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

Rajeswari¹,*, NNSSRK Prasad² and V. Satyanarayana³

¹Department of E & CE, Acharya Institute of Technology, Bangalore 560 107, India.
²ADA, Ministry of Defence, Govt. of India, Bangalore 560 017, India.
³IEEE Graduate Student Member, Bangalore, India.
e-mail: rajeswari@acharya.ac.in; nnssrkprasad2007@gmail.com; satyaece49@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 is 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 \(x_i \in \mathbb{R}^d\) and class label \(y_i \in \{-1, 1\}\), \(d\) is the dimension of input \(x_i\). A maximum margin hyperplane i.e. the distance \(w \cdot x_i + b = 0\) 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 \(y_i(w \cdot x_i + b) \geq 1\) 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:

\[
\text{Max} L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)
\]

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 \(\alpha_i\) is the Lagrange multiplier corresponding to \(i\)th sample and \(C\), the penalty parameter for error classification, \(C > 0\). When \(\alpha_i\) is not equal to 0, the corresponding \(x_i\) is called the support vector. Now after finding the support vectors \(x_i\), decision function is written as:

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

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:

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

wherein, \(K(x_i, x) = \phi(x_i) \cdot \phi(x)\), 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.
Figure 1. Block Diagram of the proposed ASR.

The kernel functions presently used are as follows:

Gaussian Radial Basis Function

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

Polynomial kernel function

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

Sigmoidal kernel function

$$ K_s(x_i, x) = \tanh[g(x_i \cdot x) + c] $$ (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,

$$ W_{k,m}^j = \text{WPT}\{y(n), j\} $$ (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

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

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

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

- Normalise the TEO coefficients

$$ M'_{i,k} = \frac{M_{i,k}}{\max(M_{i,k})} $$ (10)
Rajeswari, NNSSRK Prasad and V. Satyanarayana

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

$$\lambda_{i,k} = \lambda_i (1 - M_{i,k}^t)$$  \hspace{1cm} (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

$$T_s(\lambda, w_k) = \begin{cases} w_k & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases}$$  \hspace{1cm} (12)

Soft thresholding is given by

$$T_s(\lambda, w_k) = \begin{cases} \text{sgn}(w_k)(|w_k| - \lambda) & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases}$$  \hspace{1cm} (13)

where $w_k$ represents wavelet coefficients and $\lambda$ the threshold value.

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. 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

$$T_s(\lambda, w_k) = \begin{cases} w_k s \frac{|w_k|^\alpha - \lambda^\alpha}{|w_k|^\alpha} & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases}$$  \hspace{1cm} (14)

with careful tuning of parameter $\alpha$ for a particular signal, one can achieve best denoising effect within thresholding framework.

**Step 4: Reconstruction**

The enhanced speech is then reconstructed using the inverse wp transform

$$x'(n) = \text{WPT}^{-1}\{W_{K}^{J}, j\}$$  \hspace{1cm} (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:

$$w_{j,i}(k) = \text{pwp}(nx(n))$$  \hspace{1cm} (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.
Table 1. Comparison of recognition rates based on SVM with different kernel functions for MFCC feature vectors.

<table>
<thead>
<tr>
<th>Noise Level in dB</th>
<th>Linear Kernel (%)</th>
<th>Polynomial Kernel (%)</th>
<th>Gaussian Kernel (RBF) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>96</td>
<td>96.8</td>
<td>98.9</td>
</tr>
<tr>
<td>5</td>
<td>15.6923</td>
<td>15.6923</td>
<td>15.6923</td>
</tr>
<tr>
<td>10</td>
<td>18.962</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>38.1384</td>
<td>33.6983</td>
<td>34.1538</td>
</tr>
<tr>
<td>20</td>
<td>44.5416</td>
<td>47.5689</td>
<td>48.6154</td>
</tr>
<tr>
<td>25</td>
<td>67.5828</td>
<td>65.2358</td>
<td>67.5828</td>
</tr>
<tr>
<td>30</td>
<td>78.5698</td>
<td>82.5629</td>
<td>84.8899</td>
</tr>
<tr>
<td>35</td>
<td>89.81</td>
<td>93.5688</td>
<td>95.6581</td>
</tr>
<tr>
<td>40</td>
<td>95.1256</td>
<td>97.389</td>
<td>98.3162</td>
</tr>
</tbody>
</table>

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

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

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.
References


