Adaptive Smoothing and Wavelet Denoising for an Enhanced Speech Recognition System

Sonia Sunny\textsuperscript{1}, David Peter S\textsuperscript{2}, and K Poulouse Jacob\textsuperscript{3}
\textsuperscript{1}Dept. of Computer Science, Cochin University of Science & Technology, Kochi, Kerala, India
\textsuperscript{2} School of Engineering, Cochin University of Science & Technology, Kochi, Kerala, India
\textsuperscript{3}Dept. of Computer Science, Cochin University of Science & Technology, Kochi, Kerala, India

\textbf{Abstract} - Signals are corrupted by additive noise and removing noise from speech signals is one of the major challenges of an automatic speech recognition problem. In this paper, a speech recognition system is developed for recognizing speaker independent isolated words in Malayalam. Voice signals are sampled directly from the microphone and the background noise are reduced using wavelet denoising method based on Soft Thresholding (ST). Features are extracted using Discrete Wavelet Transforms (DWT). The extracted features are classified using Artificial Neural Networks (ANN), which produced a recognition accuracy of 88.5\%. In this paper, a new algorithm named Adaptive Smoothing Soft Thresholding (ASST) is proposed for smoothing the signals by reducing sudden spikes in the signal. When the signal after smoothing is used for recognition, the Signal to Noise Ratio (SNR) is improved and a better recognition accuracy of 91.3\% is obtained.

\textbf{Keywords:} Speech Enhancement; Discrete Wavelet Transforms; Soft Thresholding; Adaptive Smoothing; Artificial Neural Networks

1 \textbf{Introduction}

Recovering data from noise is an important area of research since it affects the recognition accuracy of a speech recognition system. Speech enhancement and noise suppression are of great importance because reducing the amount of noise is one of the main criteria for developing a good speech recognition system. The original signal should be recovered from noisy data retaining the important properties of the original signal [1]. Traditional denoising schemes are based on linear methods, where the most common choice is the Wiener filtering. Recently, nonlinear methods, especially those based on wavelets have become increasingly popular [2]. Degradation of signals by noise is a key problem in signal processing. Now a days, signal denoising has become an intensive field of study because of the increasing applications of speech enhancement in the areas like digital mobile, radio telephony systems, pay phones in noisy environments, teleconferencing systems, hearing aids etc. The most widely used method for measuring the performance of a speech recognition system is calculating the recognition accuracy. Though many parameters affect the performance of a speech recognition system, the presence of background or additive noise is one of the key factors.

Additive noise is a severe problem that greatly degrades the quality, clarity and intelligibility of the speech signals. In order to solve this, it is necessary to enhance the corrupted signal through an appropriate speech enhancement algorithm. Based on the nature and properties of the noise sources, noise can be classified into many types. Additive noises like background noise, impulse noise, speaker interfering noise and non additive noises like speaker stress, non-linearities of microphones etc. affect the quality of the speech produced. In this work, we have used wavelet denoising based on soft thresholding because it is proven that noise can be significantly reduced without reducing the edge sharpness. In soft thresholding, a threshold is estimated as a limit between the wavelet coefficients of the noise and those of the target signal [3]. But simple threshold can suppress the noise only up to an extent. So here a new algorithm is proposed in order to smooth the signal so that the magnitude of the sudden spikes in the signal are reduced. This in turn generates a signal which is smoother than the original signal. DWT and ANN are used for feature extraction and classification respectively.

The paper is organised as follows. Section 2 describes the statement of the problem. The Malayalam database used in the work is described in section 3. Section 4 gives a brief description of the architecture of the work which includes the pre-processing steps, feature extraction technique and the classification method. Section 5 presents the detailed analysis of the experiments done and the results obtained. Conclusions are given in the last section.

2 \textbf{Statement of the Problem}

The main goal of noise suppression is to improve the quality and intelligibility of the speech corrupted by background noise and to make speech recognition system more robust to input noise. This can be achieved using speech
enhancement algorithms [4]. Since speech enhancement is an intermediate stage in the implementation of a speech recognition system which in turn improves the recognition accuracy, research in this area is of great importance. Here, wavelet denoising is employed which is considered a non-parametric method [5]. In this work, we have used two methods. One using soft thresholding and the other using a new algorithm to smoothen the edges of the signal followed by the application of soft thresholding. Comparison is done in terms of SNR, spectrograms of the signals and waveform plots for ST and ASST. Feature extraction is done using DWT and classification is performed using the Multi Layer Perceptron architecture of ANN.

3 Words database

Ten isolated words from Malayalam language are chosen to create the database. 1000 speakers of age between 6 and 70 are selected to record the words. Each speaker utters 10 words. Thus the database consists of a total of 10000 utterances of the spoken words. This gives a moderate size for our study. We have recorded the speech from 400 male speakers, 400 female speakers and 200 children for creating the database. Male, female and voice of children differ in pitch, frequency, phonetics and many other factors. The samples stored in the database are recorded by using a high quality studio-recording microphone at a sampling rate of 8 KHz (4 KHz band limited). The spoken words, words in English, their International Phonetic Alphabet (IPA) format and translation in English are shown in Table 1.

Table 1. Words Stored in the Database and their IPA Format

<table>
<thead>
<tr>
<th>Words in Malayalam</th>
<th>Words in English</th>
<th>IPA format</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ഓണ</td>
<td>Onam</td>
<td>/O:nΛm/</td>
<td>Onam</td>
</tr>
<tr>
<td>ഈ</td>
<td>Chiri</td>
<td>/tʃiri/</td>
<td>Smile</td>
</tr>
<tr>
<td>വെഡ്ഡ്</td>
<td>Veedu</td>
<td>/vi:da/</td>
<td>House</td>
</tr>
<tr>
<td>കുടി</td>
<td>Kutti</td>
<td>/kuṭi/</td>
<td>Child</td>
</tr>
<tr>
<td>മരം</td>
<td>Maram</td>
<td>/mΛrΛm/</td>
<td>Tree</td>
</tr>
<tr>
<td>മയി</td>
<td>Mayil</td>
<td>/mΛjil/</td>
<td>Peacock</td>
</tr>
<tr>
<td>Lokam</td>
<td>/lokΛm/</td>
<td>World</td>
<td></td>
</tr>
<tr>
<td>Mounam</td>
<td>/maunΛm/</td>
<td>Silence</td>
<td></td>
</tr>
<tr>
<td>Vellam</td>
<td>/vellΛm/</td>
<td>Water</td>
<td></td>
</tr>
<tr>
<td>Amma</td>
<td>/ΛmmΛ/</td>
<td>Mother</td>
<td></td>
</tr>
</tbody>
</table>

4 System Architecture

In this paper, the speech recognition process is divided into three phases. In the first phase called pre-processing stage, the speech signals recorded are denoised using wavelets. The denoised signals are then given to the feature extraction stage where the relevant features are extracted. Finally, the extracted features are given for pattern classification. The different phases of this work are given below.

4.1 Pre-processing using wavelet Denoising

The speech signals are degraded by background noise. So pre-processing of speech signals is considered to be a crucial step in the development of a speech recognition system. There are different types of speech enhancement algorithms available like filtering techniques, spectral restoration techniques, model-based methods and wavelet based methods. Now more works are being done using wavelet denoising. Choice of the wavelet plays an important role in denoising. In this work, we have used the Daubechies wavelets which are the most popular wavelets that are found to be efficient in signal processing. In order to select an optimal wavelet function, the objective is to minimize reconstructed error variance and to maximize SNR. Wavelets with more vanishing moments are selected since it provides better reconstruction quality and introduce less distortion into processed speech. The speech signals are decomposed using DWT since it provides a time-frequency representation of the signal. Analysis and reconstruction of signals are done by the multi resolution filter banks and special wavelet filters of DWT [6].

4.1.1 Denoising using Soft Thresholding (ST)

The two popular thresholding functions used in denoising the signals using wavelets are the hard and the soft thresholding functions [7]. Hard thresholding sets to zero any element whose absolute value is lower than the threshold. Soft thresholding is an extension of hard thresholding. The elements whose absolute values are lower than the threshold...
are first set to zero and then shrinks the nonzero coefficients towards 0. Hard and soft thresholding can be defined as

\[
X_{\text{Hard}} = \begin{cases} 
X & \text{if } |X| > \tau \\
0 & \text{if } |X| \leq \tau 
\end{cases} \tag{1}
\]

\[
X_{\text{Soft}} = \begin{cases} 
sign(X) \cdot (|X| - |\tau|) & \text{if } |X| > \tau \\
0 & \text{if } |X| \leq \tau 
\end{cases} \tag{2}
\]

Where X represents the wavelet coefficients and \( \tau \) is the threshold value. In this paper, soft thresholding is used for wavelet denoising. There are different ways to calculate the Threshold value. In this work, the threshold used is the universal threshold developed by Donoho and Jonstone [8] which is defined as

\[
\tau = \sigma \sqrt{2 \log(N)} \tag{3}
\]

Where \( \sigma \) is the standard deviation and \( N \) is the length of the signal. Standard deviation \( \sigma \) can be calculated as \( \sigma = \text{MAD}/0.6745 \), where \( \text{MAD} \) is the median of the absolute value of the wavelet coefficients. We assume that Additive White Gaussian Noise (AWGN) is added. Then, any signal \( y(t) \) can be represented by the summation of the original \( x(t) \) and the noise \( n(t) \) as [9] [10].

\[
y(t) = x(t) + n(t) \tag{4}
\]

The outline of the de-noising algorithm can be represented using the following steps.
- Choose a wavelet and a level \( N \). Compute the wavelet decomposition of the signal at level \( N \).
- For each level from 1 to \( N \), select a threshold and apply soft thresholding to the detail coefficients.
- Compute wavelet reconstruction based on the original approximation coefficients of level \( N \) and the modified detail coefficients of levels from 1 to \( N \).

### 4.1.2 Denoising using Proposed Adaptive Smoothing Soft Thresholding Method (ASST)

Different wavelet denoising algorithms for speech signals were developed by modifying the standard threshold values available [11][12] and by combining different thresholding techniques [13]. Here, a new idea is developed to smoothen the signal before applying thresholding. Speech signals are often contaminated by sudden, abrupt noise that are represented in the form of spikes or troughs. The spikes/troughs distort the calculations and produce erroneous results. Elimination of such sudden variations has always been a challenging problem. In ordinary thresholding, only the top/bottom portions of a spike are cut out. But the steep gradient in the waveform will exist which actually accentuates the distortion. Attenuating a distortion without actually affecting the original waveform is of prime importance. In the proposed ASST method, previous values are compared with future values to determine the general trend of the signal and thereby facilitating suppression of random troughs. If this sudden spikes are reduced by smoothing, then automatically more noise components can be reduced and the original signal can be captured in its fullness. When this smoothened signal is given for denoising using ST, we get better results in terms of SNR value.

In the proposed ASST, the sign of the present value of the sample and the next value are compared. \( Y_i \) is compared with \( Y_{i+1} \). If both the values are in the same direction and in an increasing trend, the samples are reproduced in total and amplified by a smoothing factor less than 1, say 0.5 which decrease when the trend continues. When there is a reversal in trend, the factor that is added is kept high to capture the reversal in total. If \( Y_i \) and \( Y_{i+1} \) are in opposite directions, or in other words if there is a sign change in magnitude, we apply a dominant factor limiting the fall. If the trend continues, the signal is again reproduced in total plus the factor.

### 4.2 Feature Extraction using DWT

The denoised signals obtained after preprocessing are then applied to the feature extraction phase to extract the features. This is the initial signal processing front end that converts speech signal into a more compact and convenient mode called feature vectors. The main advantage of the wavelet transforms is that it has a varying window size, being broad at low frequencies and narrow at high frequencies, thus leading to an optimal time–frequency resolution in all frequency ranges [14]. In DWT, the original signal passes through a low-pass filter and a high-pass filter and emerges as two signals, called approximation coefficients and detail coefficients. In speech signals, low frequency components are of greater importance than high frequency signals as the low frequency components characterize a signal more than its high frequency components [15]. DWT provides sufficient information both for analysis and synthesis and reduce the computation time. DWT is defined by

\[
W(j, K) = \sum_j \sum_k X(k) 2^{-j/2} \Psi(2^{-j/2} n - k) \tag{5}
\]

Where \( \Psi(t) \) is the basic analyzing function called the mother wavelet. The successive high pass and low pass filtering of the signal is given by

\[
Y_{\text{high}}[k] = \sum_n x[n] g[2k - n] \tag{6}
\]

\[
Y_{\text{low}}[k] = \sum_n x[n] h[2k - n] \tag{7}
\]

Where \( Y_{\text{high}} \) (detail coefficients) and \( Y_{\text{low}} \) (approximation coefficients) are the outputs of the high pass and low pass filters obtained by sub sampling by 2. The
filtering is continued until the desired level is reached according to Mallat algorithm [16].

4.3 Classification using ANN

During classification stage, the input data is trained using information relating to known patterns and then they are tested using the test data set. Since neural networks are good at pattern recognition, we have also adopted this method for classification. The ability of Neural Networks to deal with uncertain, fuzzy, or insufficient data makes it an efficient pattern recognizer [17]. In this work, we use the MLP architecture, which consists of an input layer, one hidden layer and an output layer. In most networks, the principle of learning a network is based on minimizing the gradient of error [18]. Here, the input is presented to the network and moves through the weights and nonlinear activation functions towards the output layer, and the error is corrected in a backward direction using the error back propagation correction algorithm.

MLP is a supervised learning network as well as a fully connected network. The main problem here is to classify the speech sample feature vectors into several speech classes. The number of nodes in the input layer equals the feature dimension whereas number of nodes in output layer is same as the number of words in the database. Usually, only one hidden layer is enough for efficient classification. The number of nodes in the hidden layer is adjusted empirically for superior performance of the system. An activation function is applied to the net input to calculate the output response of a neuron. The network used in this work uses the sigmoid activation function where the output varies continuously but not linearly as the input changes.

5 Experiments and Results

During pre-processing, different Daubechies wavelets of orders db8, db12, db20 and db22 along with a noise of 5db are used for evaluating the performance of the proposed algorithm. The performance evaluation is calculated in terms of SNR value, spectrograms and waveform plots.

5.1 Evaluation using SNR

The table given below shows the comparison of SNR values using ST alone and using ASST. From the results it is clear that better results are obtained using db20.

<table>
<thead>
<tr>
<th>Input SNR (db)</th>
<th>Wavelet used</th>
<th>Using ST</th>
<th>Using ASST</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Db8</td>
<td>5.543</td>
<td>6.975</td>
</tr>
<tr>
<td>5</td>
<td>Db12</td>
<td>5.921</td>
<td>7.018</td>
</tr>
<tr>
<td>5</td>
<td>Db20</td>
<td>6.483</td>
<td>7.873</td>
</tr>
<tr>
<td>5</td>
<td>Db22</td>
<td>6.292</td>
<td>7.481</td>
</tr>
</tbody>
</table>

5.2 Evaluation using Spectrogram

Figure given below shows the spectrogram of the original signal, noisy signal, reconstructed signal using ST and reconstructed signal using ASST using db20.

![Comparison of spectrograms](image1)

5.3 Evaluation using Waveform Plots

Figure given below shows the waveform plot of the original signal, noisy signal, reconstructed signal using ST and reconstructed signal using ASST using db20.

![Comparison of waveform plots](image2)

The denoised signal after smoothing and soft thresholding are given to DWT where the features are extracted. 12 features are obtained using decomposition upto 8 levels. The
original signal and the approximation and detail coefficients of word Amma at the 8th level is given below.

![Original Signal Word -amma](image)

The feature vectors obtained are given to MLP for classification. Here we have divided the database into three. 70% of the data is used for training, 15% for validation and 15% for testing. The classification results obtained by using ST alone and ASST are shown in table 2.

Table 2. Comparison of classification results

<table>
<thead>
<tr>
<th>Pre-processing Method</th>
<th>Recognition accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Thresholding</td>
<td>88.5</td>
</tr>
<tr>
<td>Adaptive Smoothing Soft</td>
<td>91.3</td>
</tr>
<tr>
<td>Thresholding</td>
<td></td>
</tr>
</tbody>
</table>

6 Conclusion and Future Work

In this paper, a speech recognition system with improved performance is developed for recognizing isolated words in Malayalam using wavelet denoising based on soft thresholding and a new algorithm ASST for smoothing the signals. This algorithm enhances the performance of the speech recognition system by removing the sudden spikes which contain noise. When this algorithm is used along with soft thresholding, better results are obtained by an increase in the SNR value. All the 10000 samples from the database are used for evaluation. All the data gave an improvement in the results. From the results, it is clear that, smoothing the signal before applying any threshold method gives better results. The main advantage of adaptive smoothing of the signals is that, it can be applied along with any speech enhancement method. Different speech enhancement techniques can be used with ASST and the performance of these can be analyzed as an extension of this work.

7 References


