The chapter provides a novel approach to emotion recognition from facial expression and voice of subjects. The subjects are asked to manifest their emotional exposure in both facial expression and voice, while uttering a given sentence. Facial features including mouth-opening, eye-opening, eyebrow-constriction, and voice features including, first three formants: F1, F2, and F3, and respective powers at those formants, and pitch are extracted for 7 different emotional expressions of each subject. A linear Support Vector Machine classifier is used to classify the extracted feature vectors into different emotion classes. Sensitivity of the classifier to Gaussian noise is studied, and experimental results confirm that the recognition accuracy of emotion up to a level of 95% is maintained, even when the mean and standard deviation of noise are as high as 5% and 20% respectively over the individual features. A further analysis to identify the importance of individual features reveals that mouth-opening and eye-opening are primary features, in absence of which classification accuracy falls off by a large margin of more than 22%.
Emotion plays a vital role in non-verbal communication, and thus has immense applications in the next generation human-machine interactive system [28], [34]. If computers can recognize our emotion from our facial expression and voice, the scope of interactions between humans and machines can be improved significantly [3]. In [22], Picard proposed a new discipline of computing, called affective computing, that monitors the affective states of people, and provides them necessary support in critical/accident-prone environment. One interesting work in this regard is due to Li and Ji [20], where the authors proposed a probabilistic framework to dynamically model and recognize the users’ affective states for timely and efficient services to those people. Picard et al. [24] also stressed the importance of emotions on our respective affective psychological states. Among others, the works of Scheirer et al. [28], Conati [6], Kramer et al. [18], and Rani et al. [25], [26] also deserve special applause in the fields of affective computing.

Besides affective computing, the study of emotions and emotional intelligence has important applications in psychological counseling, detection of anti-societies, digital movie making with artificial agents, and many others. Details of this are available in [7]. Among interesting works on emotion recognition, the work by Ekman and Friesen [7] needs special mention. They forwarded a scheme for recognition of facial expression from different regions of face, e.g. cheek, chin, and wrinkles. It reports a direct correlation of facial expression with the eyes, the eye-brows, and the mouth. Kobayashi and Hara [14], [15], [16] came forward with a neural net based [13], [31], [32], [33] solution to facial expression based emotion recognition problem. Researchers are also keen to employ Fourier descriptor [32], template matching [2], neural network models [8], [27], [32], and fuzzy integral [12] techniques for detection of emotion.

Cohen et al. considered variations in facial expressions, displayed in a live video to recognize emotions [4], [5]. She came up with a new architecture of hidden Markov models. It automatically segments and recognizes facial expressions, thus helping in emotion recognition. Gao et al. [9] attempted to interpret facial expression from a single facial image using line-based caricatures. Lanitis et al. [19] proposed a novel technique for automatic recognition and coding of face images using flexible models. Sinéad McGilloway [21] proposed a method for recognition of emotion from voice of the subjects with the help of discriminant analysis that uses linear combinations of variables to separate samples that belong to different categories using neural net classifiers. Valery A. Petrushin [23] proposed an approach of emotion recognition from voice data. He used a part of a corpus for extracting the features and for training computer-based emotion recognizers. He then took some of the features and fed them as inputs to different types of recognizers, ultimately concluding that the neural network based recognizers work best.

In this chapter, we intend to classify the emotions from the data obtained on different parameters of facial expression and voice. The experiment was conducted with 7 subjects, where each subject was asked to utter a given sentence with emotions in both his/her voice and facial expression. The voice and facial expressions are recorded for subsequent feature extraction and classification of emotions. We also check the level of noise-acceptance of these data. Further, an analysis on the importance of individual facial and voice features on emotion classification is also studied using Support Vector Machines. So, in short, this chapter deals with multi-modal emotion recognition from voice and facial expression of a subject.

This chapter includes six sections. In section 4.2, 4.3 and 4.4, experimental set-ups and the processes for extraction of facial and voice features have been discussed. Section 4.5 introduces a brief theory of Support Vector Machines.
Vector Machine and its application in recognition of emotion on the basis of facial expression and voice and also the corresponding results are shown. The conclusions are drawn in section 4.6.

4.2. **Experimental Set-Up**

This section provides an overview of the proposed experiment to study the composite benefits of face and voice features on emotion recognition in comparison to its independent counterparts. An experiment was conducted at Jadavpur University with 7 speakers, each attempting to express an English statement: ‘What are you doing here?’ with manifestation of the emotion on facial expression and voice. Each subject was pre-trained to arouse a specific emotion by acting with proper manifestation on facial expression and voice, so that the emotion can be retrieved manually by a group of 10 evaluators. The evaluators use the following policy to examine the emotion content in the videos of facial expressions and voice spectrum. Each evaluator assigns a score to one of seven emotions considered here, such that the sum of the assignments of scores to all the seven emotions is hundred. A given facial video and voice spectrum falls in emotion class $j$, if the scores $s_j$ obtained from all the 10 experts for these videos, denoted by $\sum_{j=1}^{10} s_j \geq \sum_{k=1}^{10} s_k$ for all emotion class $k$. In case the emotion class acquired now by the composite opinion of 10 evaluators does not match with the actual emotion, the subject attempted to manifest by acting, the facial video and the voice spectrum of the subject is abandoned. The facial expression video and voice spectrum of subjects, which have not been discovered by the judgment of the evaluators are passed by for recognition of emotion automatically. Fig. 4.1 and 4.2 below illustrates samples of facial expression and voice spectrum for a few selected subjects, which have been recommended by the evaluators appropriate for automatic recognition of emotion.

The automatic recognition process includes two fundamental steps. The first step includes extraction of facial features from the facial video and voice. Since the numbers of features in the present context are just a few, we do not go for feature reduction and straight away employ classification technique to classify the given features to one of seven emotion classes. Here, we use the well-known Support Vector Machine (SVM) classifier technique to classify features into one of seven emotion classes. SVM is selected for its reportedly high classification accuracy and small training time (in comparison to non-linear perception learning).

One interesting aspect of this chapter lies in determining relative performance of the SVM classifier (under the given setting of features) for three distinct problems: i) only facial expression, ii) only voice data and iii) both facial and voice data. Interestingly noticed that only voice data based classification has a very poor performance, while only facial expression based classification still survives, while the composite feature based classifier has a very good classification accuracy. The above result vividly reminds us the common observations in our regular life that both facial and voice data are required to correctly recognize emotion.

The last interesting issue undertaken in the chapter is to study the effect of noise on the classification process. To study this we add Gaussian noise of small mean (w.r.t signal mean) and increasing variance, and plot the classification accuracy for increasing noise mean/signal mean and noise variance/signal variance. It is apparent from this plots that classification accuracy falls off by bigger margins as the noise variance/signal variance and noise mean/signal mean increases. It is further observed from these plots that a sharp fall-off in the plot takes place for smaller noise mean/signal mean and noise variance/signal variance.
This fall-off, however, is not so stiff for bimodal facial plus voice and only facial expression based recognition. The above observation reminds us that only voice based emotion recognition is not very robust.
Fig 4.1 Voice waveform for emotion (a) Relax, (b) Happy, (c) Sad, (d) Fear, (e) Disgust, (f) Anger, (g) Surprise
Fig. 4.2 Facial Expression pictures of emotions (row-wise) Relax, Happy, Sad, Fear, Anger, Disgust and Surprise respectively.
4.3. FACIAL FEATURES

4.3.1. Filtering, Segmentation, Localization of Facial Components

Recognition of facial expressions by pixel-wise analysis of images is both tedious and time consuming. This chapter attempts to extract significant attributes of facial expressions by segmenting the image into components of interest. Because of the differences in the regional profiles on an image, simple segmentation algorithms, such as histogram-based thresholding techniques, do not yield good results. After conducting several experiments, we determined that for the segmentation of the mouth region, a colour-sensitive segmentation algorithm is most efficient. Further, because of apparent non-uniformity in the lip colour profile, a fuzzy segmentation algorithm is preferred. A colour-sensitive fuzzy C-means clustering algorithm is therefore selected for the segmentation of the mouth region.

Segmentation of eye-regions, here has been performed successfully by the traditional thresholding method. The hair region in human face has been easily segmented by the thresholding technique. Segmentation of the mouth and the eye regions is required to measure mouth-opening and eye-opening respectively. Segmentation of the eyebrow region is equally useful to determine the length of eyebrow constriction. The details of the segmentation techniques of different regions are presented below.

4.3.1.1. Segmentation of mouth region

Before segmenting the mouth region, we first transform the image into L*a*b space from the given RGB space. The L*a*b representation has the additional merit of ensuring a perceptually uniform color space. It defines a uniform matrix space representation of color, so that a perceptual color difference can be described by Euclidean distance measure. The color information, however, is not adequate to identify the lip region. The position information of pixels, along with their color together is a good feature to segment the lip region from the face. The fuzzy-C Means (FCM) Clustering algorithm that we would employ to detect human lip region is therefore supplied with both color and pixel-position information of the image.

The FCM clustering algorithm is an established technique for pattern classification. But its use in image segmentation in general and lip region segmentation in particular still remains a virgin area of research until today. The FCM clustering algorithm is available in any books on fuzzy pattern recognition [35], [36], [37]. In this chapter, we just illustrate how to use the FCM clustering algorithm in the proposed application.

A pixel in this chapter is denoted by 5 attributes: 3 color information and 2 pixel position information. 3 color information are L*a*b and 2 pixel position information are (x, y). The objective of this clustering algorithm is to classify the above set of 5 dimensional data points into 2 classes/partitions namely the lip region and non-lip region. Initial membership values are assigned to each 5 dimensional pixel data such that sum of membership in the lip and the non-lip region is equal to one. Mathematically for the kth pixel xk,

\[
\mu_L(x_k) + \mu_{NL}(x_k) = 1
\]  

where \( \mu_L \) and \( \mu_{NL} \) denote the membership of \( x_k \) to fall in the lip and the non-lip regions respectively.

Given the initial membership values of \( \mu_L(x_k) \) and \( \mu_{NL}(x_k) \) for \( k = 1 \) to \( n^2 \) (assuming that the image is of size \( n \times n \)), we use the FCM algorithm to determine the cluster centers, \( V_L \) and \( V_{NL} \) of the lip and the non-lip regions:

\[
V_L = \frac{\sum_{k=1}^{n^2} [\mu_L(x_k)]^m x_k}{\sum_{k=1}^{n^2} [\mu_L(x_k)]^m}
\]

102
and 

\[ V_{NL} = \frac{\sum_{k=1}^{n^2} [\mu_{NL}(x_k)]^m x_k}{\sum_{k=1}^{n^2} [\mu_{NL}(x_k)]^m} \]  

(4.3)

Expressions (4.2) and (4.3) indicate centroidal measures of the lip and non-lip clusters, computed over all data points \( x_k \) for \( k=1 \) to \( n^2 \). The parameter \( m (>1) \) is any real number that affects the membership grade. The membership values of pixel \( x_k \) in the image for the lip and the non-lip regions are obtained from the following formulae:

\[
L(x_k) = \left( \sum_{j=1}^{2} \left( \frac{\| x_k - v_L \|^2}{\| x_k - v_j \|^2} \right)^{\frac{1}{(M-1)}} \right)^{-1} 
\]  

(4.4)

\[
NL(x_k) = \left( \sum_{j=1}^{2} \left( \frac{\| x_k - v_{NL} \|^2}{\| x_k - v_j \|^2} \right)^{\frac{1}{(M-1)}} \right)^{-1} 
\]  

(4.5)

where \( V_j \) denotes the \( j\)-th cluster center for \( j \in \{L, NL\} \).

Fig. 4.3 The Original Face Image  
Fig. 4.4 The Median-filtered Image

Fig. 4.5 The image after applying fuzzy C-means clustering.

Fig. 4.6 Measurement of mouth opening from the dips in average intensity plot
The two step process of determination of the cluster centers by expression (4.2) and (4.3) and membership function evaluation by (4.4) and (4.5) are repeated several times following FCM algorithm until the position of the cluster centers do not shift further.

Fig. 4.3 presents a section of a facial image with a large mouth opening. This image is passed through a median filter and the resulting image is shown in Fig. 4.4. Application of FCM algorithm on Fig. 4.4 yields Fig. 4.5. In Fig. 4.6, we demonstrate the computation of mouth opening.

4.3.1.2. Localization of the Eye-Region

The eye-region in a monochrome image has a sharp contrast with respect to the rest of the face. Consequently, the thresholding method can be employed to segment the eye-region from the image. Images grabbed at poor illuminating conditions have a very low average intensity value. Segmentation of eye region in these images is difficult because of the presence of dark eyebrows in the neighborhood of the eye region. To avoid this problem, we consider the images grabbed under good illuminating condition. After segmentation is over, we localize the left and the right eyes on the image. In this chapter, we use template-matching scheme to localize the eyes. The eye template we used looks like Fig. 4.7. The template-matching scheme we used here is taken from our previous works [38], [39]. It attempts to minimize the Euclidean distance between a fuzzy descriptor of the template with the fuzzy descriptor of the part of the image where the template is located. It needs mention here that in case the template is not a part of the image, the nearest matched location of the template in the image can be traced.

![A Symmetric Eye Template](image)

Fig. 4.7 A Symmetric Eye Template

4.3.1.3. Segmentation of the Eyebrow Constriction

In a facial image, eyebrows are the 2nd dark region next to the hair region. The hair region is easily segmented by setting a very low threshold in the histogram-based thresholding algorithm. The eye regions are also segmented by the method discussed earlier. Naturally, a search algorithm invoked to determine a dark narrow template can localize the eyebrows. It is indeed important to note that localization of eyebrows is essential to determine its length. This will be taken up in the next section.

4.3.2. Determination of Facial Attributes

In this section, we present a scheme for on-line measurements of facial extracts such as mouth-opening (MO), eye-opening (EO) and the length of eyebrow-constriction (EBC).

4.3.2.1. Determination of the Mouth-Opening

After segmentation of the mouth region, we consider a plot of the average intensity profile against the mouth opening. Determination of MO in a black and white image is easier due to the presence of the white teeth. A plot of average intensity profile against the MO reveals that the curve will have several minima, out of which the first and the third correspond to one the inner region of the top lip and the inner region of the bottom lip respectively. The difference along the MO axis (Y-axis) of the above two measures gives a measurement of the MO. An experimental instance of MO is shown in Fig. 4.6. In this Fig, the pixel count between the thick horizontal lines gives a measure of MO.
4.3.2.2. Determination of the Eye-opening

After the localization of the eyes, the count of dark pixels (having intensity <30) plus the count of the white pixels (having intensity > 225) is plotted against the x-position. Suppose that the peak of this plot occurs at x= a. Then the ordinate at x= a gives a measure of the eye-opening.

4.3.2.3. Determination of the Length of Eyebrow-constriction

Constriction in the forehead region can be explained as a collection of white and dark patches called hilly and valley regions respectively. The valley regions usually are darker than the hilly regions. Usually the width of the patches is around 10-15 pixels for a given facial image of (512 x512) pixels.

Let I_{av} be the average intensity in a selected rectangular profile on the forehead and I_{ij} be the intensity of pixel (i, j). To determine the length of eyebrow constriction on the forehead region, we scan for variation in intensity along the x-axis of the selected rectangular region. The maximum x-width that includes variation in intensity is defined as the length of eyebrow-constriction. The length of the eyebrow constriction has been measured in Fig. 4.8 by using the above principle.

An algorithm for eyebrow-constriction is presented below.

1. Take a narrow strip over the eyebrow region with thickness two-third of the eye-opening.

2. The length l of the strip is determined by identifying its intersection with the hair regions at both ends. Determine the center of the strip, and select a window of x-length 2l/ 3 symmetric with respect to the center.

3. For x-positions central to window-right-end do
   Select 9 vertical lines in the window and compute average intensity on each line;
   Take variance of the 9 average intensity values;
   If the variance is below a threshold, stop;
   Else shift one pixel right.

4. Determine the total right shift

5. Similar to step 3, determine the total left shift.

6. Compute length of eyebrow-constriction = total left shift + total right shift.
4.3.3. **Extraction of facial expression data for experiment**

Extraction of feature plays an important role in recognizing a particular emotion. In order to extract features for facial expression, we need to restrict our attention to mouth region, eye region and eyebrow region in order to obtain the required data for mouth-opening, eye-opening and eyebrow constriction [37]. Next, we fuzzify these features in three fuzzy sets as HIGH, MODERATE, and LOW in the following way:

\[
\begin{align*}
\mu_{\text{HIGH}}(x) &= 1 - \exp(-a \cdot x), \quad a>0, \\
\mu_{\text{LOW}}(x) &= \exp(-b \cdot x), \quad b>0, \\
\mu_{\text{MODERATE}}(x) &= \exp \left[-\frac{(x - x_{\text{mean}})^2}{2\sigma^2}\right]
\end{align*}
\]

where \( x \in \{\text{mo, eo, ebc}\} \) and \( a, b, c, x_{\text{mean}} \) and \( \sigma \) are fixed experimentally. These fuzzified features are used to represent facial expressions in the subsequent part of this work. The measured facial features for emotion relaxation are listed in Table 4.1.

**Table 4.1**
Feature Extracted for Emotion Relaxation

<table>
<thead>
<tr>
<th>Features</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye opening in pixels</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>11</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Mouth opening in pixels</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>Eye-brow constriction in pixels</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Fig. 4.8 Determination of eye-brow constriction
4.4. Voice Feature Extraction

4.4.1. Voice Features

4.4.1.1. Pitch

The former definition of Pitch given by Webster Dictionary states: “the property of a sound and especially a musical tone that is determined by the frequency of the waves producing it: highness or lowness of sound”. But, a rather simple definition of Pitch is: Pitch represents the perceived fundamental frequency of a sound. Fundamental frequency is defined as the frequency at which the vocal cords vibrate during speech.

A frequently used method to estimate pitch (fundamental frequency) from the waveform directly is to use autocorrelation. This method is based on detecting the highest value of the autocorrelation function in the region of interest. The autocorrelation function for a signal shows how well the waveform shape correlates with itself at a range of different delays. The correlation between two waveforms is a measure of their similarity. The waveforms are compared at different time intervals, and their similarity is calculated at each interval. The autocorrelation function itself is periodic.

The autocorrelation function of discrete time deterministic signal is defined as,

\[ R(m) = \sum_{n=-\infty}^{N} x(n)x(n + m) \]  

(4.6)

If the signal is periodic, the appropriate definition is,

\[ R(m) = \lim_{N \to \infty} \left( \frac{1}{(2N + 1)} \right) \sum_{n=-N}^{N} x(n)x(n + m) \]

(4.7)

If the signal is periodic with period T samples, then,

\[ R(m+T) = R(m) \]

We can see from the following figure that the autocorrelation function peaks at zero delay and at delays corresponding to 1T, 2T, etc.
We can estimate the fundamental frequency or pitch by looking for a peak in the delay interval corresponding to the normal pitch range in speech.

4.4.1.2 Formant

A formant is a peak in a frequency spectrum that causes from the resonant frequencies of any acoustical system. Most of the formants are produced by tube and chamber resonance. For human voice, formants are recognized as the resonance frequencies of the vocal tracts. Formant regions are not directly related to the fundamental frequency and may remain more or less constant as the fundamental changes. If the fundamental is low in the formant range, the quality of the sound is rich, but if the fundamental is above the formant regions the sound is thin.

There are several formants, at different frequencies, such as F1, F2, F3. F1 is the lowest frequency, F2 is the second, F3 is the third and so on. Most often, 1st three formants (F1, F2, and F3) are enough to disambiguate the speech. Area of the major constriction determines the location of F1, decrease the area of major constriction results in decrease of F1. Distance from the glottis to the major constriction determines the location of F2, increase the distance results increase of F2.

We have taken first three formants & their corresponding peak power obtained by Fourier analysis as the speech recognition parameters. In following figures FFT of speech wave for different emotions have been shown.
Fig. 4.10 First three formant regions of a speech spectrum

Here the spectrum shows three formant regions. The vertical lines represent harmonics produced by vibration of the vocal cords and based on a low fundamental. These harmonics are resonated by the vocal tract to create the speech’s characteristic spectral shape.

4.4.1.3 Power Spectral Density

The amount of power per unit (density) of frequency (spectral) as a function of the frequency is called Power Spectral Density. Power Spectral density describes the distribution of power of a speech signal with frequency and also shows the strength (signal energy is strong or weak at different frequency) of the signal as a function of frequency. The energy or power (average energy per frame) in a formant comes from the sound source (vibration of the vocal folds, frequency of the vocal tract, movement of lips and jaw).

The energy in the speech signal \( x(n) \) is computed as,

\[
E_x = \sum_{n=0}^{N-1} X_n^2
\]

(4.8)

The power of the signal \( x(n) \) is the average energy per frame

\[
P_x = \frac{1}{N} \sum_{n=0}^{N-1} X_n^2
\]

(4.9)

\( N=\)total no of samples in a frame.

4.4.2 Extraction of Voice data

To obtain data for voice, the audio-visual clip was first read by using the MATLAB command “wavread” that reads WAVE (.wav) file. Then subsequent calculations were done in order to obtain the pitch, formants 1, 2 and 3 along with the powers at respective formants. Pitch was calculated using the MATLAB code “xcorr” that gave the auto-correlation between the sequences, one obtained using command "wavread" and
the other, with the minimum speech at 50 Hz with the sequence being normalized so that the autocorrelations at zero lag are identically 1.0. [10]

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Time duration of speech (sec)</th>
<th>Pitch (Hz)</th>
<th>Formnt 1 (Hz)</th>
<th>Formnt 2 (Hz)</th>
<th>Formnt 3 (Hz)</th>
<th>Power at Formant 1</th>
<th>Power at Formant 2</th>
<th>Power at Formant 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral</td>
<td>3.93</td>
<td>182.944</td>
<td>265.625</td>
<td>1263.7</td>
<td>2043.0</td>
<td>0.000841</td>
<td>0.0000339</td>
<td>0.0000140</td>
</tr>
<tr>
<td>happy</td>
<td>2.85</td>
<td>196.721</td>
<td>265.625</td>
<td>1310.5</td>
<td>2105.5</td>
<td>0.000482</td>
<td>0.0000581</td>
<td>0.0000125</td>
</tr>
<tr>
<td>sad</td>
<td>3.87</td>
<td>175.862</td>
<td>109.375</td>
<td>1169.9</td>
<td>2699.2</td>
<td>0.0018</td>
<td>0.00002854</td>
<td>0.0000207</td>
</tr>
<tr>
<td>fear</td>
<td>3.57</td>
<td>142.433</td>
<td>281.25</td>
<td>1482.4</td>
<td>2043.0</td>
<td>0.0010</td>
<td>0.0006623</td>
<td>0.0000200</td>
</tr>
<tr>
<td>disgust</td>
<td>3.21</td>
<td>204</td>
<td>171.875</td>
<td>1388.7</td>
<td>2449.2</td>
<td>0.0024</td>
<td>0.0000182</td>
<td>0.00002097</td>
</tr>
<tr>
<td>anger</td>
<td>3.21</td>
<td>298.763</td>
<td>109.375</td>
<td>1044.9</td>
<td>2308.6</td>
<td>0.0075</td>
<td>0.0000687</td>
<td>0.0000278</td>
</tr>
<tr>
<td>surprise</td>
<td>3.21</td>
<td>257.07</td>
<td>140.625</td>
<td>1888.7</td>
<td>2683.6</td>
<td>0.0006339</td>
<td>0.00002779</td>
<td>0.00000479</td>
</tr>
</tbody>
</table>

4.5. SUPPORT VECTOR MACHINE AND ITS APPLICATION

A. Support Vector Machine (SVM)

An SVM has been successfully used for both linear and non-linear classification. However, as non-linear operation yields results with lesser accuracy, in this chapter, we focus on the linear operation only. To understand the basic operation of SVM, let us consider Fig. 4.11 where X is the input vector and y is the desired scalar output that can take +1 or -1 values, indicating linear separation of the pattern vector X.

The function f (X, W, b) can be represented as follows:

\[ f(X, W, b) = \text{sign}(WX + b) \]

where

\[ W=[w_1 \ w_2 \ldots \ w_n] \] is the weight vector
\[ X=[x_1 \ x_2 \ldots \ x_n]^T \] represents the input vector
\[ b=[b_1 \ b_2 \ldots \ b_n] \] represents the bias vector
Fig 4.11 Defining support vector for a linear SVM system.

The function \( f \) classifies the input vector \( X \) into two classes denoted by +1 or -1. The straight line that segregates the two pattern classes is usually called a hyperplane. Further, the data points that are situated at the margins of the two boundaries of the linear classifier are called support vectors. Fig. 4.11 describes a support vector for a linear SVM.

Let us now select two points \( X^+ \) and \( X^- \) as two support vectors. Thus by definition

\[
WX^+ + b = +1
\]

and

\[
WX^- + b = -1
\]

which jointly yields

\[
W(X^+ - X^-) = 2.
\]

Now, the separation between the two support vectors lying in the class +1 and class -1, called marginal width is given by

\[
M = \frac{(WX^+ + b) - (WX^- + b)}{\|W\|} = \frac{2}{\|W\|}
\]

The main objective in a linear Support Vector Machine is to maximize \( M \), i.e., to minimize \( \|W\| \), which is same as minimizing \( \frac{1}{2}W^TW \). Thus, the linear SVM can be mathematically described by:
Minimize $\mathcal{L}(W) = \frac{1}{2} W^T W$ subject to $y_i (WX_i + b) \geq 1$ for all $i$, where $y_i$ is either 1 or -1 depending on the class which $X_i$ belongs to.

Here, the objective is to solve $W$ and $b$ that satisfies the above equation. In this chapter, we are not presenting the solution to the optimization problem, referred to above. This is available in standard text in neural network [11]. One important aspect of SVM is the kernel function selection. For linear SVM, the kernel $K$ for two data points $X_i$ and $X_j$ is defined by

$$K(X_i, X_j) = X_i^T X_j$$

B. Application of SVM

After the feature extraction is complete, we constructed training instances using both facial and voice features, and used `svmtrain` code of MATLAB to train the support vector machine with the training instances. The classification of features into one of 7 classes is then studied with MATLAB code `svmclassify`. The trained instances were correctly classified with 100% accuracy.

Next we studied the effect of Gaussian noise on emotion classification. We added Gaussian noise of specific mean and standard deviation over the mean and variance of individual components of the feature vector, and the classification was performed using MATLAB code `svmclassify`. The percentage classification noisy feature vector into emotion classes is then studied with varied ratio of noise mean to data mean and noise variance to data variance as in Table 4.3.

![Accuracy Plot for Linear SVM with various range of noise](image)

Fig. 4.12 Classification Accuracy of Noisy Feature vectors over emotion classes
### Table 4.3
Accuracy of linear classification of Facial Expression and Speech data

<table>
<thead>
<tr>
<th>Mean of Noise / Mean of Data</th>
<th>0</th>
<th>0.02</th>
<th>0.04</th>
<th>0.06</th>
<th>0.08</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>97.95</td>
<td>87.75</td>
</tr>
<tr>
<td>0.02</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>94.57</td>
<td>87.73</td>
</tr>
<tr>
<td>0.04</td>
<td>100.0</td>
<td>100.0</td>
<td>99.97</td>
<td>99.59</td>
<td>93.93</td>
<td>87.24</td>
</tr>
<tr>
<td>0.06</td>
<td>100.0</td>
<td>99.95</td>
<td>99.73</td>
<td>97.85</td>
<td>93.00</td>
<td>87.26</td>
</tr>
<tr>
<td>0.08</td>
<td>99.97</td>
<td>99.77</td>
<td>98.89</td>
<td>96.75</td>
<td>92.34</td>
<td>87.40</td>
</tr>
<tr>
<td>0.1</td>
<td>99.81</td>
<td>99.63</td>
<td>98.32</td>
<td>96.42</td>
<td>91.69</td>
<td>86.79</td>
</tr>
<tr>
<td>0.12</td>
<td>99.53</td>
<td>99.02</td>
<td>98.36</td>
<td>94.89</td>
<td>90.59</td>
<td>86.28</td>
</tr>
<tr>
<td>0.14</td>
<td>98.97</td>
<td>98.65</td>
<td>96.97</td>
<td>94.30</td>
<td>89.53</td>
<td>85.53</td>
</tr>
<tr>
<td>0.16</td>
<td>98.57</td>
<td>97.63</td>
<td>96.12</td>
<td>92.95</td>
<td>90.22</td>
<td>85.55</td>
</tr>
<tr>
<td>0.18</td>
<td>97.51</td>
<td>97.22</td>
<td>95.02</td>
<td>91.61</td>
<td>89.10</td>
<td>84.48</td>
</tr>
<tr>
<td>0.2</td>
<td>96.53</td>
<td>96.46</td>
<td>94.22</td>
<td>91.53</td>
<td>88.53</td>
<td>84.00</td>
</tr>
<tr>
<td>0.22</td>
<td>95.91</td>
<td>94.83</td>
<td>93.40</td>
<td>90.95</td>
<td>86.87</td>
<td>83.57</td>
</tr>
<tr>
<td>0.24</td>
<td>94.04</td>
<td>93.42</td>
<td>92.65</td>
<td>90.40</td>
<td>87.48</td>
<td>83.30</td>
</tr>
<tr>
<td>0.26</td>
<td>93.18</td>
<td>92.91</td>
<td>91.04</td>
<td>88.75</td>
<td>85.44</td>
<td>82.18</td>
</tr>
<tr>
<td>0.28</td>
<td>91.83</td>
<td>90.73</td>
<td>89.89</td>
<td>87.85</td>
<td>86.10</td>
<td>81.26</td>
</tr>
<tr>
<td>0.3</td>
<td>91.34</td>
<td>90.44</td>
<td>89.02</td>
<td>87.40</td>
<td>83.59</td>
<td>80.63</td>
</tr>
<tr>
<td>0.32</td>
<td>89.32</td>
<td>89.32</td>
<td>87.04</td>
<td>86.06</td>
<td>82.46</td>
<td>80.36</td>
</tr>
<tr>
<td>0.34</td>
<td>88.22</td>
<td>88.12</td>
<td>86.73</td>
<td>85.77</td>
<td>83.00</td>
<td>79.06</td>
</tr>
<tr>
<td>0.36</td>
<td>86.77</td>
<td>85.95</td>
<td>85.57</td>
<td>84.22</td>
<td>81.55</td>
<td>78.59</td>
</tr>
<tr>
<td>0.38</td>
<td>85.55</td>
<td>85.36</td>
<td>84.77</td>
<td>82.85</td>
<td>80.73</td>
<td>77.55</td>
</tr>
<tr>
<td>0.4</td>
<td>84.91</td>
<td>84.95</td>
<td>83.69</td>
<td>81.87</td>
<td>80.26</td>
<td>77.91</td>
</tr>
<tr>
<td>0.42</td>
<td>84.44</td>
<td>83.34</td>
<td>82.32</td>
<td>81.42</td>
<td>78.71</td>
<td>75.75</td>
</tr>
<tr>
<td>0.44</td>
<td>81.89</td>
<td>83.08</td>
<td>80.55</td>
<td>80.14</td>
<td>78.32</td>
<td>75.26</td>
</tr>
<tr>
<td>0.46</td>
<td>80.61</td>
<td>81.04</td>
<td>80.16</td>
<td>79.48</td>
<td>76.97</td>
<td>74.48</td>
</tr>
</tbody>
</table>

The classification accuracy obtained in Table 4.3 is plotted in Fig. 4.12 for the sake of convenience. It is clear from Fig. 4.12 that larger is the ratio of mean noise to mean data, and standard deviation of noise to standard deviation of data, the larger is the fall-off in percentage classification accuracy. It is clear from Fig. 4.12 that for a 10% ratio of mean noise to data noise and 1% ratio of standard deviation of noise to standard deviation of data the percentage accuracy in emotion classification is as high as 90.14%. This proves the robustness of the SVM classification.

### C. Classification with Only facial and Only Voice Features

To study the significance of both voice and facial features, we study the classification with only facial and only voice features, and plot the classification accuracy with varied ratio of mean noise to mean data, and
standard deviation of noise to standard deviation of data. The results are tabulated in Table 4.4 and 4.5. Fig. 4.13 and 4.14 present the classification accuracy surfaces for only facial and only voice feature sets.

**Fig 4.13** Accuracy plot for linear SVM of all the features of facial expression data

**Fig 4.14** Accuracy plot for linear SVM of all the features of voice data

It is apparent from Fig. 4.13 and 4.14 that the fall-off in percentage classification in only voice is much steeper than only facial expression based classification. Further, in either case the classification is poorer than both facial expression and voice based classification.
Table 4.4
Accuracy of Linear classification of facial expression data

<table>
<thead>
<tr>
<th>Mean of noise/ Mean of data</th>
<th>Standard deviation of noise/ Standard deviation of data</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>100</td>
<td>99.8</td>
<td>95.5</td>
<td>91.6</td>
<td>84.1</td>
<td>77.1</td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>100</td>
<td>99.4</td>
<td>95.3</td>
<td>88.4</td>
<td>80</td>
<td>75.9</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>100</td>
<td>99.2</td>
<td>93.5</td>
<td>87.4</td>
<td>80.6</td>
<td>77.4</td>
</tr>
<tr>
<td>0.15</td>
<td></td>
<td>100</td>
<td>96.7</td>
<td>91.4</td>
<td>86.3</td>
<td>79.8</td>
<td>74.5</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td>100</td>
<td>96.5</td>
<td>87.8</td>
<td>84.5</td>
<td>77.4</td>
<td>74.1</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>100</td>
<td>93.3</td>
<td>86.7</td>
<td>82.5</td>
<td>77.6</td>
<td>69.6</td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>100</td>
<td>89.8</td>
<td>87.8</td>
<td>85.1</td>
<td>82.2</td>
<td>74.7</td>
</tr>
<tr>
<td>0.35</td>
<td></td>
<td>89.8</td>
<td>85.1</td>
<td>80.2</td>
<td>76.5</td>
<td>73.3</td>
<td>70</td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td>81.6</td>
<td>78</td>
<td>76.9</td>
<td>74.7</td>
<td>74.5</td>
<td>67.6</td>
</tr>
<tr>
<td>0.45</td>
<td></td>
<td>71.4</td>
<td>71.4</td>
<td>73.9</td>
<td>74.7</td>
<td>70.2</td>
<td>69.6</td>
</tr>
</tbody>
</table>

Table 4.5
Accuracy of Linear classification of Voice data

<table>
<thead>
<tr>
<th>Mean of noise/ Mean of data</th>
<th>Standard deviation of noise/ Standard deviation of data</th>
<th>0</th>
<th>0.02</th>
<th>0.04</th>
<th>0.06</th>
<th>0.08</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>53.3</td>
<td>49</td>
<td>44.3</td>
<td>41.4</td>
<td>37.3</td>
<td>32.8</td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>51.3</td>
<td>46.9</td>
<td>43.1</td>
<td>38.6</td>
<td>37.3</td>
<td>33.5</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>49.7</td>
<td>44.8</td>
<td>41.3</td>
<td>38.8</td>
<td>37.7</td>
<td>32.7</td>
</tr>
<tr>
<td>0.15</td>
<td></td>
<td>46.7</td>
<td>43.1</td>
<td>41.4</td>
<td>38.2</td>
<td>36.2</td>
<td>33.5</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td>45.1</td>
<td>43.4</td>
<td>40.4</td>
<td>38.8</td>
<td>35.6</td>
<td>33.8</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>44.1</td>
<td>41.2</td>
<td>39.3</td>
<td>36.8</td>
<td>35.5</td>
<td>31.8</td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>43.6</td>
<td>40.1</td>
<td>38.5</td>
<td>35.8</td>
<td>33.2</td>
<td>32.2</td>
</tr>
<tr>
<td>0.35</td>
<td></td>
<td>41</td>
<td>40.4</td>
<td>37.8</td>
<td>35.7</td>
<td>34.4</td>
<td>32.4</td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td>42</td>
<td>39.2</td>
<td>37.1</td>
<td>34.3</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>0.45</td>
<td></td>
<td>42.5</td>
<td>39.1</td>
<td>37.5</td>
<td>34.3</td>
<td>31.6</td>
<td>31</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>39.5</td>
<td>38.2</td>
<td>35.7</td>
<td>33.5</td>
<td>32.7</td>
<td>29.2</td>
</tr>
</tbody>
</table>

It is clear from Table 4.4 and 4.5 that for a 10% ratio of mean noise to data noise and 1% ratio of standard deviation of noise to standard deviation of data, the percentage accuracy in emotion classification by only facial features and only voice features are as low as 88% and 37% respectively. This justifies the significance of composite facial and voice features. It is also noted that facial features are more important than the voice features from the emotion classification point of views.
D. Study of Importance of Features

The experiment started with several facial and voice features to recognize emotion of the subject. It is, however, questionable whether all the facial and voice features are equally useful. One way to determine this is to drop one feature from the application instance, and measure the % accuracy in classification. In case there is a significant (say 5% or more) fall-off in percentage classification of emotions, then that particular feature is regarded as an essential feature. If the fall-off is less than 5%, we can ignore the specific feature, as dropping it out of the list does not make significant changes in decision-making from the classifier point of views. The process of dropping one feature from the application (measurement) instance and continuing classification in absence of the selected feature is thus repeated for each feature one by one, and the significant changes in the result are tabulated in Table 4.6 below. A cross under a column in Table 4.6 means that the particular feature is dropped from the application instance to recognize emotion class.

### Table 4.6

Classification of emotions after removal of features

<table>
<thead>
<tr>
<th></th>
<th>eye-opening</th>
<th>mouth-opening</th>
<th>eyebrow-constriction</th>
<th>pitch formants</th>
<th>accuracy (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

It is apparent from Table 4.6 that mouth-opening and eye-opening are fundamental facial features, and first formant is also a very important voice feature, in absence of which percentage classification falls off by a large margin.

An experiment with limited 7 features, as indicated in Table 4.7, was used to test the % misclassification of emotion for unknown application instances (i.e. those instances not used for training). The experimental results are tabulated in Table 4.7.

### Table 4.7

Table for misclassification of emotions (IN %)

<table>
<thead>
<tr>
<th>Determined Class</th>
<th>Neutral</th>
<th>Happy</th>
<th>Sad</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>89.72</td>
<td>10.14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0</td>
</tr>
<tr>
<td>Happy</td>
<td>0.29</td>
<td>99.43</td>
<td>0.14</td>
<td>0.14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sad</td>
<td>18.43</td>
<td>3.86</td>
<td>77.71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.71</td>
<td>0</td>
<td>98.71</td>
<td>0.58</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>
It is clear from Table 4.7 that anger and disgust are correctly classified even with unknown application instance. Fear is misclassified rarely as anger. Sadness is misclassified as neutral or happy. Neutral is sometimes misclassified as happy, and happy is mis-classified as neutral or sad. On a whole, sadness is not correctly recognized by more than 20%.

4.6. CONCLUSION

The chapter proposed a new approach to emotion classification from facial expression and voice using linear support vector machine. Facial and voice features are extracted from the subjects, having arousal of emotions in both their voice and facial expression while uttering a given sentence.

Significance of the chapter lies in studying the effect of noisy features on the classification of emotion. Gaussian noise is added to the features with varying mean and standard deviation, and the percentage fall-off in classification is recorded with increased noise variance. It follows from the experiment that 95% classification accuracy can be maintained even when the mean and variance of noise is as high as 5% and 20% respectively over their individual features.

One more interesting study undertaken in the chapter is to study the importance of individual feature. This was studied by dropping one feature one at a time, and by measuring the drop-up in percentage classification accuracy. Those features giving rise to high drop-up are considered as essential features. By this approach, we identified eye-opening, mouth-opening and the first formant as essential features for emotion classification from facial expression and voice.

The significance of composite voice and facial expression over the individual expressions is also studied, and the results show that the composite feature is more robust for classification, when noise is picked up in measurements. The percentage fall-off surfaces with increased mean and variance of noise reveal that the fall-off is much sharper when only voice features is used, is nominal when facial features only is used, and slow when composite features of face and voice is used.

References


