patiala, India. He obtained his His research interests include soft computing, neural networks, and statistical methods in NLP.

Teena Mittal received her MS degree of engineering from M.I.T.S. Gwalior in 2006. Currently, she is pursuing her PhD degree from Thapar University, Patiala, India. Her research interests include natural language processing, speech recognition, and machine learning.