Proceedings of International Conference on
"Emerging Research in Computing, Information, Communication and Applications"
ERCICA 2013
ISBN: 9789351071020

A Noise Robust Speech Recognition System Using Wavelet Front End and Support Vector Machines

Rajeswari[1,*], NNSSRK Prasad[2] and V. Satyanarayana[3]

[1]*Department of E & CE, Acharya Institute of Technology, Bangalore 560 107, India.*
[2]*ADA, Ministry of Defence, Govt. of India, Bangalore 560 017, India.*
[3]*IEEE Graduate Student Member, Bangalore, India.*
e-mail: rajeswari@acharya.ac.in; nnssrkprasad2007@gmail.com; satyaec49@gmail.com

Abstract. Recent works in speech recognition technology, classification techniques is focused on models, such as support vector machines (SVMs), in order to improve the generalization ability of the machine learning for noisy environments. However kernel function plays a vital role in the generalization ability of the SVMs. This paper address, the issue of noise robustness for an Automatic Speech Recognition (ASR) system focusing on a wavelet domain front end and SVM based classifier with different kernel functions. The proposed ASR has a front end that exploits the benefits of wavelet techniques for speech enhancement and feature extraction along with a comparison of different kernel functions for classification. The experiments are performed on speaker independent TIMIT database which are trained in a clean environment and later tested in the presence of AWGN for various Signal to Noise Ratio (SNR) levels. Experiments indicate that for large vocabulary the wavelet front end and the Radial Basis Function (RBF) kernel has more convergence area when compared to the polynomial kernel and the linear kernel for classification and robustness.

Keywords: Support vector machines, Automatic speech recognition, Kernel functions, Perceptual wavelet packet transform, Hidden markov models.

1. Introduction

In the state-of-art ASRs, the focus is either at the front end in enhancement and extracting robust features along with the back end building a robust classifier. Traditionally, the Linear Predictive Coding (LPC) and the Mel Frequency Cepstral Coefficients (MFCC) features are used for parameterization of speech, however, MFCC and LPC do not give a good representation of noisy speech, especially at low signal to noise ratio (SNR). Wavelet Transforms, the rapidly developed mathematical tool, with its flexible time-frequency resolution, is becoming the most widely used tool for the front end of the ASR. The recent approach of Wavelet Packets which segment the frequency axis and makes uniform translation in time is been proposed. Wavelet coefficients provide flexible and efficient manipulation of a speech signal localized in the time–frequency plane which is an alternative to MFCC [1]. 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 (PWPs) decompose only critical bins [2,13].

Hidden Markov Models (HMM) which use the probability distribution associated with each state in an HMM to model the temporal variability that occurs in speech via an underlying markov process are the most significantly used modeling techniques over the last decades for ASR [10–12]. The widespread use of these models lies in the availability of efficient parameter estimation procedures to maximize the likelihood (ML) of data given the model. Lot many approaches have been added to improve the estimation of HMM procedures. However, it demands for static knowledge about speech recognition in advance and is well suited for large training sample number. Artificial Neural Networks (ANN) also represents class of discriminative techniques for speech recognition. However, ANN have shortfall in case of network structure, its optimization along with demanding for large training samples for better generalization ability [9,10].

*Corresponding author

© Elsevier Publications 2013.

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. 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 [6–9]. The proposed ASR has a front end that exploits the benefits of wavelet techniques for speech enhancement and feature extraction along with a comparison of different kernel functions for classification.

The rest of the paper is organized as follows. Section 2 presents the theory of support vector machines. Section 3 describes the proposed wavelet front end and SVM based ASR. In section 4 experimental results are reported and analysed. Finally, the conclusion is drawn in section 5.

2. Support Vector Machines

2.1 *Principal*

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification, function estimation and density estimation [6]. The basic principal of SVM is, suppose some given data points each belong to one of two classes, and the goal is to decide which class a *new* data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a $(p-1)$-dimensional hyperplane. This is called a linear classifier.

Given some training data, a linear classifier $\{(x_i, y_i), i = 1, 2, \ldots, l\}$, where $x_i \in R^d$ and class label $y_i \in \{-1, 1\}$, d is the dimension of input x_i. A maximum margin hyperplane *i.e. the distance* $w.x_i + b = 0$ can be determined where $1/w$ is the distance between sample and hyperplane. An optimization can be put as minimizing w giving classification margin as maximum by the condition $y_i(w.x_i + b) \geq 1$ for $i = 1, 2, \ldots, l$.

The solution of the best separating hyperplane using Lagrange method in the dual form is given by the function:

$$\mathrm{Max}L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(X_i, X_i) \tag{1}$$

Subject to (for any $i - 1, 2, \ldots l$) and $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$.

Here α_i is the Lagrange multiplier corresponding to ith sample and C, the penality parameter for error classification, $C > 0$. When α_i is not equal to 0, the corresponding x_i is called the support vector. Now after finding the support vectors $_{xi}$ decision function is written as:

$$f(x) = \mathrm{sgn}\left[\sum_{i=1}^{l} \alpha_i^* y_i (x_i, x) + b^*\right] \tag{2}$$

where $\alpha_i^* (\alpha_i \neq 0)$ is the best value, b^* is the separating threshold value, x is the sample to be recognized.

2.2 *Kernel functions*

The original optimal hyperplane algorithm proposed by Vapnik in 1963 was a linear classifier. However, in 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik suggested a way to create nonlinear classifiers by applying the kernel trick (originally proposed by Aizerman *et al.*) to maximum-margin hyperplanes. Here the training data x is mapped into a higher dimensional linear feature space Z by a non linear function $\phi(.)$[6,8,14]. The decision function is given by:

$$f(x) = \mathrm{sgn}\left[\sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x) + b^*\right] \tag{3}$$

wherein, $K(x_i, x) = \varphi(x_i).\varphi(x)$, the dot product is replaced by a nonlinear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be nonlinear and the transformed space high dimensional; thus though the classifier is a hyperplane in the high-dimensional feature space, it may be nonlinear in the original input space.

© Elsevier Publications 2013.

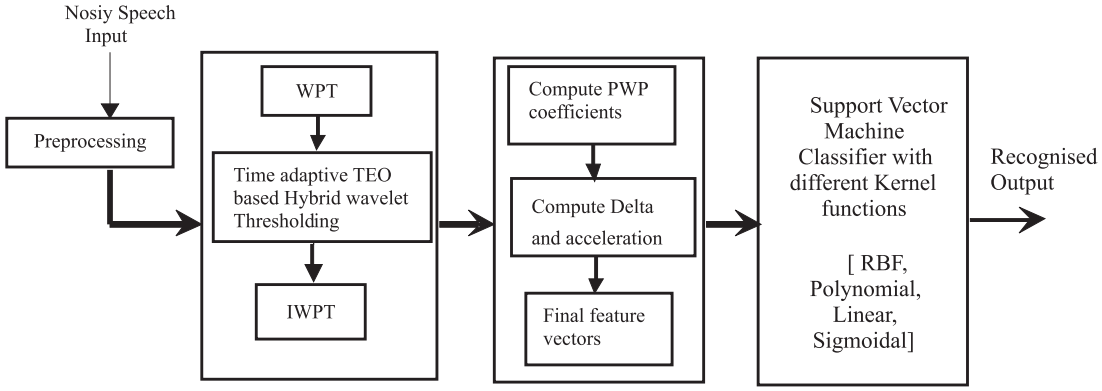

Figure 1. Block Diagram of the proposed ASR.

The kernel functions presently used are as follows:
Gaussian Radial Basis Function

$$K_{rbf}(x_i, x) = \exp(-\gamma \|x_i - x\|^2) \tag{4}$$

Polynomial kernel function

$$K_{poly}(x_i, x) = [(x_i.x) + 1]^q \tag{5}$$

Sigmoidal kernel function

$$K_s(x_i, x) = \tan h[g(x_i.x) + c] \tag{6}$$

The effectiveness of SVM depends on the selection of kernel, the kernel's parameters, and soft margin parameter C.

3. Proposed ASR with Wavelet Front End and SVM

3.1 *Speech enhancement*

In the proposed method as shown in figure 1 Wavelet Packet Transform (WPT) is applied to each input frame. The coefficients obtained are then subjected to Teager Energy approximation, where the threshold is adapted with respect to the voiced/unvoiced segments of the speech data. A Hybrid thresholding process is adopted which is a compromise for the conventional hard and soft thresholding in preserving both the edges and reducing the noise [3–5].

3.2 *Algorithm*

Step 1: Wavelet Packet analysis

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

$$W^J_{K,m} = \text{WPT}\{y(n), j\} \tag{7}$$

where $n = 1, \ldots N, m = 1, \ldots, N/2^j$ and $k = 1, \ldots 2^j$.

Step 2: Teager Energy Operator (TEO) on wavelet coefficients

Teager energy approximation for each WPT subband i is computed

$$\text{TEO}_{i,k} = Y^2_{i,k} - Y_{i,k-1} \tag{8}$$

- TEO coefficients are smoothened in order to reduce the sensitivity to noise

$$M_{i,k=}\text{TEO}_{i,k} * H_p \tag{9}$$

- Normalise the TEO coefficients

$$M'_{i,k=}\left[\frac{M_{i,k}}{\max(M_{i,k})}\right] \tag{10}$$

© Elsevier Publications 2013.

- Time scale adaptive threshold based on Bayes shrink for each subband k is computed

$$\lambda_{i,k=}\lambda_i(1 - M'_{i,k}) \tag{11}$$

Step 3: Denoising By Thresholding

Denoising using wavelet packet coefficients is performed by thresholding. i.e. the coefficients which fall below the specific value are shrunk and the later retained. Different thresholding techniques have been proposed. However, there are two popular thresholding functions used in the speech enhancement systems which are the hard and the soft thresholding functions.

Hard thresholding is given by

$$T_s(\lambda, w_k) = \begin{cases} w_k & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{12}$$

Soft thresholding is given by

$$T_s(\lambda, w_k) = \begin{cases} \text{sgn}(w_k)(|w_k| - \lambda) & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{13}$$

where w_k represents wavelet coefficients and λ the threshold value.

However, Hard thresholdingis best in preserving edges but worst in denoising while soft thresholding is best in reducing noise but worst in preserving edges. In order to have a general case of both reducing noise as well as preserving edges a hybrid thresholding is used.

Hybrid thresholding is given as

$$T_s(\lambda, w_k) = \begin{cases} w_k * \frac{|w_k|^\alpha - \lambda^\alpha}{|w_k|^\alpha} & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{14}$$

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

Step 4: Reconstruction

The enhanced speech is then reconstructed using the inverse wp transform

$$x'(n) = \text{WPT}^{-1}\{W'^{J}_{K}, j\} \tag{15}$$

3.3 *Dynamic perceptual wavelet packet feature extraction*

Wavelet coefficients provide flexible and efficient manipulation of a speech signal localized in the time–frequency plane [1]. 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 PWP decompose only critical bins [2]. The size of the decomposition tree is directly related to the number of critical bins. The decomposition is implemented by an efficient 7 level tree structure. The PWP transform is used to decompose $nx(n)$ into several frequency bands that approximate the critical bands. The PWP coefficients for the sub-bands are generated as follows:

$$wj, i(k) = pwpt(nx(n)) \tag{16}$$

where $n = 1, 2, 3, \ldots L$ (L is the frame length)

$j = 0, 1, 2, \ldots, 7$ (j is the no. of levels)
$i = 1, 2, 3, \ldots (2^j - 1)$ (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.

© Elsevier Publications 2013.

Table 1. Comparison of recognition rates based on SVM with different kernel functions for MFCC feature vectors.

Noise Level in dB	Linear Kernel (%)	Polynomial Kernel (%)	Gaussian Kernel (RBF) (%)
Clean	96	96.8	98.9
5	15.6923	15.6923	15.6923
10	18.962	22	24
15	38.1384	33.6983	34.1538
20	44.5416	47.5689	48.6154
25	67.5828	65.2358	67.5828
30	78.5698	82.5629	84.8889
35	89.81	93.5688	95.6581
40	95.1256	97.389	98.3162

Table 2. Comparison of recognition rates for proposed wavelet front end and SVM with different kernel functions.

Noise Level in dB	Linear Kernel (%)	Polynomial Kernel (%)	Gaussian Kernel (RBF) (%)
Clean	98.6532	99.7932	99.9216
5	25.0231	37.4269	53.4318
10	40.7818	52.2007	71.1911
15	55.8018	67.5592	84.0259
20	67.4054	79.8707	91.7821
25	75.5002	89.3813	96.4297
30	80.9788	94.6137	98.8612
35	86.6728	97.9378	99.6922
40	90.6741	99.0459	99.8461

4. Experimental Setup & Results

To evaluate the performance of the proposed method, recognition experiments were carried out using the 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 word segment used in this work were obtained by applying 64ms rectangular window to each waveform (of variable length) at its center which is at 16 kHz sampling frequency with 8 bit quantization. The speech signal is pre-emphasized by first order filter by coefficient 0.95 and multiplied by 20 ms hamming window having overlap of 10 ms. The feature vector is of 39 dimension, which includes 12d static mel frequency cepstral coefficient, log energy, delta and acceleration coefficients. Word recognition rate is obtained with multiclass SVM classifier. The performance of the SVM is tested with different kernel function in the presence of Additive White Gaussian Noise with different SNR levels.

5. Conclusion

The issue of noise robustness for an Automatic Speech Recognition (ASR) system focusing on a wavelet domain front end and SVM based classifier with different kernel functions is discussed. The proposed ASR has a front end that exploits the benefits of wavelet techniques for speech enhancement and feature extraction along with a comparison of different kernel functions for classification. The experimental results validates that Gaussian Radial Basis Function Kernel (RBF) along with wavelet fronted has a better recognition accuracy when compared with the other kernel functions as shown in table 2. Robustness of ASR can be addressed by exploiting both the features of HMM and SVM, hybrid classifiers with the wavelet front end as further work.

Acknowledgements

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

© Elsevier Publications 2013.

References

[1] Maya Gupta and Anna Gilbert: Robust speech recognition using wavelet coefficient features. *In IEEE Transactions on Speech and Audio Processing*, pp. 445–448, 2002.

[2] Hai Jiang *et al.*: Feature Extraction Using wavelet Packet Strategy. *In Proc. of 42nd IEEE Conference on Decision and Control*, pp. 4517–4520, December 2003.

[3] Donoho D. L. and Johnstone I. M.: De-noising by soft-thresholding, *In IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–27, 1995.

[4] Bahoura M. and Rouat J.: Wavelet speech enhancement based on the teager energy operator. *In Signal Process Lett. IEEE*, vol. 8, no. 1, 2001.

[5] Hesham Tolba: A Time-Space Adapted Wavelet de-noising Algorithm for Robust ASR in Low SNR Environments. *In IEEE Trans. On Speech and Audio Processing*, pp. 311–314, 2004.

[6] C. J. C. Burges: A tutorial on support vector machines for pattern recognition. *Knowledge discovery and Data Mining*, vol. 2, no. 2, pp. 121–167, 1998.

[7] V. N. Vapnik: statistical Learning Theory. John Wiley and Sons, New York, 1998.

[8] N. Cristianini and J. Shawe-Taylor: An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, 2000.

[9] A. Ganapathiraju: Support vector machines for speech recognition. Phd Thesis, Mississipi State University, 2002.

[10] Jing Bai *et al.*: Application of support vector machine with modified Gaussian kernel in a noise-robust speech recognition system. *In IEEE Transactions on Speech and Audio Processing*, pp. 502–505, 2008.

[11] L. Rabiner and B. H. Juang: Fundamentals of Speech Recognition. Prentice Hall Englewood Cliffs, New Jersy, vol. 103, 1996.

[12] D. O'Shaughnessy: Speech Communication Human and Machine. In IEEE Press, 2001.

[13] S. Mallat: A Wavelet Tour of Signal Processing. Academic Press, 2001.

[14] Jibran Yousafzai *et al.*: Combined Features and Kernel Design for noise Robust Phoneme Classification Using SVMs. *In IEEE Trans. On Speech and Audio Processing*, pp. 1396–1407, 2011.

© Elsevier Publications 2013.