Chapter 4

Speech Based Watermarking Techniques

4.1 Introduction

This chapter discusses techniques developed to embed a representation of a biometric trait, such as speech which corresponds to the identity of owner, in digital images. As a biometric of human being, speech is inherent and does not change drastically along with time; it is universal and easily quantifiable. The novelty of our approach is that it embeds temporal media in a spatial media. Figure 4.1 depicts the generic diagram of biometric watermarking system. It differs from conventional watermarking system at stages which are marked in red color. Watermark extraction is emphasized rather than simple watermark detection. The overheads associated with biometric watermarking are preparation of biometric watermark from the given biometric trait. After extraction reconstruction and/or template matching is required for authentication and ownership assertion. Each technique evolved in this research work is presented in the following manner.

- Watermark preparation through feature extraction from biometric trait
- Watermark embedding
- Watermark extraction
- Experimentation and results
- Observations
Fig. 4.1: Generic block diagram of biometric watermarking scheme

4.2 Organization of Research Work

4.2.1 Hardware and Software Requirement

The experimentation carried out in this thesis is implemented on Dual core Intel 2.5 GHz CPU with 3GB RAM. All the algorithms are developed in Matlab ver.7.1 and 7.5 at different stages of implementation in Windows Xp operating environment.

4.2.2 Setup for Experimentation

4.2.2.1 Selection of host images

Gray scale as well as color images used for testing purpose are referred from site [86]. These images are of various resolutions with different formats like .JPEG, .tif, .bmp etc. Various images with different statistical properties are chosen for experimentation. Facial image ‘Lena’ is the classical image in the field of image processing. It represents low as well as moderate spatial frequency content with features like hair, eyes. This image con-
tains the smooth region on her shoulder and contrast with the background makes it difficult to embed the watermark without visible distortion. On the other hand images 'Baboon' and 'Matherann' contain high spatial frequency components. The 'Cameraman' image has the large and smooth background and it is difficult to make a tradeoff between the capacity and imperceptibility. Figure 4.2 shows various test images used for experimentation.

![Various test images](image)

**Figure 4.2:** Various gray scale and color images

### 4.2.2.2 Generation of speech watermark

For speech based watermarking algorithms speech samples are acquired in the laboratory using Pulse Coding Modulation(PCM) technique at 8 KHz sampling rate, bit rate 512 kbps, 16 bit sample size, and mono channel. Various speech sample files which are used as a watermark for experimentation are given in Table 4.1. Speech samples are English spoken words. For evaluation of watermarking techniques, speech samples of varying duration of same person are considered as a watermark.
<table>
<thead>
<tr>
<th>Speech files</th>
<th>Duration</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample-1.wav</td>
<td>1 sec</td>
<td>15.6KB</td>
</tr>
<tr>
<td>Sample-4.wav</td>
<td>4 sec</td>
<td>60.8KB</td>
</tr>
<tr>
<td>Sample-7.wav</td>
<td>7 sec</td>
<td>109KB</td>
</tr>
<tr>
<td>Sample-9.wav</td>
<td>9 sec</td>
<td>140KB</td>
</tr>
<tr>
<td>Sample-11.wav</td>
<td>11 sec</td>
<td>171KB</td>
</tr>
<tr>
<td>Sample-13.wav</td>
<td>13 sec</td>
<td>203KB</td>
</tr>
<tr>
<td>Sample-15.wav</td>
<td>15 sec</td>
<td>234KB</td>
</tr>
<tr>
<td>Sample-18.wav</td>
<td>18 sec</td>
<td>281KB</td>
</tr>
<tr>
<td>Sample-20.wav</td>
<td>20 sec</td>
<td>312KB</td>
</tr>
</tbody>
</table>

### 4.2.3 Evaluation of Watermarking Scheme

Evaluation of watermarking scheme is based on subjective evaluation of watermarked image, extracted watermark and robustness of watermarking scheme to different intentional and malicious attacks. The subjective measurement depends on various factors such as the expertise of the viewer, the experimental conditions etc. Objective measures are essential to quantify the fidelity of extracted watermark as well as watermarked image. Different measurement parameters used for speech based techniques are discussed as follow.

#### 4.2.3.1 Peak signal to Noise Ratio (PSNR)

Objective measure used to evaluate the fidelity of the watermarked image is the Peak signal to Noise Ratio (PSNR). It is the ratio of the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of image. As many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale. It is defined as follows:

$$MSE = \frac{1}{M*N} \sum_{i=1}^{M} \sum_{j=1}^{N} |I(i,j) - I^w(i,j)|^2$$  \hspace{1cm} (4.1)

The PSNR in terms of MSE is defined as

$$PSNR = 10.log_{10}\left( \frac{MAX^2}{MSE} \right) = 20.log_{10}\left( \frac{MAX}{\sqrt{MSE}} \right)$$ \hspace{1cm} (4.2)

Here, MAX is the maximum pixel value of the image.
4.2.3.2 Similarity Factor (S.F)

Quantitative measure used to evaluate the fidelity of extracted speech watermark with original one is the Similarity Factor (S.F) which is defined by equation.

\[
S.F \left(W_s, W_s^*\right) = \frac{W_s \times W_s^*}{\sqrt{W_s^* \times W_s^*}}
\]

(4.3)

4.2.3.3 Subjective evaluation

Subjective evaluation of watermarked image and extracted speech watermark is done with ten different persons. Subjective quality of watermarked image is tested from a distance of one foot to 3 feet. Intelligible contents of extracted speech watermark are verified through audible testing.

4.2.4 Types of Attacks Considered for Robustness

Most commonly used attacks for image watermarking scheme are removal attacks and geometric attacks. Under removal attacks, common types of signal processing operations which are used to obliterate the watermark are intensity adjustment, histogram equalization, salt and pepper noise, Gaussian noise, Wiener filtering, Median filtering, unsharpen and JPEG compression. Under geometric attacks row column blanking, row column copying, cropping, scaling and rotations are considered.

4.2.5 Presentation of Results and Observations

Each technique developed in the research work is rigorously tested using various test images and biometric watermarks. Results are organized in following manner.

- Different parameters set like data payload, strength of watermark are mentioned.
- Original image, original watermark, watermarked image and extracted watermark are displayed for two test images. In some techniques output snapshots of GUI are displayed directly.
- Quantitative metrics PSNR for watermarked image and S.F for corresponding extracted watermark are mentioned.
• To represent the results in succinct form tables/graphs are used for repeated experimentation with varying parameters.

• Attacked watermarked image with extracted watermark mentioning corresponding quantitative metric is displayed.

• Summary of attacks for which technique survives is displayed in tabular form.

• Any additional experimentation done is narrated in the relevant section.

### 4.3 Speech Based Watermarking Techniques

Using speech as watermark, following two techniques are developed in this research work.

I. Speech based watermarking technique using wavelet transform

II. Speech based watermarking technique using stationary wavelet transform

The major challenge in embedding speech watermark is its exorbitant file size, which increases the data payload of watermarking system. To circumvent this problem, dimensionality reduction technique is applied on speech watermark.

### 4.4 Speech Based Watermarking Technique Using Wavelet Transform

This technique aims at development of watermarking scheme to investigate the feasibility of integrating temporal media with spatial media.

#### 4.4.1 Watermark Preparation

The main objective of this step is to compute a parsimonious sequence of speech watermark through feature extraction process and it is achieved through Voice excited Linear Predictive Coding (LPC) [87, 88, 89, 90]. The speech signal is first encoded using PCM encoding technique at 8 KHz sampling rate. The Voice excited LPC employs the conventional method of calculating linear predictive parameters. Plain LPC coder is efficient in exploiting the parametric redundancy [91, 92].
4.4.1.1 Linear Predictive Coding Technique

The basic concept behind LPC analysis is that the speech samples can be approximated as a linear combination of past speech samples in a short time segment. A linear predictor model forecasts the amplitude of a signal $x(n)$ at time $n$, using a linearly weighted combination of past speech samples $x(n-1), x(n-2), \ldots, x(n-k)$. The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech signal and estimated speech signal over a finite duration.

The predictor is given as:

$$\hat{x}(n) = a_1 x(n-1) + a_2 x(n-2) + \ldots + a_p x(n-p)$$  \hspace{1cm} (4.4)

$$\hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k)$$  \hspace{1cm} (4.5)

Where integer variable ‘$n$’ is the discrete time index, $\hat{x}(n)$ is the predictor of $x(n)$ and $a_k$ are the predictor coefficients. The prediction error $e(n)$, defined as the difference between the actual sample value $x(n)$ and its predicted value $\hat{x}(n)$, is given by

$$e(n) = x(n) - \hat{x}(n)$$  \hspace{1cm} (4.6)

From Equation 4.6, a signal generated or modeled by a linear predictor can be described by the following feedback equation.

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + e(n)$$  \hspace{1cm} (4.7)

In this model, the random input excitation i.e. the prediction error is $e(n) = Gu(n)$, where $u(n)$ is the zero-mean, unit-variance random signal, and $G$ is a gain term. Taking the Z transform of the Equation 4.7, the transfer function is given as

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$  \hspace{1cm} (4.8)

Thus the system is excited by an impulse train for voiced speech or a random noise sequence for unvoiced speech. The parameters of this model are voiced/unvoiced classification, pitch period for voiced speech (time duration between consecutive excitation pulses),
gain parameter $G$, and the coefficients $a_k$ of the digital filter. All these parameters are slowly varying with time. Autocorrelation method is used to estimate the predictor coefficients. Levinson’s-Durbin recursion is used to compute the coefficients for the autocorrelation method. Figure 4.3 shows the simplified model for speech reproduction. There are some drawbacks of plain LPC. A simple periodic pulse as an excitation signal for voiced speech has limitations in reproducing complex human speech sounds. It is also very difficult to recognize and distinguish human speech as voiced or unvoiced perfectly. These problems result in synthesized speech having a strong synthetic quality. It sounds buzzy and mechanical or has annoying thumps and tonal noises. To overcome this problem, voice excited LPC is implemented.

### 4.4.1.2 Voice excited LPC

The main idea behind the voice-excitation is to avoid the imprecise detection of the pitch and the use of an impulse train while synthesizing the speech. One should try to come up with a better estimate of the excitation signal. In voice excited LPC, the input speech signal in each frame is filtered with the estimated transfer function of LPC analyzer. This filtered signal is called the residual signal. If this signal is transmitted to the receiver, one can achieve a very good quality of reconstructed sound. The tradeoff, however, is paid by a higher bit rate, although there is no longer a need to transfer the pitch frequency and the voiced/unvoiced information. The system is robust since there is no need to analyze whether the sound is voiced or unvoiced or to analyze the pitch period.
Figure 4.4 shows the block diagram of voice excited LPC. In voice excited LPC, excitation detection involves generation of residual signal and compression of residual signal. Hence the excitation of residual signal $u[n]$ is generated using filter coefficients and signal $x[n]$ for each frame. The residual signal is generated on frame basis with a frame size of 30 msec. with 10 msec. overlapping. Voice excitation is used on the receiver side to reconstruct the speech. But due to nature of source coding, whole of residual signal need not be transmitted. As Discrete Cosine Transform (DCT) has excellent energy compaction property, it is applied on the residual signal. Only first fifty coefficients of DCT are transmitted along with LPC coefficients. These DCT Residual coefficients along with LP coefficients are used as a watermark for embedding. Parameters set for encoding are:

- Speech signal bandwidth = 4 KHz
- Sampling Rate = 8 KHz
- Frame size = 30 msec. (20 msec. window length and 10 msec. overlapping)
- There are 50 frames per second
- Prediction order = 12

LPC encoded speech watermark is denoted by $w_s$.

**4.4.1.3 Reconstruction of the signal at the receiver side**

The excitation input for the filter can be obtained from Inverse Discrete Cosine Transform (IDCT) of the DCT coefficients of residual signal. The inverse of the Finite Impulse Response (FIR) filter $H(z)$ called Autoregressive (AR) model $A(z)$, is used for the synthesis
filter of the decoder. The residual signal reconstructed by taking the IDCT of the DCT coefficients transmitted is used as the excitation component for the AR model for each frame. The output of the AR model is the reconstructed speech signal on a framewise basis.

### 4.4.2 Watermark Embedding

In order to increase the security of the algorithm, LPC coefficients are randomized using secret key before embedding. Single level wavelet decomposition using db2 filter is applied on host image. The two dimensional discrete wavelet transform decomposes an image into four subbands namely LL, LH, HL, and HH as shown in Figure 4.5. The decomposed subbands correspond to the coarse approximation, horizontal, vertical and diagonal details of the image signal respectively. Large coefficients in HL band are selected for watermark insertion. Watermark embedding algorithm is as follow:

![Figure 4.5: Second level wavelet decomposition](image)

1. Apply single level wavelet decomposition on host image ‘I’ to decompose it into four bands.

2. Select the HL subband of decomposed image and generate the perceptual mask that identifies the significant perceptual components (watermark indices) of the wavelet coefficients. The method employs the largest $L^w_w$ wavelet coefficients, where $L^w_w$ is chosen to be equal to length of watermark signal.

3. Insert the watermark into selected wavelet coefficients using Cox’s algorithm [12].

$$\hat{I}_{HL}(i, j) = I_{HL}(i, j)(1 + \alpha W^i_s) \quad (4.9)$$

where $W^i_s$ = the $i^{th}$ value of speech watermark

$I_{HL}(i, j)$ = the original wavelet coefficients at $i^{th}$ location in HL subband
\( \hat{I}_{HL}(i,j) \) = modified wavelet coefficients at \( i^{th} \) location in HL subband
\( \alpha \) = the strength of the watermark

4. Generate the watermarked image \( I^w \) by applying the inverse wavelet transform.

To represent the image in spatial domain and its transformed coefficients, the same notation ‘\( I \)’ is used. This nonlinear watermark insertion procedure given by Equation 4.9 adapts the watermark to the energy present in each wavelet coefficients proportional to the value of coefficients. It draws an analogy to spread spectrum communications. The advantage of such insertion is that when the coefficient \( I(i,j) \) is small, the watermark energy is also small, thereby avoiding artifacts; and when \( I(i,j) \) is large, the watermark energy is increased for robustness. The value of the watermark strength is chosen to strike a balance between perceptibility and robustness.

4.4.3 Watermark Extraction

Watermark extraction process is a nonblind method as it requires the original image. Watermark extraction algorithm is as follow:

1. Apply the wavelet transform on original image and watermarked image to get the four bands.

2. Apply the perceptual mask to extract the wavelet coefficients from HL band that contain the watermark. The perceptual mask used in the watermark insertion process identifies the locations of wavelet coefficients in the recovered image to look for watermark information.

3. Extract the watermark by using the inverse of the insertion Equation 4.9.

\[
W^*_i = \frac{\hat{I}_{HL}(i,j)}{I_{HL}(i,j)} - 1
\]

(4.10)

Where \( W^*_i \) = the \( i^{th} \) recovered speech watermark value. The extracted watermark is in randomized form.

4. Generate the ordered watermark using the same secret key.

5. Decode the extracted watermark coefficients to synthesize speech.
4.4.4 Experimentation and Results

4.4.4.1 Data payload

Speech watermark which is in the .wav file format is LPC encoded for dimensionality reduction before embedding. Table 4.2 shows the size of original speech file and length of LPC encoded stream. On an average, payload is reduced by 61% through wavelet encoding technique.

<table>
<thead>
<tr>
<th>Duration of speech in sec.</th>
<th>Length of PCM encoded stream</th>
<th>Length of LPC encoded stream</th>
<th>Compression Factor</th>
<th>Percentage of compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8000</td>
<td>3968</td>
<td>2.612</td>
<td>61.743</td>
</tr>
<tr>
<td>4</td>
<td>32000</td>
<td>12398</td>
<td>2.581</td>
<td>61.545</td>
</tr>
<tr>
<td>7</td>
<td>56000</td>
<td>21638</td>
<td>2.588</td>
<td>61.360</td>
</tr>
<tr>
<td>9</td>
<td>72000</td>
<td>27838</td>
<td>2.586</td>
<td>61.336</td>
</tr>
<tr>
<td>11</td>
<td>88000</td>
<td>34038</td>
<td>2.585</td>
<td>61.320</td>
</tr>
<tr>
<td>13</td>
<td>104000</td>
<td>40238</td>
<td>2.584</td>
<td>61.309</td>
</tr>
<tr>
<td>15</td>
<td>120000</td>
<td>46438</td>
<td>2.584</td>
<td>61.301</td>
</tr>
<tr>
<td>18</td>
<td>144000</td>
<td>55738</td>
<td>2.583</td>
<td>61.293</td>
</tr>
</tbody>
</table>

4.4.4.2 Strength of watermark

Strength of watermark plays an important role in maintaining the tradeoff between robustness and imperceptibility. For this algorithm strength is varied between 2 to 5.

4.4.4.3 Results

Figure 4.6 shows the original watermarked image with original and extracted speech watermark for two test images Lena and Matherann for different speech watermarks of varying durations. Table 4.3 and Table 4.4 show the PSNR and S.F for different watermarks for different host images. For high resolution image like Matherann speech embedding capacity is up to 18 sec. Though the perceptual quality of watermarked image for low resolution images above 9 sec. duration of speech watermark is good, the quality of extracted watermark is not recognizable. The corresponding entries for S.F in the Table 4.4 are kept blank.
Figure 4.6: continued
(g) Lena image with 9 sec watermark
(h) Matherann image with 9 sec watermark

(i) Matherann image with 13 sec watermark
(j) Matherann image with 18 sec watermark

Figure 4.6: Results of wavelet based technique for different speech watermarks

Table 4.3: PSNR for different test images for wavelet based technique

<table>
<thead>
<tr>
<th>Duration of Watermark</th>
<th>Lena</th>
<th>Pepper</th>
<th>Baboon</th>
<th>Goldhill</th>
<th>Hat</th>
<th>Matherann</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49.12</td>
<td>45.56</td>
<td>41.32</td>
<td>45.83</td>
<td>51.23</td>
<td>46.81</td>
</tr>
<tr>
<td>4</td>
<td>48.03</td>
<td>46.51</td>
<td>40.89</td>
<td>45.47</td>
<td>50.09</td>
<td>45.2</td>
</tr>
<tr>
<td>7</td>
<td>46.93</td>
<td>46.25</td>
<td>39.61</td>
<td>44.81</td>
<td>49.87</td>
<td>44.32</td>
</tr>
<tr>
<td>9</td>
<td>47.25</td>
<td>46.01</td>
<td>39.25</td>
<td>44.31</td>
<td>49.59</td>
<td>44.07</td>
</tr>
<tr>
<td>11</td>
<td>46.91</td>
<td>45.81</td>
<td>38.06</td>
<td>43.12</td>
<td>44.67</td>
<td>44.08</td>
</tr>
<tr>
<td>13</td>
<td>45.23</td>
<td>43.27</td>
<td>37.59</td>
<td>42.81</td>
<td>43.36</td>
<td>43.91</td>
</tr>
<tr>
<td>15</td>
<td>44.67</td>
<td>47.78</td>
<td>36.21</td>
<td>41.80</td>
<td>42.81</td>
<td>44.06</td>
</tr>
<tr>
<td>18</td>
<td>43.71</td>
<td>41.03</td>
<td>34.76</td>
<td>40.57</td>
<td>41.06</td>
<td>44.25</td>
</tr>
</tbody>
</table>
4.4.4.4 Robustness against attacks

Figure 4.7 shows the watermarked image and recovered watermark against different signal processing and geometric attacks for Lena and Matherann image with speech watermark of 4 sec. duration.

![Figure 4.7: continued](image)

(a) Lena image with brightness attack

(b) Matherann image brightness attack

(c) Lena image with histogram equalization

(d) Matherann image with histogram equalization

Figure 4.7: continued
Figure 4.7: continued
Figure 4.7: continued

(k) Lena image with Wiener filtering

(l) Matherann image with Wiener filtering

(m) Lena image with unsharpening

(n) Matherann image with unsharpening

(o) Lena image with row column copy

(p) Matherann image with row column copy
Figure 4.7: Watermarked images under different attacks for wavelet based technique
Table 4.5 shows average of the S.F for various test images by varying parameters and duration of speech watermark under different attacks. In the intensity adjustment attack, if the range of parameter is greater than one, image will become brighter and if it is less than one image will become darker. For salt and pepper noise, the noise density is varied from 0.02 to 0.2. The Gaussian noise is added with zero mean and 0.01 variance. For median and Wiener filtering window size of $3 \times 3$ and $5 \times 5$ are used for testing purposes. For row column blanking copying, random rows and columns are selected as mentioned in the table.

Table 4.5: Summary of attacks for wavelet based technique for speech watermark

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Type of Attack</th>
<th>Parameters</th>
<th>Lena</th>
<th>Pepper</th>
<th>Baboon</th>
<th>Goldhill</th>
<th>Hat</th>
<th>Matherann</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intensity adjustment</td>
<td>0.2 to 1.5</td>
<td>0.9114</td>
<td>0.941</td>
<td>0.932</td>
<td>0.938</td>
<td>0.925</td>
<td>0.901</td>
</tr>
<tr>
<td>2</td>
<td>Histogram equalization</td>
<td>-</td>
<td>0.7021</td>
<td>0.7671</td>
<td>0.7321</td>
<td>0.7841</td>
<td>0.772</td>
<td>0.785</td>
</tr>
<tr>
<td>3</td>
<td>Salt and pepper noise</td>
<td>0.02 to 0.2</td>
<td>0.6072</td>
<td>0.6123</td>
<td>0.7134</td>
<td>0.6345</td>
<td>0.683</td>
<td>0.7022</td>
</tr>
<tr>
<td>4</td>
<td>Gaussian noise</td>
<td>0 mean, 0.01 variance</td>
<td>0.4196</td>
<td>0.4345</td>
<td>0.5497</td>
<td>0.5123</td>
<td>0.524</td>
<td>0.5478</td>
</tr>
<tr>
<td>5</td>
<td>Wiener Filtering</td>
<td>3x3 window , 5x5 window</td>
<td>0.7328</td>
<td>0.5421</td>
<td>0.56291</td>
<td>0.5221</td>
<td>0.523</td>
<td>0.7579</td>
</tr>
<tr>
<td>6</td>
<td>Median Filtering</td>
<td>3x3 window , 5x5 window</td>
<td>0.5319</td>
<td>0.643</td>
<td>0.6743</td>
<td>0.626</td>
<td>0.601</td>
<td>0.5683</td>
</tr>
<tr>
<td>7</td>
<td>Unsharpen</td>
<td>0.1</td>
<td>0.804</td>
<td>0.793</td>
<td>0.752</td>
<td>0.819</td>
<td>0.823</td>
<td>0.8271</td>
</tr>
<tr>
<td>8</td>
<td>Row Column copying</td>
<td>Cols= 211-123, 256-11, 455-169, 359-50 and Rows = 123-211, 11-256, 169-455, 359-50</td>
<td>0.8561</td>
<td>0.8594</td>
<td>0.8794</td>
<td>0.825</td>
<td>0.869</td>
<td>0.8084</td>
</tr>
<tr>
<td>9</td>
<td>Row Column blanking</td>
<td>Cols= 50,99,192,300 and Rows = 123-211, 11-256, 169-455, 359-50</td>
<td>0.7117</td>
<td>0.765</td>
<td>0.792</td>
<td>0.7543</td>
<td>0.7456</td>
<td>0.8182</td>
</tr>
<tr>
<td>10</td>
<td>Cropping</td>
<td>-</td>
<td>0.8197</td>
<td>0.803</td>
<td>0.812</td>
<td>0.7934</td>
<td>0.845</td>
<td>0.8155</td>
</tr>
</tbody>
</table>

The watermarked image is compressed with JPEG compression. The Q.F is varied from 100 to 40 and the effect of varying Q.F is observed on PSNR and S.F. Figure 4.8 shows
the graphs plotted for Q.F against average PSNR and average S.F for different images for varying speech watermark.

Figure 4.8: PSNR and S.F by varying Q.F for wavelet based technique

4.4.5 Observations

- Average PSNR of watermarked image for the technique discussed is 45 dB and S.F is 0.92.

- The maximum speech embedding capacity for low resolution images of size $512 \times$
512 is upto 9 sec., while for higher resolution image Matherann (1024 × 1024) embedding capacity is upto 18 sec.

- The extracted speech watermark from watermarked image is intelligible and listening tests performed are satisfactory.

- As the strength of watermark is increased the robustness of watermark increases and its recovery is good but the image suffers visible degradation. For smooth and low frequency images strength of the watermark has to be kept on lower side in the range of 2 to 3, while for high frequency images it is in between 3 to 5.

- As length of watermark is increased, PSNR of watermarked image is decreased. However there is no significant effect of length watermark on S.F of extracted watermark.

- In case of JPEG compression attack, when the Q.F is decreased, PSNR and S.F are decreased. For high frequency images like Matherann and Baboon, decrease in PSNR and S.F is more compared to other smooth and low frequency images.

- Robustness against signal processing attack
  - Method is resilient to all common signal processing attacks.
  - Method shows robustness for JPEG compression up to quality factor 40 only.

- Robustness against geometric attacks
  - Method is resistant to cropping up to 5% only.
  - Method is not resistant to scaling and rotation attacks.

- For extracted speech watermark from attacked image listening test is done with different persons. It is observed that when S.F is below a factor 0.4, extracted speech watermark is perceptually not recognizable.
4.5 Speech Based Watermarking Technique Using Stationary Wavelet Transform

For steganography/covert communication application higher embedding rate is desirable in order to communicate proper message. It is also necessary to hide the very fact that communication is taking place. The cover work containing the covert message should be inconspicuous to an adversary. The two main requirements of watermarking system, data payload and fidelity are conflicting to each other. The scheme is developed to increase the embedding capacity for speech watermark, without degrading the fidelity of the host image. This objective is met by considering two aspects:

- Stationary wavelet transform is used, as each band is of same size as that of original image, it provides more capacity for data embedding.
- More compression of speech signal is achieved by applying wavelet compression on speech watermark.

4.5.1 Watermark Preparation

The audio signal is first encoded using PCM encoding technique at 8 KHz sampling rate. Wavelet compression technique is applied to reduce the size of audio watermark. Wavelet concentrate speech information (energy and perception) into a few neighboring coefficients [93]. Therefore as a result of taking the wavelet transform of a signal, many coefficients will either be zero or have negligible magnitudes. Human psychoacoustic model is also considered for compression. Human Audio System (HAS) is more sensitive to low frequencies than high frequencies and hearing threshold is very high in the high frequency regions. High frequency coefficients above some threshold are truncated. The process of compressing a speech signal using wavelets involves a number of different stages as mentioned below:

i) Deciding optimal wavelet

ii) Deciding level of decomposition

iii) Deciding threshold criteria for the truncation of coefficients
iv) Encoding of wavelet coefficients

i) Deciding optimal wavelet

The choice of the mother wavelet function in designing the high quality speech coders is of prime importance. Choosing a wavelet that has compact support in both time and frequency in addition to a significant number of vanishing moments is essential for an optimum wavelet speech compression [94]. Several different criteria can be used for selecting an optimal wavelet function. The objective is to minimize reconstructed error variance and maximize signal to noise ratio (SNR). In general optimum wavelets can be selected based on the energy conservation properties in the approximation part of the wavelet coefficients. In [95] it was shown that the Battle-Lemarie wavelet concentrates more than 97.5% of the signal energy in the approximation part of the coefficients. This is followed very closely by the Daubechies db20, db12, db10, db8 and db6 wavelets, all concentrating more than 96% of the signal energy in the Level 1 approximation coefficients. In the proposed implementation db6 is used as a mother wavelet.

ii) Deciding level of decomposition

Choosing a decomposition level for the DWT usually depends on the type of signal being analyzed or some another criteria such as entropy. For processing the speech signal decomposition level upto five is adequate, with no further advantage gained in processing beyond level five [95]. Using this reference, a decomposition level up to four is selected for present implementation.

iii) Deciding threshold criteria for the truncation of coefficients

After calculating the wavelet transform of the speech signal, compression involves truncating wavelet coefficients below a threshold. Applying the wavelet transform to the speech signal, signal energy is concentrated into few detail coefficients. More than 90% of the wavelet coefficients are found to be insignificant, and their truncation to zero makes an imperceptible difference to the signal. Two different approaches, global thresholding and level dependent thresholding are used for truncation of small valued coefficients. Thresholding can modify the coefficients to produce consecutive zeros, which can be later encoded. By applying hard threshold, coefficients below this threshold level are zeroed.
• **Global thresholding**

It involves taking the wavelet expansion of the signal and keeping the largest absolute value coefficients. It involves thresholding every subband with the same threshold value.

• **Level dependent thresholding**

Threshold for each subband is different and is obtained using the Birge-Massart strategy [96]. If \( l \) is the decomposition level, \( m \) the length of the coarsest approximation coefficients and \( CP \) is the compression parameter, then \( l, m \) and \( CP \) define the strategy for keeping number of coefficients at each level.

1) At level \( l + 1 \) (and coarser levels), all coefficients are retained.

2) For level \( i \) from 1 to \( l \), the number of largest coefficients to be kept are defined as follow:

\[
   n_i = m(l + 2 - i)^{-CP}
\]

Proposed method uses level dependent thresholding with compression parameter \( CP = 1.5 \).

iv) **Encoding of wavelet coefficients**

Run length encoding is applied to encode consecutive zero valued coefficients.

### 4.5.2 Watermark Embedding

Watermarking scheme is based on stationary wavelet transform domain. Single level decomposition using db2 filter is applied to cover image. The two dimensional stationary discrete wavelet transform decomposes an image in to four subbands namely \( LL, LH, HL, \) and \( HH \). As stationary wavelet transform is used, each band is of same size as that of original image, which provides more room for watermark insertion. Algorithm for watermark embedding is as follow:

1) Apply the stationary wavelet transform on image to decompose it into four bands

2) Select the HL subband of decomposed image and generate the perceptual mask that identifies the significant perceptual components (watermark indices) of the wavelet co-
coefficients. The method employs the largest $'L'_w$ wavelet coefficients, where $'L'_w$ is chosen to be equal to length of watermark signal.

3) Insert the wavelet coefficients of speech watermark into selected wavelet coefficients of host image using additive multiplicative Equation 4.12.

$$\hat{I}_{HL}(i,j) = I_{HL}(i,j) + \alpha W^i_s$$  \hspace{1cm} (4.12)

where $W^i_s$ = the $i^{th}$ value of speech watermark

$I_{HL}(i,j)$ = the original wavelet coefficients at $i^{th}$ location in HL subband

$\hat{I}_{HL}(i,j)$ = modified wavelet coefficients at $i^{th}$ location in HL subband

$\alpha$ = the strength of the watermark

4) Generate the watermarked image by applying the inverse stationary wavelet transform.

### 4.5.3 Watermark Extraction

Watermark extraction algorithm is as follow:

1) Apply the stationary wavelet transform on original image and watermarked image to get the four bands.

2) Apply the perceptual mask to extract the wavelet coefficients from HL band that contain the watermark. The perceptual mask used in the watermark insertion process identifies the locations of wavelet coefficients in the recovered image to look for watermark information.

3) Extract the watermark by using Equation 4.13.

$$W^{i*}_s = \frac{\hat{I}_{HL}(i,j) - I_{HL}(i,j)}{\alpha}$$  \hspace{1cm} (4.13)

Where $W^{i*}_s$ = the $i^{th}$ recovered speech watermark value

The extracted watermark is in randomized form.

4) Generate the ordered watermark using the same secret key.
4.5.4 Experimentation and Results

4.5.4.1 Data payload

Speech watermark which is in the .wav form is encoded using wavelet transform for dimensionality reduction before embedding. Table 4.6 shows the size of original speech file and length of wavelet encoded stream. On an average payload is reduced by 85% through wavelet encoding technique of speech watermark.

Table 4.6: Length of .wav file with wavelet compression

<table>
<thead>
<tr>
<th>Duration of speech watermark in sec.</th>
<th>No. of samples in .wav file</th>
<th>No. of Samples after wavelet compression</th>
<th>Compression Factor</th>
<th>% of Compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8000</td>
<td>1282</td>
<td>6.240</td>
<td>83.941</td>
</tr>
<tr>
<td>4</td>
<td>32000</td>
<td>4794</td>
<td>6.490</td>
<td>84.593</td>
</tr>
<tr>
<td>7</td>
<td>56000</td>
<td>8211</td>
<td>6.820</td>
<td>85.337</td>
</tr>
<tr>
<td>9</td>
<td>72000</td>
<td>10619</td>
<td>6.780</td>
<td>85.251</td>
</tr>
<tr>
<td>11</td>
<td>88000</td>
<td>12979</td>
<td>6.780</td>
<td>85.251</td>
</tr>
<tr>
<td>13</td>
<td>104,000</td>
<td>15226</td>
<td>6.830</td>
<td>85.359</td>
</tr>
<tr>
<td>15</td>
<td>120000</td>
<td>17595</td>
<td>6.820</td>
<td>85.337</td>
</tr>
<tr>
<td>18</td>
<td>144,000</td>
<td>21270</td>
<td>6.770</td>
<td>85.229</td>
</tr>
<tr>
<td>20</td>
<td>160,000</td>
<td>23597</td>
<td>6.780</td>
<td>5.250</td>
</tr>
</tbody>
</table>

4.5.4.2 Strength of watermark

For this algorithm strength is varied between 2 to 5.

4.5.4.3 Results

Figure 4.9 shows the original and watermarked image with original and extracted speech watermark.
Figure 4.9: continued
Figure 4.9: continued
Figure 4.9: Results of stationary wavelet based technique for different speech watermarks

Table 4.7 and Table 4.8 show the PSNR and S.F for varying duration of speech watermarks for different test images.
Table 4.7: PSNR for different test images for stationary wavelet based technique

<table>
<thead>
<tr>
<th>Duration of Watermark</th>
<th>Lenna</th>
<th>Pepper</th>
<th>Baboon</th>
<th>Goldhill</th>
<th>Hat</th>
<th>Matherann</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60.12</td>
<td>58.44</td>
<td>56.08</td>
<td>57.59</td>
<td>53.44</td>
<td>57.05</td>
</tr>
<tr>
<td>4</td>
<td>50.23</td>
<td>57.92</td>
<td>55.59</td>
<td>56.82</td>
<td>51.11</td>
<td>55.92</td>
</tr>
<tr>
<td>7</td>
<td>46.69</td>
<td>56.20</td>
<td>54.67</td>
<td>55.76</td>
<td>49.51</td>
<td>52.68</td>
</tr>
<tr>
<td>9</td>
<td>44.89</td>
<td>54.71</td>
<td>54.11</td>
<td>52.41</td>
<td>48.28</td>
<td>51.02</td>
</tr>
<tr>
<td>11</td>
<td>44.76</td>
<td>53.23</td>
<td>52.78</td>
<td>53.93</td>
<td>47.16</td>
<td>51.42</td>
</tr>
<tr>
<td>13</td>
<td>44.11</td>
<td>48.98</td>
<td>51.96</td>
<td>53.44</td>
<td>46.01</td>
<td>50.17</td>
</tr>
<tr>
<td>15</td>
<td>43.45</td>
<td>47.84</td>
<td>50.05</td>
<td>52.61</td>
<td>44.97</td>
<td>50.11</td>
</tr>
<tr>
<td>18</td>
<td>42.91</td>
<td>46.02</td>
<td>49.12</td>
<td>51.032</td>
<td>43.54</td>
<td>49.24</td>
</tr>
<tr>
<td>20</td>
<td>41.67</td>
<td>45.32</td>
<td>48.65</td>
<td>49.87</td>
<td>42.064</td>
<td>48.64</td>
</tr>
</tbody>
</table>

Table 4.8: S.F for different test images for stationary wavelet based technique

<table>
<thead>
<tr>
<th>Duration of Watermark</th>
<th>Lenna</th>
<th>Pepper</th>
<th>Baboon</th>
<th>Goldhill</th>
<th>Hat</th>
<th>Matherann</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9314</td>
<td>0.9351</td>
<td>0.9381</td>
<td>0.9243</td>
<td>0.9621</td>
<td>0.9125</td>
</tr>
<tr>
<td>4</td>
<td>0.9022</td>
<td>0.9211</td>
<td>0.9647</td>
<td>0.9294</td>
<td>0.9294</td>
<td>0.9087</td>
</tr>
<tr>
<td>7</td>
<td>0.9421</td>
<td>0.9045</td>
<td>0.961</td>
<td>0.9072</td>
<td>0.9173</td>
<td>0.9257</td>
</tr>
<tr>
<td>9</td>
<td>0.921</td>
<td>0.9201</td>
<td>0.9534</td>
<td>0.9104</td>
<td>0.9367</td>
<td>0.9187</td>
</tr>
<tr>
<td>11</td>
<td>0.8914</td>
<td>0.9345</td>
<td>0.9741</td>
<td>0.9267</td>
<td>0.9284</td>
<td>0.8926</td>
</tr>
<tr>
<td>13</td>
<td>0.9327</td>
<td>0.9278</td>
<td>0.9173</td>
<td>0.8993</td>
<td>0.9134</td>
<td>0.9093</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9072</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9187</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8719</td>
<td></td>
</tr>
</tbody>
</table>

4.5.4.4 Robustness against attacks

Figure 4.10 shows the watermarked image and recovered watermark for couple of test images under different attacks.
Figure 4.10: continued
Figure 4.10: continued
Figure 4.10: continued
The watermarked image is compressed with JPEG compression. The Q.F is varied from 100 to 40 and the effect of varying Q.F is observed on PSNR and S.F. Figure 4.11 shows the graphs plotted for Q.F against average PSNR and average S.F for different images for varying speech watermark.
4.5.5 Observations

- Average PSNR of watermarked image for the technique discussed is 51 dB and S.F is 0.92.

- The maximum speech embedding capacity for low resolution images of size 512 × 512 is upto 13 sec., while for higher resolution image Matherann (1024 × 1024) embedding capacity is upto 20 sec.

- The extracted speech watermark from watermarked image is intelligible and listening tests performed are satisfactory.
- As the strength of watermark is increased the robustness of watermark increases and its recovery is good but the image suffers visible degradation. For smooth and low frequency images strength has to be kept on lower side in the range of 2 to 3, while for high frequency images it is in between 3 to 5.

- As length of watermark is increases, PSNR of watermarked image is decreases. However there is no any significant effect of length watermark on S.F of extracted watermark.

- In case of JPEG compression attack, when the Q.F is decreased, PSNR and S.F decreases. For high frequency images Matherann and Baboon, PSNR and S.F decreases more compared to other smooth and low frequency images.

- Robustness against signal processing attack
  - Method is resilient to all common signal processing attacks.
  - Method shows robustness for JPEG compression up to quality factor 40 only.

- Robustness against geometric attacks
  - Method is resistant to cropping up to 5% only, below this extraction fails.
  - Method is not resistant to scaling and rotation attacks.

- For extracted speech watermark from attacked image, listening test is done with different persons. It is observed that when S.F is below a factor 0.4 extracted speech watermark is perceptually not recognizable.

### 4.6 Summary

In this chapter techniques developed during the course of research work using speech as watermark were discussed. The techniques were developed to fulfill the objectives set at the beginning. Watermark preparation was carried out through classical compression techniques. The experimental set up and results derived out of these methods were also presented. The maximum speech embedding capacity of wavelet based techniques which uses LPC encoding of speech is 9 sec. for low resolution images and 18 sec. for images
of higher size. However, in case of second technique which is based on stationary wavelet transform, maximum speech embedding capacity is 13 sec. for low resolution images and 20 sec. for high resolution images without degrading the perceptual quality of image. There is no any significant difference between the S.F of original and extracted speech watermark. This technique can be used for ownership verification and covert communication. Techniques are applicable for color as well gray scale images.