

A Hybrid HMM/SVM Classifier for Wavelet Front End Robust Automatic Speech Recognition

Rajeswari
Dept. of E&CE
Acharya Institute of Technology
Bangalore, India

N.N.S.S.R.K.Prasad
ADA, Ministry of Defence
Govt. of India
Bangalore, India

Sathyanarayana V
IEEE Graduate
Student Member
Bangalore, India

ABSTRACT

Noisy ambient conditions pose a challenge to speech recognition, increasing the acoustic confusability, thereby looking for powerful acoustic models to improve the generalization ability of the machine learning and improve the recognition accuracy. This paper discusses a hybrid classifier that harness the power of hidden markov models (HMM) and the discriminative support vector machines (SVM) applied to a wavelet front end based automatic speech recognition (ASR) system. The experiments are performed on speaker independent TIMIT database which are trained in a clean environment and later tested in the presence of additive white gaussian noise (AWGN) for various SNR levels using the HTK toolkit, SVMLib and SVMLight software tool. Experiments indicate that for large vocabulary the classification power of SVMs and the elegant iterative training algorithms for the estimation of HMMs together as a hybrid classifier with the wavelet front end performs better than the conventional classifiers.

General Terms

Pattern Recognition, Machine Learning.

Keywords

Hidden Markov Models, Support Vector Machines, Automatic Speech Recognition, Perceptual Wavelet Packets.

1. INTRODUCTION

The tremendous demand for voice interface applications, with the internet revolution has made ASR being used beyond the primary forms of man-machine communication [1]. Owing to the recent progress in automatic speech recognition (ASR) systems, they have been embedded in diversity of applications such as mobile devices, medical diagnosis, automotive vehicles, industry and military applications etc., which makes an ASR system to perform recognition under unknown conditions. The performance of an ASR system degrades substantially in the presence of a mismatch between the training and testing environments. This mismatch may be caused by different noises, microphone, channel characteristics etc., these conditions need to be dealt in order to have a recognizer deliver reliable results [1, 2]. In an ASR, classifier plays a vital role, especially embedded in noisy environments. HMMs have been the prevailing statistical modeling techniques for ASR [4]. Its success largely lies on HMMs simplicity, flexible modeling ability, and the efficient learning algorithms. The power of HMM lies in the fact that the parameters used to model the speech signals can be well optimized, and this

results in lower computational complexity in decoding procedure as well as improved recognition accuracy. Maximum likelihood (ML) is the optimization criterion which is typically the probabilistic definition of speech recognition. The expectation-maximization (EM) algorithm provides an iterative framework for ML estimation [4]. The ML based optimization improves the classifiers ability to represent a specific class but does not discriminate a class from the others [2, 4, 16].

However, it does not provide an adequate representation of speech due to some inaccurate assumptions about the speech production process, such as output independence assumption, assumption in modeling of state duration, and the markov assumption itself [16]. A lot of research has been put in the last decades using HMM modeling to improve the recognition accuracy by exploiting the power of HMMs [5, 6, 7]. Though the above approaches have had significant success in terms of recognition performance, their use gives opportunities to researchers to look into options of classifiers to be more discriminative.

Some alternative approaches to HMMs are based on artificial neural networks (ANN) and Fuzzy logic [8, 11]. ANNs represent an important class of discriminative techniques, very well suited for classification problems. They learn according to discriminative criteria. Although other classifiers like HMMs can be trained in a discriminative framework, ANN training is inherently discriminative and do not require strong assumptions about the underlying statistical properties of the input data and the functional form of the output density. On the contrary, HMMs usually assume that, successive acoustic vectors are uncorrelated and follow a gaussian (or mixture of gaussians) distribution. Despite the advantages of ANNs on static classification problems, they present notable limitations to deal with the classification of time sequences as is the case of speech signals [14].

To overcome these difficulties, several researchers have proposed the so-called Hybrid ANN/HMM-based ASR systems [9, 10], which combines the ability of HMMs to model the time variability of the speech signal and the discrimination ability provided by ANNs. However, hybrid ANN/HMM have not been yet widely applied to speech recognition, very likely because some problems still remain open, for example, the design of optimal network architectures or the difficulty of designing a joint training scheme for both, ANNs and HMMs [16].

The need for discriminative and classifiers with good generalization as well as convergence properties for speech recognition has paved way to look at a new tool that has

appeared in the field of machine learning. They have proved to be able to cope with hard classification problems in several fields of application, the support vector machines (SVMs). Support vector machine (SVM) is a new machine learning method proposed by Vapnik in 1995. This method based on dimension theory and structural risk minimization (SRM) can solve the problems associated with sample number, generalization ability and classification [12, 13]. The principal is that the input data are mapped into a higher dimensional space from a lower dimensional space by the kernel function. However selecting a good kernel function is also important for better classification [14, 17]. This paper discusses a hybrid classifier that harness the power of hidden markov models (HMM) and the discriminative support vector machines (SVM) applied to a wavelet front end based automatic speech recognition (ASR) system.

2. HMM FRAMEWORK FOR SPEECH RECOGNITION

A HMM is a stochastic finite state automation built from a finite set of possible states $Q = \{q_1, \dots, q_k\}$, by instantaneous transitions with certain probabilities between these states. Each of these states is associated with a specific emission probability distribution $P(x_n/q_k)$. Thus, HMMs can be used to model a sequence of feature-vectors X as a piecewise stationary process where each stationary segment is associated with a specific HMM state. This approach defines two concurrent stochastic processes, the sequence of HMM-states modeling the temporal dynamics of speech and a set of state output processes modeling the locally stationary property of the speech signal. In speech recognition it is to find the HMM model M^* which maximizes the posterior probability $P(M/X)$ of the hypothesized HMM model M , given a sequence X of feature-vectors. Since this probability cannot be computed directly, it is usually split using Bayes' rule into the acoustic model (likelihood) $P(X/M)$ and a prior $P(M)$ representing the language model [2, 4].

The general structure of a HMM is specified by the set of the five parameters given as (N, O, π, A, B) where N is the set of hidden states, O is the set of observation symbols, π being the initial state distribution, A is the state transition probability distribution and B being the observation symbol probability distribution. A HMM based recognizer is realized as follows [2, 4].

2.1 Probability Evaluation-Forward Backward Procedure

For a given observation sequence $O=(o_1,o_2,o_3,\dots,o_T)$ and a HMM model λ , the task here is to find the probability of the observation sequence i.e., $P(O/\lambda)$. For a given HMM in speech recognition the probability is given by,

$$P_{ij} \left(\frac{O_i}{\lambda_j} \right) = \text{Log} \sum_{i=1}^N \lambda_t(i) \quad (1)$$

Where $\lambda_t(i)$ is the forward component of the forward-backward procedure.

2.2 Optimal State Sequence-Viterbi Algorithm

The task here is to find the optimal sequence of the states to a given observation sequence and the model. This procedure is taken care by the viterbi algorithm. The aim of the algorithm

is to find the best state sequence, $q=(q_1,q_2,q_3,\dots,q_T)$, for the given observation sequence, $O=(o_1,o_2,o_3,\dots,o_T)$ and a HMM model λ .

2.3 HMM model Parameter Optimisation-Baum Welch Algorithm

For a given model $\lambda=(A, B, \pi)$, the algorithm aims at adapting the probability measures or the parameters so as to maximize the probability $P(O/\lambda)$, which is called as the training procedure. The training problem is the crucial task for speech recognition with HMM, since it will optimally adapt the model parameters to the observed training data.

3. SVM FRAMEWORK FOR SPEECH RECOGNITION

A SVM is inherently a binary nonlinear classifier capable of guessing whether an input vector x belongs to class 1 or to class 2 based on the value of the class label $y \in \{+1, -1\}$ i.e if input vector x belongs to class 1 when the desired output would be then $y = +1$ otherwise input vector x belongs to class 2 when the desired output would be $y = -1$. This algorithm was first proposed by Vapnik [12, 13] and it is a nonlinear version of a much older linear algorithm, the optimal hyperplane based generalized decision rule.

Given a set of separable data, the goal is to find the optimal decision function. It can be easily seen that there is an infinite number of optimal solutions for this problem, in the sense that they can separate the training samples with zero errors. However, since a decision function able to generalize for unseen samples is looked, an additional criterion to find the best solution among those with zero errors can be opted.

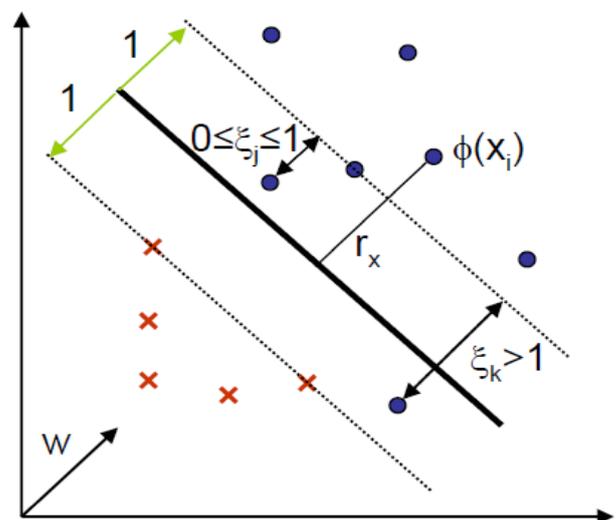


Fig 1: Soft Decision Margin

If probability densities of the classes are known, then the maximum a posteriori (MAP) criterion to find the optimal solution can be applied. Unfortunately, in most practical cases this information is not available, another simpler criterion is adopted i.e., among those functions without training errors, it is chosen that, with the maximum margin, being this margin the distance between the closest sample and the decision boundary defined by that function. Of

course, optimality in the sense of maximum margin does not imply necessarily optimality in the sense of minimizing the number of errors in test, but it is a simple criterion that yields to solutions which, in practice, turn out to be the best ones for many problems [12, 13, 14].

From figure 1, the non linear discriminate function is given as

$$f(X_i) = \mathbf{W}^T \cdot \Phi(X_i) + b \quad (2)$$

Where $\Phi(X_i)$ is a nonlinear function which maps input vector X_i into higher dimensionality space also called feature space where the classes are assumed to be linearly separable. The vector represents the separating hyperplane in such a space. Let r_x be the distance between the transformed sample $\Phi(X_i)$ and the separating hyperplane, and $\|W\|$ Euclidean norm of W . The vectors closest to the decision boundary called support vectors, these vectors define the margin and are the only samples that are needed to find the solution. For every sample X_i , has $r_x = \frac{f(X_i)}{\|W\|}$, hence optimum classifier is achieved by minimizing $\|W\|$ with the condition that all samples are being correctly classified, $y_i(\mathbf{W}^T \cdot \Phi(X_i) + b) \geq 1$. This can be formulated as a problem of quadratic optimization.

$$\min_{W,b} \|W\|^2 \text{ subject to } y_i(\mathbf{W}^T \cdot \Phi(X_i) + b) \geq 1 \quad (3)$$

In order to get a classifier with a better generalization ability and capable of handling the non-separable case, a number of misclassified data has to be allowed. This is accomplished by introducing a penalty term in the function to be minimized.

$$\min_{W,b,\xi_i} [L_P = 1/2 (\|W\|)^2 + \sum_{i=1}^N \xi_i] \quad (4)$$

subject to, $y_i(\mathbf{W}^T \cdot \Phi(X_i) + b) \geq 1 - \xi_i$, $\xi_i \geq 0$ for $i = 1, \dots, N$.

Where X_i be the training vectors corresponding to the labels $y_i \in \{\pm 1\}$ and the variables ξ_i are called slack variables. ξ_i verifies $0 \leq \xi_i \leq 1$ for those samples well classified but inside the margin, and $\xi_i > 1$ for those samples wrongly classified. The C term, on the other hand, expresses the trade-off between the number of training errors and the generalization capability. This problem is usually solved introducing the restrictions in the function to be optimized using Lagrange multipliers, leading to the maximization of the Wolfe dual.

$$\max_{\alpha_i} L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j \Phi^T(X_i) \Phi(X_j) \quad (5)$$

$$\text{subject to } \sum_{j=1}^n \alpha_j y_j = 0 \text{ and } 0 \leq \alpha_i \leq C \quad (6)$$

This problem is quadratic and convex, so its convergence to a global minimum is guaranteed using quadratic programming (QP) schemes. The resulting decision boundary W , given by

$$\mathbf{W} = \sum_{j=1}^n \alpha_j y_j \Phi(X_j) \quad (7)$$

The vectors with an associated $\alpha_i \neq 0$ will contribute to determine the weight vector \mathbf{W} and, therefore, the separating boundary. These are the support vectors that define the separation border and the margin. Generally, the function $\Phi(X_j)$ is explicitly unknown. However, it is not actually required to know it, since it only evaluates the dot products $\Phi^T(X_i) \cdot \Phi(X_j)$ which, by using what has been called the kernel trick, can be evaluated using a kernel function $K(x_i, x_j)$.

By using this method and replacing W in (2) by the equation in (7), the form that a SVM finally adopts is the following:

$$f(X_i) = \sum_{j=1}^n \alpha_j y_j K(x_i, x_j) + b \quad (8)$$

The most popularly used kernel functions are:

$$\text{Linear kernel: } K_L(x_i, x_j) = x_i^T \cdot x_j \quad (9)$$

Radial basis function kernel:

$$K_{\text{RBF}}(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (10)$$

which is proportional to the inverse of the variance of the gaussian function and whose associated feature space is of infinite dimensionality

$$\text{Polynomial kernel: } K_P(x_i, x_j) = (1 + x_i^T \cdot x_j)^P \quad (11)$$

whose associated feature space are polynomials up to grade P ,

Sigmoid kernel:

$$K_{\text{SIG}}(x_i, x_j) = \tanh(\alpha x_i^T \cdot x_j + b) \quad (12)$$

4. HYBRID HMM/SVM FRAMEWORK FOR SPEECH RECOGNITION

The HMMs are used to generate phonetic level alignments that are treated individually by the SVM to perform phoneme identification. Since each segment will have a different duration, some method is needed to convert them to fixed length vectors. This can be achieved by dividing the segment into three regions according to a pre-established proportion; thus, the vectors of the parameterized signal can be split into three groups according to a distribution of 30%-40%-30%. The vectors into every region are averaged and finally concatenated as depicted in figure 2.

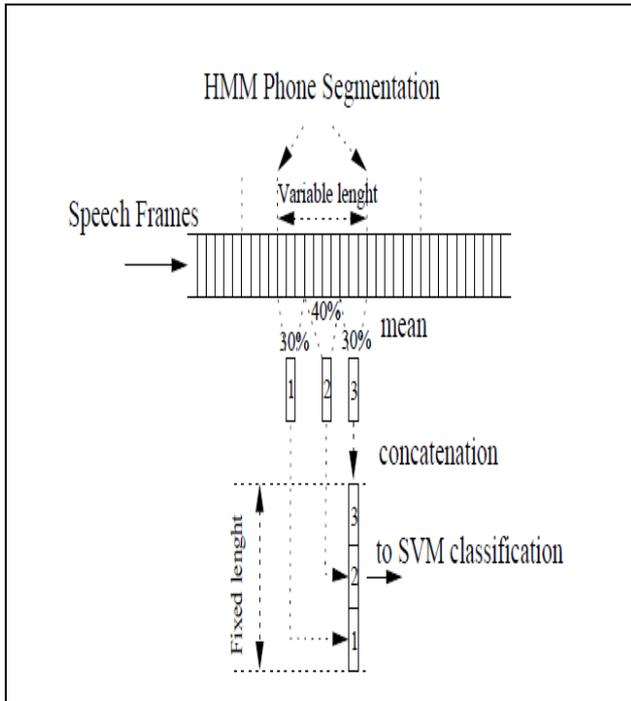


Fig 2: Segmental Features of SVM

Once having the SVM output the immediate probability is estimated [16, 25] as follows

$$P(y = 1/f) = \frac{1}{1 + \exp(Af + B)} \quad (13)$$

Where A and B are the parameters that needs to be estimated. The commonly used estimation method is Levenberg-Marquardt estimation method. The next procedure is to incorporate SVM classifier into standard Viterbi decoder available from ViterbiAlign function in HTK toolkit designed for HMM based system. This can be done as follows

- Train the SVM classifier for each Phone.
- N best list generation using conventional HMM system.
- Hypothesis of model level alignment in the N best list is generated using HMM system.
- Using SVMs, frames are classified based on the alignments.
- Posteriors are computed using sigmoid approximation.
- Utterance likelihood of each hypothesis in the N-best list is computed using these probabilities.
- The n best list is reordered based on the likelihood.

5. Wavelet Front End

The wavelet domain front end which can handle the non-stationary noise in case of speech enhancement and also suitable for robust features, detecting sudden bursts in speech signals [18, 19] has been used for the ASR in this paper.

5.1 Speech Enhancement

Denoising using wavelet packet (WP) coefficients is performed by thresholding. ie. the coefficients which fall below the specific value are shrunk and the latter retained. Different thresholding techniques have been proposed [19, 20, 21]. There are two popular thresholding functions used in the speech enhancement systems which are the hard and the soft thresholding functions. However, hard thresholding is best in preserving edges but worst in denoising while, soft thresholding is best in reducing noise but worst in preserving edges. The algorithm for denoising is as follows,

Step 1: Wavelet packet analysis

For a j level WP transform, the noisy speech signal with frame length N is decomposed into 2^j subbands. The m-th WP coefficient of the k-th subband is expressed as,

$$W_{K,m}^j = \text{WPT} \{x(n), j\} \quad (14)$$

Where $n=1, \dots, N$, $m=1, \dots, N/2^j$ and $k=1, \dots, 2^j$.

Step 2: Denoising by thresholding

In order to have a general case of both reducing noise as well as preserving edges a hybrid thresholding [22] is used.

Hybrid thresholding is given as

$$T_s(\lambda, w_k) = \begin{cases} w_k * \frac{|w_k|^{\alpha - \lambda}}{|w_k|^{\alpha}} & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \quad (15)$$

Where w_k represents wavelet coefficients and λ the threshold value. The threshold value or the shrinkage factor is found using the universal or the bayes threshold technique [19,20,21].

With careful tuning of parameter α for a particular signal, one can achieve best denoising effect within thresholding framework.

Step 3: Reconstruction

The enhanced speech is then reconstructed using the inverse wp transform

$$x'(n) = \text{WPT}^{-1} \{W_K^j, j\} \quad (16)$$

5.2 Feature Extraction

Wavelet coefficients provide flexible and efficient manipulation of a speech signal localized in the time–frequency plane [24]. The perceptual wavelet filter bank is built to approximate the critical band responses of the human ear. Wavelet packets decompose the data evenly into all bins but perceptual wavelet packets (PWP) decompose only critical bins. The decomposition is implemented and depicted in figure 3 to construct 24 filter banks. The size of the decomposition tree is directly related to the number of critical bins. The PWP transform is used to decompose $n_x(n)$ into several frequency bands that approximate the critical bands[23]. The PWP coefficients for the sub-bands are generated as follows:

$$w_j, i(k) = \text{pwpt}(nx(n)) \quad (17)$$

Where $n=1, 2, 3, \dots, L$ (L is the frame length),

$j=0, 1, 2, \dots, 6$ (j is the no. of levels)

$i=1, 2, 3, \dots, (2^j-1)$ (i is the subband index in each level of j)

The static PWP coefficients are made more robust by computing the delta and the acceleration coefficients.

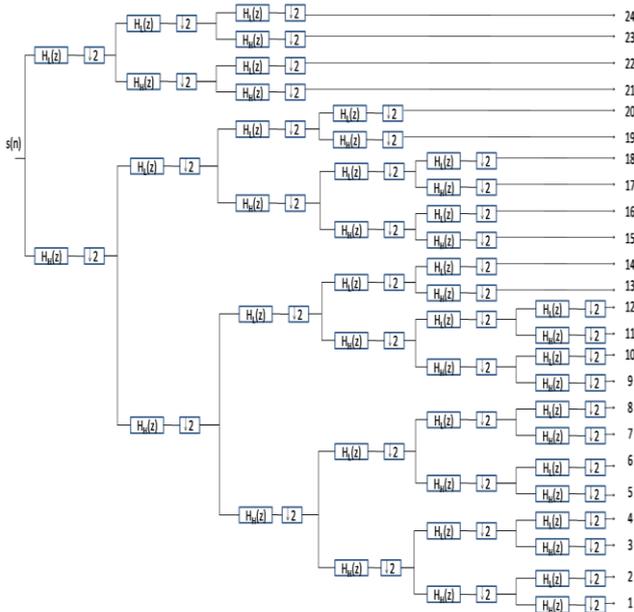


Fig 3: Tree Structure of PWPT

6. EXPERIMENTAL SETUP AND RESULTS

The evaluation of the speech recognition experiment is performed on the realization of isolated words extracted from TIMIT database. This set includes 50 isolated words uttered by 120 male and 120 female speakers. 80% database is used for training and 20% database is used for testing. The database is sampled at 16kHz under clean condition for the baseline experiment with the HMM classifier, a hamming window with a width of 20 ms is used and the feature vectors (consisting of 12 MFCC, the log energy, 12 delta-MFCC and the delta-log energy) are extracted once every 10 ms. Wavelet derived feature vectors are also applied as the same way as the MFCC feature vectors to test the performance of the system. In the testing phase the speech waveforms are corrupted with different levels of noise to test the robustness of the system. The speech signals are denoised using wavelet thresholding technique to further improve the robustness of the system.

The baseline HMM-based ASR system is an isolated-word, speaker-independent system, developed using the HTK package [3]. A five state HMM for each model, three gaussian mixtures per state with continuous observation densities are used. The SVM classifier is built using open source SVMlib and SVMlight package available from [15].

The SVM classifier is trained on the composite vectors with the ratio of 30-40-30. The recognition of the hybrid system is also done using these segment level composite vectors. The posterior probability from the SVMs are calculated using sigmoid approximation and using these probabilities the N-best list alignment is reordered and then the top hypothesis is constructed to recognize the word. The hybrid HMM/SVM classifier is evaluated with a wavelet front and a comparison is drawn between the HMM based ASR and the hybrid HMM/SVM ASR as shown in table 1 and figure 4.

TABLE 1. Comparison of recognition rates for hybrid HMM/SVM classifier for wavelet frontend ASR with HMM based recognizer.

SNR in dB	HMM based ASR (%)	HMM/SVM based ASR (%)
Clean	91.85	99.9321
5	20.0231	53.5326
10	39.1825	72.0432
15	53.8018	85.1232
20	66.5456	91.8831
25	78.2897	96.6120
30	82.8458	98.8800
35	86.6728	99.6922
40	90.4164	99.8451

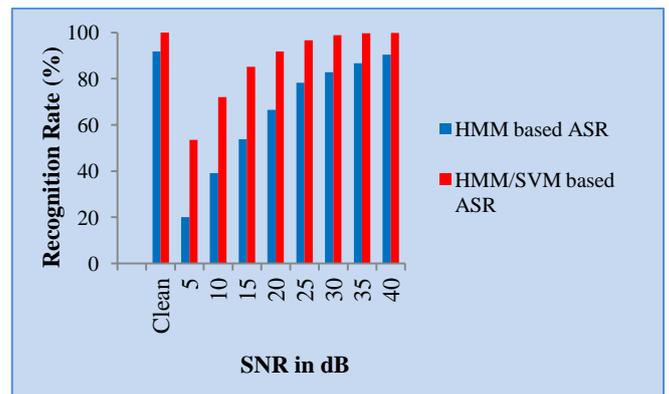


Fig 4: SNR versus recognition rates of HMM-ASR and HMM/SVM-ASR for timit database

7. CONCLUSIONS

In this paper, a hybrid HMM/SVM classifier and its evaluation with a wavelet front end based ASR has been described. The experimental results indicates that the elegant mechanism to model both the acoustic variability and the temporal evolution of speech by the HMMs along with the SVMs with good generalization and high dimensionality

input handling ability improves the robustness of ASR against noise. The wavelet domain front end which can handle the non-stationary noise in case of speech enhancement and also suitable for robust features, detecting sudden bursts in speech signals, aids in improving the robustness of the ASR system.

8. ACKNOWLEDGEMENTS

We would like to express our sincere thanks to Aeronautical Development Agency, (Ministry of Defence, Govt. of India), Bangalore, India for supporting to do our research work.

9. REFERENCES

- [1] Deng, L., Redmond, W.A., Li, X., 2013. Machine Learning Paradigms for Speech Recognition: An Overview. In IEEE Transactions on Audio, Signal and Language Processing, 1060-1089.
- [2] Rabiner, L. and Juang, B.H. 1993. Fundamentals of Speech Recognition, Prentice Hall.
- [3] Young, S. 2009. The HTK Book. Version 3.4, Cambridge University Engineering Department. Cambridge, UK.
- [4] Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Proceedings of the IEEE, 257-286.
- [5] Jiang, H., Li, X., Liu, C. 2006. Large Margin HMMs for Speech Recognition. In IEEE Transactions on Speech and Audio Language Processing, 14, 1584-1595.
- [6] Hoon, C. and Lee, S.Y. 2007. Noise Robust Speech Recognition Using Top-down selective Attention with an HMM Classifier. In IEEE Signal Processing Letters, 14, 489-491.
- [7] MIN, S. Y. and BAE, M. J. 2001. On a Study of Decreasing the Processing Time in the Speech Recognition System Using The HMM Algorithm. In IEEE Proceedings, 1224-1228.
- [8] Prasad, D. P. et al. 2001. Isolated Speech Recognition Using ANN, In Proceedings of the 23rd Annual EMBS International Conference, 1731-1734.
- [9] Cong, L. et al. 2001. Robust Speech Recognition Using Neural Networks and HMM's. In Proceedings of ICASSP.
- [10] Gemello, R., Albesano, D., Mana, F. 2000. CSELT Hybrid HMM/NN Technology for Continuous Speech Recognition. In IEEE International Conference on Neural Networks, 103-108.
- [11] Tsenov, G.T. and Mladenov, V.M. 2010. Speech Recognition using Neural Networks, Neural Network applications in Electrical Engineering, 181-186.
- [12] Vapnik, V. 1995. The Nature of Statistical Learning Theory, Springer-verlag, New York, USA.
- [13] Vapnik, V. N. 1998. Statistical Learning Theory, John Wiley & Sons, New York, USA.
- [14] Burges, C. J. C. 1999. A Tutorial on Support Vector Machines for Pattern Recognition. <http://svm.research.bell-labs.com/SVMdoc.html>, AT&T Bell Labs.
- [15] Joachims, T. 1999. SVMlight: Support Vector Machine, http://www.ai.informatik.uni-dortmund.de/FORSCHUNG/VERFAHREN/SVM_LIGHT/svm_light.eng.html, University of Dortmund.
- [16] Ganapathiraju, A., Hamaker, J. and Picone, J. 2000. Hybrid HMM/SVM Architectures for Speech Recognition, In Proceedings of the Department of Defense Hub 5 Workshop, College Park, Maryland, USA.
- [17] Wu, K.P. and Wang, S. D. 2009. Choosing the Kernel Parameters for Support Vector Machines by the Inter-Cluster Distance in the Feature Space, In IEEE Transactions on Pattern Recognition, 42, 710-717.
- [18] Rajeswari, Prasad, N.N.S.S.R.K. and Satyanarayana, V. 2012. Robust Speech Recognition using Wavelet Domain Front End and Hidden Markov Models. In Proceedings of International Conference on Emerging Research in Electronics, Computer science and Technology-ICERECT, Springer Lecture Notes-10.1007/978-81-322-1157-0_44
- [19] Chang, S., Yu, B., Vetterli, M. 2000. Adaptive Wavelet Thresholding for Image Denoising and Compression. In IEEE Transactions on Image Processing, 9, 1532-1546.
- [20] Donoho, D.L. and Johnstone, I.M. 1995. De-noising by Soft-thresholding. In IEEE Transactions on Information Theory, 41, 613-627.
- [21] Donoho, D.M. and Johnstone, I.M. 1995. Adapting to Unknown Smoothness via Wavelet Shrinkage. In Journal of the American Statistical Association, 90, 1200-1224.
- [22] Sumithra, M. G. and Thanuskodi, K. 2009. Wavelet based Speech Signal De-Noiseing using Hybrid Thresholding, In conference proceedings of the International conference on Control, Automation, Communication and Energy Conservation.
- [23] Jiang, H. et al. 2003. Feature Extraction Using wavelet Packet Strategy. In Proceedings of 42nd IEEE Conference on Decision and Control, 4517-4520.
- [24] Gupta, M. and Gilbert, A. 2002. Robust Speech Recognition using Wavelet Coefficient Features, In IEEE Transactions on Speech and Audio Processing, 445-448.
- [25] Florin, R. and Mititaru, D. 2010. A HMM/SVM Hybrid Method for Speaker Verification. In Proceedings of IEEE International Conference on Communication, 111-114.