SIViP
DOI 10.1007/s11760-013-0522-6

ORIGINAL PAPER

Recognition of emotions from video using acoustic and facial features

K. Sreenivasa Rao · Shashidhar G. Koolagudi

Received: 31 January 2011 / Revised: 26 June 2013 / Accepted: 27 July 2013

Abstract In this paper, acoustic and facial features extracted from video are explored for recognizing emotions. The temporal variation of gray values of the pixels within eye and mouth regions is used as a feature to capture the emotion-specific knowledge from the facial expressions. Acoustic features representing spectral and prosodic information are explored for recognizing emotions from the speech signal. Autoassociative neural network models are used to capture the emotion-specific information from acoustic and facial features. The basic objective of this work is to examine the capability of the proposed acoustic and facial features in view of capturing the emotion-specific information. Further, the correlations among the feature sets are analyzed by combining the evidences at different levels. The performance of the emotion recognition system developed using acoustic and facial features is observed to be 85.71 and 88.14 %, respectively. It has been observed that combining the evidences of models developed using acoustic and facial features improved the recognition performance to 93.62 %. The performance of the emotion recognition systems developed using neural network models is compared with hidden Markov models, Gaussian mixture models and support vector machine models. The proposed features and models are evaluated on real-life emotional database, Interactive Emotional Dyadic Motion Capture database, which was recently collected at University of Southern California.

K. S. Rao (✉)
School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India
e-mail: ksrao@iitkgp.ac.in

S. G. Koolagudi
Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal 575025, Karnataka, India
e-mail: koolagudi@nitk.ac.in

Keywords Emotion recognition · Autoassociative neural network (AANN) · Spectral and prosodic features · Facial features, Acoustic features

1 Introduction

Human beings exploit emotions extensively for conveying messages and their intentions. There are numerous applications based on emotion recognition (ER). Emotion recognition can provide natural interface between the humans and machines. If machines could recognize the emotions through either speech or facial expressions, they could provide appropriate help to users, particularly for handicapped people. Automatic emotion recognition and analysis are used in call centers to handle customer queries by machine based on customer's mood [1]. In case of story telling and e-tutoring applications, the system should automatically analyze the students' behavioral characteristics based on their emotions and respond accordingly with desired emotions [2,3]. The automatic schemes of analysis of emotions in multimedia documents are useful for indexing and retrieving the multimedia files based on emotion-specific information [4]. Emotion analysis of tapped telephonic conversation and surveillance video tapes of criminals or terrorists helps crime investigation departments to predict their extremist activities. Conversation with robotic pets and humanoid partners would be more realistic and enjoyable, if they are able to express and understand emotions in the same way as humans do [5,6].

Although several automatic emotion recognition systems have been explored the use of either facial expressions or speech, relatively few efforts have focused on emotion recognition using both modalities. It is known that the multimodal approach may give not only better performance, but also more

Published online: 24 July 2013

robustness when one of these modalities acquired in a noisy environment.

In this paper, we propose a bimodal emotion recognition system using acoustic and facial features. In this work, spectral and prosodic features are used to represent the emotion-specific information embedded in speech. Here, spectral features represent the variations in vocal tract shapes, movements and positioning of articulators. Prosodic features used in this work represent both global and local variations of duration, intonation and intensity patterns. Here, the term *global* refers to gross statistics of prosodic parameters, and the term *local* refers to time-varying or dynamic characteristics of prosodic parameters. Most of the existing multimodal ER systems use simple spectral and prosodic features for characterizing emotions from speech, and they have not considered the dynamic characteristics of prosodic parameters, because of the difficulties in their acquisition and modeling. But, human beings mostly exploit the dynamics of the prosody for experiencing emotions from spoken utterances [5]. With this reason, we have proposed the dynamic characteristics of prosody to explore the emotion-specific characteristics. The proposed dynamic characteristics of prosody are, in fact, the time-varying prosodic parameters, which emphasize the local variations of prosody with respect to time. In this work, for recognizing emotions from visual data, features derived from eye and mouth regions are explored. From human perspective also, it is observed that humans realize emotions mainly by focusing on the movements of eyes and mouth [5,7,8]. In this work, we are conducting systematic studies to explore the contribution of each of the proposed features from the acoustic and facial components of video data toward the recognition of emotions. In addition to the analysis of individual features, we also examine the complementary evidence provided by the proposed acoustic and facial features. Here, the term *complementary* refers to emotion-specific information captured by different feature sets are different, and hence, their combination may improve the recognition accuracy. Concurrently, we also investigate whether the individual components of acoustic features and facial features possess any complementary information for enhancing the discrimination of emotions.

For capturing the emotion-specific information from acoustic and facial features, autoassociative neural network (AANN) models [9,10] are explored in this work. The proposed AANN-based models are used at different levels for capturing the emotion-specific information. The main reason for choosing the neural network models for developing ER systems is that, they capture the complex nonlinear relations present in data. The tuning parameters such as the number of hidden layers, the number of units in each layer, the initial weights, the type of nonlinear function associated with each neuron and the number of iterations during training are not critical in this study, for achieving the reasonable performance.

The rest of the paper is organized as follows. Section 2 provides brief review of existing literature related to emotion recognition using acoustic and facial features. The description of the video database used for this study is provided in Sect. 3. The details of feature extraction process are discussed in Sect. 4. Details of the proposed neural network model are given in Sect. 5. Development and evaluation of the proposed emotion recognition models are discussed in Sects. 6 and 7, respectively. Section 8 discusses the advantages of the proposed features and their combinations for discriminating the emotions with respect to existing works. Conclusions of the paper and the future directions to extend the present work are provided in Sect. 9.

2 Related work

In this section, relevant existing works in the literature on emotion recognition using video are briefly discussed along the following dimensions: (i) facial features, (ii) features derived from speech (acoustic features) and (iii) combination of facial and acoustic features.

2.1 Facial features

Facial expression is one of the most important non-verbal cues for human communication. The Facial Action Coding System (FACS) is a commonly used method for determining facial expression [7]. In FACS, action units (AUs) are used as the intermediate features for understanding different facial feelings and expressions. A systematic review of attempts to develop automatic action unit recognition system is given in [11]. The recognition accuracy of the developed AU recognition system was about 87 %. Two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action (action unit) detection [8]. The existing approaches to facial expression analysis can be classified as geometric and appearance based. Systems based on these features are proposed in [12–17]. Spatial Gabor energy filters are proposed to represent facial expressions in computer vision applications, including face recognition and expression analysis [17]. Pose-invariant facial expression recognition is addressed in [16] using coupled Gaussian process regression model. Active appearance models are proposed in [15] for recognizing the pain expression. The accuracy of recognizing pain expression was found to be around 81 %. Affects and expressions are identified from face and body parts by using temporal dynamics [14]. The overall recognition accuracy achieved over 12 different classes was found to be about 78 % using both facial and body cues. Head

movements and facial expressions are used in [13] for laughter detection. The proposed audiovisual clues have achieved over 90 % recall rate and over 80 % precision. Spontaneous human behavior is analyzed in real time using automatic FACS [12]. In [18], bank of multilayer perceptron neural networks are used for the classification of six different facial expressions. In this study, logarithmic Gabor filters were applied to extract the features. The classification accuracy has been observed to be about 70 % using optimized subset of log-Gabor features. Filko and Martinovic [19] have proposed a system for human emotion recognition by analyzing key facial regions using principal component analysis and neural networks. In this work, 15 neural networks were used for building the overall system. Out of which, one was used for region detection and the other 14 are used to recognize 7 universal emotions over eyes and mouth regions. The average recognition accuracy was found to be about 70 % on the FEEDTUM database. Ioannou et al. [20] have designed a rule-based neuro-fuzzy system for recognizing the emotions using facial expression analysis. Facial animation parameters derived by robust facial analysis system are used as features for the proposed system. The proposed neuro-fuzzy system is evaluated for both discrete and continuous (2D activation–emotion) emotional spaces. Performance of this system is found to be around 78 %.

2.2 Acoustic features

Several approaches to recognize emotions from speech have been reported in the literature. A comprehensive review of these approaches can be found in [5]. Most researchers have used global prosodic features as acoustic cues for emotion recognition. Recognition accuracy of 77 % was observed for discriminating the neutral speech from emotional speech. Pitch-related features are used in [21], for classifying four emotions. Peaks and troughs in the profile of fundamental frequency and intensity and durations of pauses are explored in [22], for identifying four emotions: fear, anger, sadness and joy. In their work, the recognition accuracy was observed to be about 55 %. Prosodic and phonetic features are used in [23], for recognizing 8 emotions. The recognition performance was found to be about 50 %. Log frequency power coefficients (LFPC) are used for classifying the emotions using discrete hidden Markov model [24]. In this study, emotion recognition task was carried out with 7 emotions, and the recognition performance was found to be about 80 %. In addition to pitch-related information, log energy, formants and Mel frequency cepstral coefficients (MFCC) are used for classifying the emotions [25]. Fifty-five acoustic features comprising of 25 prosodic, 24 MFCCs and 6 formant frequencies are used for discriminating six emotions using Fisher's linear discriminate analysis [26]. The recognition accuracy was observed to be about 67.22 %, using both prosodic and spectral features together. Features extracted from glottal air flow signal are used in [27], for classifying the emotions using optimal path classifier. The recognition accuracy for four classes was found to be in between 55 and 67 % with different combinations of glottal and spectral parameters. Modulation spectral features are proposed in [28], for discriminating the emotions. Combination of modulation spectral features and prosodic features was found to increase the recognition accuracy to 91.6 % for discriminating seven emotions. Articulatory features are explored in [29], for recognizing emotions from speech. Nicholson et al. [30] have used neural networks for recognizing emotions from speech. In their work, separate neural network is used for each of the emotions. In this work, energy and pitch are used as prosodic features, and linear prediction coefficients and their deltas are used as spectral features or phonetic features for discriminating the emotions. The recognition accuracy is found to be about 50 % for discriminating 8 emotions. Rao and Koolagudi [31], Koolagudi and Rao [32] have explored neural network models on IITKGP-SESC and Berlin emotion databases for discriminating emotions. In their works, excitation source features represented by glottal pulse parameters and epoch parameters [31], and vocal tract parameters extracted from sub-syllabic regions [32] are used for capturing the emotion-specific information. The recognition accuracy is found to be in between 60 and 70 %.

2.3 Acoustic and facial features

Relatively, few efforts have focused on implementing emotion recognition systems using both acoustic and facial features [8,33–35]. Results of expression recognition from face and emotion recognition from speech are combined at higher level in [34], to predict the emotions effectively. The recognition accuracy was found to be 86, 67 and 91 % using facial, acoustic and facial plus acoustic features, respectively. A multimodal emotion recognition system proposed in [36] uses not only speech and visual information, but also thermal distribution acquired by infrared camera. A bimodal emotion recognition system to recognize six emotions has been proposed in [37], and it uses prosodic features from speech, and position and movement of facial organs from video for representing the emotion-specific information. The best features from both unimodal systems were used as input in the bimodal classifier. They showed that the performance is significantly increased from 69.4 % (video system) and 75 % (audio system) to 97.2 % (bimodal system). A bimodal emotion recognition system, proposed in [38], divides speech into energy and pitch features. These speech features are combined with motion features using multistream fused HMM.

3 Database

For developing the emotion recognition system in the present study, video data were collected using 20 subjects (10 males and 10 females), recorded in a studio environment. The emotions considered in this study are anger, fear, happy, neutral and sad. In each session, about 3–4 min of video data are collected for each emotion from each of the subjects. The total video data were collected in three sessions, and the sessions were separated by two-week interval. For expressing the emotions, the subjects were asked to speak the emotion-specific text in front of a camera. The subjects were given the choice to choose the appropriate text for expressing the desired emotions. All the subjects chose text from story books, novels, dramas and cinema stories. The text material was not common across the sessions and speakers. The number of sentences were not same across the emotions. The duration of video for each emotion was around 10 –12 min for each subject. In the database, sentences contained 3–10 words and words contained mostly 2–5 syllables. Altogether, the total number of sentences in the database was 6,530 (anger 1,720; fear 1,570; happy 1,140; neutral 1,230; and sad 870). In this study, all the subjects had chosen Hindi text for expressing the emotions. Video data were recorded at 30 frames per second, with a resolution of 640 $\times$ 480. The speech component corresponding to video was captured through a built-in microphone in video camera in both mono and stereo channels at 48 KHz sampling rate with 16-bit linear pulse code modulation format. Later, speech signal was down-sampled to 8 KHz, for extracting acoustic features. The distance between the video camera and the subject, lighting conditions and the background were maintained at the same level for all the recordings. After collecting the video database, it was evaluated by 25 research students, who had not participated in the database collection process. The average emotion recognition performance by human subjects was observed to be about 96, 98 and 100 % using speech, video without speech and video with speech, respectively.

In this work, two sessions of video data in each emotion were used for developing the models, and the remaining session was used for testing the models. The overall recognition performance was obtained by averaging the recognition performance across three sessions.

4 Feature extraction

Extraction of facial features from the video and acoustic (spectral and prosodic) features from the speech are discussed in Sects. 4.1 and 4.2, respectively.

4.1 Extraction of facial features

Human beings capture the emotions from the facial features, such as movement of head, eyebrows, fore head, eye, cheek and mouth regions. Among these features, eye and mouth regions are more significant in discriminating the emotions and easier to extract the features [35]. Therefore, in this work, we have considered the features derived from eye and mouth regions for discriminating the emotions. For supporting the above intuition, we have carried out some informal subjective emotion recognition tests by masking eye and mouth regions in video. From this study, it is observed that subjects felt difficulty in recognizing emotions by masking the eye and mouth regions. Subjects have marked the emotions randomly. The overall accuracy is observed to be less than chance level. From this study, we may conclude that visual clues from eye and mouth regions are crucial in recognizing the emotions from facial expressions. Figure 1 shows five frames of video, indicating four different emotions (anger, fear, happy, sad and neutral). In each frame, it is observed that most of the emotion-specific information is perceived from eye and mouth regions.

For extracting the features from eye and mouth regions, first, we need to estimate the face region from video frame sequence and then detect eye and mouth regions. In this work, motion information is used to track the face region in the sequence of video frames. Face regions are extracted from the upper head contour points, which are derived from the thresholded difference image. The details of deriving the

Fig. 1 Illustration of facial features for five emotions

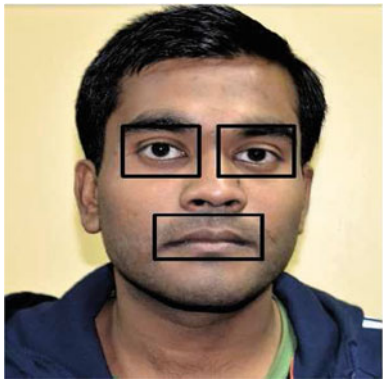

Fig. 2 *Rectangular bounding boxes* over eyes and mouth regions for extracting the features

thresholded difference image, head contour points and the bounding box to cover the face region are discussed in [39].

The eye regions have low intensity (Y), low red chrominance (C_r) and high blue chrominance (C_b) when compared to the forehead region of the face. Using this fact, the face region is thresholded using average Y, C_r and C_b values of the pixels in the forehead region. Morphological closing operation is applied to the thresholded face image, and the centroids of the blobs are estimated. The relative positions of the centroids with respect to the rectangular bounding box enclosing the face region and the contrast information in the eyebrow region are used to determine the locations of the eyes.

The mouth region is estimated from the locations of the eyes and the center of the mouth. The center of the mouth is estimated by modeling the color distribution of the non-lip region of the face using Gaussian distribution [39]. The non-lip regions are extracted relative to the locations of the eyes. The Y, C_r and C_b values of the pixels in these regions are used to estimate the parameters of the Gaussian distribution. The Y, C_r and C_b values of the pixels in the lip region may not fall into the distribution, and hence, the parameters of the Gaussian distribution are used to detect the pixels in the lip region.

The temporal dynamics of the gray values of the pixels of eye and mouth regions are captured using the gradient or local extreme in that region. The local maxima and minima are the largest and smallest intensity values of an image within some small local neighborhood, respectively. The key facial features such as hair, eyebrows, eyes, nostrils and end points of the lips are associated with local minima, and the shape of the lip contour and mouth region corresponds to local maxima. The local maxima and minima can be extracted using the gray scale morphological operations dilation and erosion, respectively [39]. In this work, rectangular rigid grids are placed over left eye, right eye and mouth regions. Figure 2 shows the marked rectangular bounding boxes over the eyes and mouth regions. Since the features of the eye regions are associated with local minima, the multiscale morphological erosion is used for feature extraction from the eyes. The multiscale morphological erosion operation is applied at each grid node for extracting the features from the eye region, as described below.

The multiscale morphological erosion operation is based on the gray scale morphology erosion. Let $\mathbb{Z}$ denote the set of integer numbers. Given an image $\mathbf{I} : \mathcal{D} \subseteq \mathbb{Z}^2 \longrightarrow \mathbb{Z}$ and a structuring function $\mathbf{G}_\sigma : \mathcal{G}_\sigma \subseteq \mathbb{Z}^2 \longrightarrow \mathbb{Z}$ at scale σ, the erosion of the image $\mathbf{I}$ by the structuring function $\mathbf{G}_\sigma$ is denoted as $(\mathbf{I} \ominus \mathbf{G}_\sigma)$, and it is defined by

$$(I \ominus G_\sigma)(i, j) = \min_{x,y}\{I(i + x, j + y) - G_\sigma(x, y)\} \tag{1}$$

where $-M_a \leq x, y \leq M_b$, with $1 \leq i \leq W, 1 \leq j \leq H$. The quantities W and H are the width and height of the eye region, respectively. The size of the structuring function is decided by the parameters M_a and M_b and is given by $(M_a + M_b + 1) \times (M_a + M_b + 1)$. The structuring functions such as flat, hemisphere and paraboloid are commonly used in morphological operations. The flat structuring function $G_\sigma(x, y) = 0$ is used in this work. For a flat structuring function, the expression for erosion reduces to

$$(I \ominus G_\sigma)(i, j) = \min_{x,y}\{I(i + x, j + y)\} \tag{2}$$

where $-M_a \leq x, y \leq M_b$. The erosion operation (2) is applied at each grid node for $\sigma = 1, 2, \ldots, P$ to obtain P feature vectors from the eye region. The height and width of the eye region are used to determine the parameters M_a, M_b and P. The values $M_a = \lfloor d/32 \rfloor + \lfloor (\sigma - 1)/2 \rfloor$, $M_b = \lfloor d/32 + 0.5 \rfloor + \lfloor \sigma/2 \rfloor$ and $P = 3$ have been used in our experiments. Here d is derived from W and H as $d = \frac{W+H}{2}$. These parameters are chosen in such a way that $M_a + M_b + 1$ for $\sigma = P$ is less than or equal to the minimal distance between two nodes of the grid. Each feature vector from the eye region $\mathbf{f} = (f_1, f_2, \ldots, f_{20})$ is normalized to $[-1, 1]$ as follows:

$$y_i = \frac{2(f_i - f_{\min})}{(f_{\max} - f_{\min})} - 1 \tag{3}$$

where $f_{\max}$ and $f_{\min}$ are the maximum and minimum values in the feature vector. The normalized facial feature vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_{20})$ is less sensitive to variation in the image brightness.

The features of the mouth region are associated with local maxima; therefore, multiscale morphological dilation is used for extracting the features from the mouth region. The features are extracted at each grid node using the multiscale morphological dilation operation as described below. Given an image $\mathbf{I} : \mathcal{D} \subseteq \mathbb{Z}^2 \longrightarrow \mathbb{Z}$ and a structuring function $\mathbf{G}_\sigma$: $\mathcal{G}_\sigma \subseteq \mathbb{Z}^2 \longrightarrow \mathbb{Z}$ at scale σ, the dilation of the image $\mathbf{I}$ by the structuring function $\mathbf{G}_\sigma$ is denoted as $(\mathbf{I} \oplus \mathbf{G}_\sigma)$, and it is

defined by

$$(I \oplus G_\sigma)(i, j) = \max_{x,y}\{I(i - x, j - y) + G_\sigma(x, y)\} \quad (4)$$

where $-M_a \leq x, y \leq M_b$, with $1 \leq i \leq W, 1 \leq j \leq H$. For a flat structuring function, the dilation can be expressed as

$$(I \oplus G_\sigma)(i, j) = \max_{x,y}\{I(i - x, j - y)\} \quad (5)$$

where $-M_a \leq x, y \leq M_b$. The dilation operation (5) is applied at each grid node for $\sigma = 1, 2, \ldots, P$ to obtain P feature vectors from the mouth image. The detailed descriptions of the multiscale morphological erosion and dilation operations are given in [39]. In this work, three rectangular grids are used for extracting the features from the left eye, right eye and mouth regions. The rectangular grids covering the eye regions consists of 20 nodes, and the positions of these nodes are determined with respect to the centroid of the eye. The nodes are uniformly spaced in four rows and five columns for covering the eye regions. The rectangular grid covering the mouth region consists of 35 nodes, placed in five rows and seven columns. From each video frame, three feature vectors are extracted for representing the left eye, right eye and mouth regions. The length of feature vectors of eye and mouth regions are 20 and 35 respectively, according to the number of nodes in their respective grids. The sequence of steps for extracting various facial features is given in Table 1.

4.2 Extraction of acoustic features

The emotion-specific information is present in speech at different levels. At the segmental level, emotion-specific information can be observed in the form of a unique sequence of shapes of the vocal tract for producing the sound units. The shape of the vocal tract is characterized by the spectral envelope. In this work, spectral envelope is represented by Mel frequency cepstral coefficients (MFCC). At the suprasegmental level, emotion-specific knowledge is embedded in duration patterns and the temporal variations of pitch and energy contours of the sequence of syllables. At the subsegmental level, emotion-specific information may be present in the shape of the glottal pulse and durations of open and close phases of vocal folds.

In this work, we have explored segmental and suprasegmental features for identification of emotions from speech. Usually, segmental features are extracted by analyzing the speech segments of duration 20–30 ms. Mostly, these features are extracted from the frequency spectrum of the speech segment. Hence, these features are known as spectral features. Suprasegmental features, also known as prosodic features, are extracted from speech segments of duration greater than 100 ms. In this work, 13 MFCC features are derived from a speech frame of 20 ms with a frame shift of 10 ms. For deriving the MFCCs, 24 filter bands are used. In this study, VOICEBOX: Speech processing toolbox for MATLAB has been used for extracting the MFCC features. The sequence of steps used for extracting MFCCs are given below:

(1) Pre-emphasize the speech signal.
(2) Divide the speech signal into a sequence of frames with a frame size of 20 ms and a shift of 10 ms. Apply the Hamming window over each of the frames.
(3) Compute magnitude spectrum for each windowed frame by applying DFT.
(4) Compute Mel spectrum by passing the DFT signal through Mel filter bank.
(5) Compute the desired MFCCs by applying the DCT over log Mel frequency coefficients (log Mel spectrum).

For deriving the prosodic features, speech data are segmented into phrases using the knowledge of longer pauses. The average phrase duration was observed to be about 2.5 s, and maximum and minimum phrase durations were observed to be 4.2 and 0.9 s respectively. The maximum number of syllables in a phrase was found to be 23. Therefore, the size of the duration vector was fixed to 23 dimensions indicating 23 duration values. If the number of syllables in a phrase is less than 23, then the tail portion of the duration vector was appended with zeros, to maintain the size of the duration vector to be 23. Syllable dura-

Table 1 Basic steps for deriving the facial features

1. Face regions are extracted from the upper head contour points, which are derived from the thresholded difference image.
2. Eye locations are estimated by applying the morphological closing operation to the thresholded face image. The face region is thresholded using the average Y, Cr and Cb values of the pixels in the forehead region.
3. The mouth region is estimated from the locations of the eyes and the center of the mouth. The center of the mouth is estimated by modeling the color distribution of the non-lip region of the face using Gaussian distribution.
4. The multiscale morphological erosion is used for extracting the features from the eye regions.
5. The multiscale morphological dilation is used for extracting the features from the mouth region.

Table 2 Basic steps for deriving the acoustic features

Spectral features
1. Thirteen-dimensional MFCC feature vectors are derived from the sequence of speech frames having a frame size of 20 ms and a shift of 10 ms.
Dynamic prosodic features
2. Duration contour is represented by sequence of durations of the syllables present in the sentence. Syllable boundaries are determined using VOPs.
3. Pitch and energy contours are derived from the sequence of speech frames having a frame size of 20 ms and a shift of 10 ms.
4. Pitch is computed by performing the autocorrelation on the Hilbert envelope of the LP residual signal.
5. Frame energies are computed by summing the squared sample amplitudes.
Global prosodic features.
6. Global prosodic parameters represents minimum, maximum, mean and standard deviation of frame-level prosodic features. Global prosodic feature vector is derived from each phrase using frame-level pitch and energy values, and syllable-level duration values.

tions are determined automatically, using vowel onset points [40].

The sequence of fundamental frequency values constitutes a pitch contour. In this work, pitch contours are extracted from speech using the autocorrelation of the Hilbert envelope of the linear prediction residual signal [41]. Energy contour of a speech signal is derived from the sequence of frame energies. Frame energies are computed by summing the squared sample amplitudes. In this study, we have chosen the frame size of 20 ms and a frame shift of 10 ms for extracting the pitch and energy values at frame level. A sequence of frame-level pitch and energy values constitute the pitch and energy contours. The size of pitch and energy contours of the phrases are proportion to the length of the utterance. To obtain the fixed dimensional feature vectors, resampling has been used. The dimension of pitch and energy contours was chosen to be 25. The choice of dimension 25 for pitch and energy contours is not crucial. The reduced size of pitch and energy contours has to be chosen such that the dynamics of the original contours are retained in the resampled versions. The basic reasons for reducing the dimensionality of the original pitch and energy contours are (i) need for the fixed dimensional input feature vector for developing the AANN models, and (ii) the number of feature vectors required for the training is proportion to the size of the feature vector.

For capturing the gross behavior of prosody over an utterance, gross statistical parameters are derived from the frame-level prosodic parameters of the utterance. The pitch and energy parameters at the gross (global) level are represented by maximum, minimum, mean and standard deviation of the frame-level values of the utterance. The duration parameters at the gross level are the number of syllables, and the maximum, minimum and mean durations of the syllables present in the utterance. A 12-dimensional feature vector was used to represent the prosodic information at the gross level. The sequence of steps for extracting various acoustic features from speech is given in Table 2.

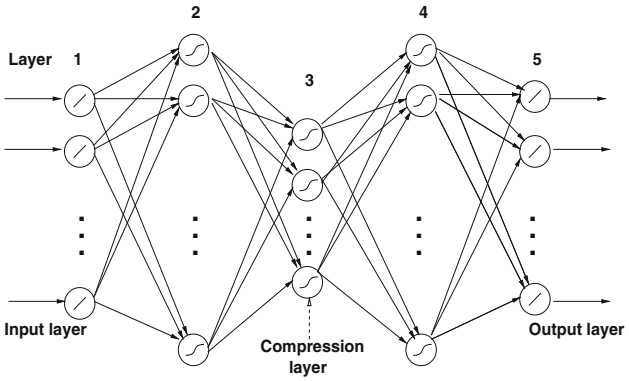

Fig. 3 Five-layer autoassociative neural network (AANN)

5 Autoassociative neural network (AANN)

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space and are used to capture the distribution of the input data [9]. The performance of AANN models can be interpreted in different ways, depending on the application and the input data. If the data constitute a set of feature vectors in the feature space, then the performance of AANN models can be interpreted as linear or nonlinear principal component analysis (PCA) or capturing the distribution of input data [42]. On the other hand, if the AANN is presented directly with signal samples, such as LP residual signal, the network captures the implicit linear/nonlinear relations among the samples [10].

In this work, a five-layer AANN model, as shown in Fig. 3, was used to capture the distribution of the feature vectors. The input and output (first and fifth) layers have the same number of units. The second and the fourth layers of the network have more units than the input layer. The third layer has fewer units than the input or output layers. The activation functions of second, third and fourth layers are nonlinear, whereas the first and the fifth (the input and the output) layers are linear. The nonlinear units use $\tanh(s)$ as

the activation function, where s is the activation value of that unit.

In this study, the network structures at different levels are arrived empirically. The performance of the network does not critically depends on the structure (the number of nodes in various layers of the network) of the network. In the literature, AANNs are used with symmetrical structure having the number of nodes in the outer and middle hidden layers is in the range of 1.4–2 times and 0.4–0.8 times the number of nodes of the input layer [39,10]. With this background, we have examined various network structures and finalized the appropriate structures suitable for the feature sets explored in this work. All the input and output features are normalized to the range [−1, +1] before presenting them to the neural network. The standard backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector [9].

6 Development of emotion models

In this work, autoassociative neural network models are explored for capturing the emotion-specific information from the facial features as well as acoustic features. These models acquire emotion-specific information by capturing the true distribution of feature vectors. The basic reason for choosing AANN model at different levels is that it will capture the distributions of feature vectors close to the actual (true) distributions present in data.

6.1 Emotion models using facial features

For each emotion, 3 models were developed using the feature vectors extracted from left eye, right eye and mouth regions. For training the models, the video data collected during sessions 1 and 2 were used. The number of feature vectors used for training the models was 216,000 (20 speakers × 2 sessions × 3 min × 60 s × 30 frames). The extracted feature vectors were given to both input and output of the respective AANN models. The reason for giving the feature vectors to input and output was to capture the distribution of the feature vectors. The structure of the AANN models used for capturing the distribution of feature vectors of eye and mouth regions were 20L 40N 10N 40N 20L and 35L 50N 15N 50N 35L, respectively. The integer value indicates the number of nodes present in the layer, and L and N indicates the linear and nonlinear activation functions of the nodes. The number of epochs needed for training depends on the behavior of the training error. It was found that 100 and 150 epochs are adequate for the AANN models correspond to the eye and the mouth regions, respectively. In this work, recognition performance of AANN models was analyzed with different hidden layers. It was observed that AANN models with 1 and 2 hidden layers have about 4.13 and 1.62 % less recognition accuracy compared to the proposed 3 hidden layer AANN models. Therefore, in the present work, AANNs with 3 hidden layers are considered for developing emotion models.

6.2 Emotion models using acoustic features

For each emotion, five AANN models were developed using (1) spectral features (MFCCs), (2) gross prosodic features, (3) durations of the sequence of syllables (duration contour), (4) sequence of pitch values (pitch contour) and (5) sequence of frame energies (energy contour). Models 3-5 represent the emotion models correspond to local prosodic information. For training the models, features extracted from the speech recorded during sessions 1 and 2 were used. The number of MFCC feature vectors used for training the models was about 500,000 (20 speakers × 2 sessions × 3 min × 60 s × 70 frames). Spectral feature vectors were extracted from speech frames for every 10 ms shift. Prosodic features were extracted at the phrase level (i.e., from each speech phrase, one feature vector will be derived). In this paper, the terms speech phrase and speech utterance were used interchangeably. The number of prosodic feature vectors used for training the models was about 2880 (20 speakers × 2 sessions × 3 min × 24 phrases). The structure of the AANN models correspond to spectral, gross prosodic, duration pattern, pitch contour and energy contour features are 13L 20N 7N 20N 13L, 12L 20N 6N 20N 12L, 23L 40N 10N 40N 23L, 25L 40N 15N 40N 25L and 25L 40N 15N 40N 25L, respectively. It was found that 75 epochs are adequate for training the AANN models using spectral and gross prosodic features, and about 100 epochs were found to be sufficient using local prosodic features.

7 Evaluation of the emotion models

The developed emotion-specific models are evaluated using the video data collected in the third session. The test data are segmented into pieces of 5, 10 and 15 s duration. In this study, the number of test video clips per emotion is 720 ((20 speakers × 3 min × 60 s)/(5 s)), 360 and 240 using 5-, 10- and 15s duration segments, respectively. The performance of ER systems using 10- and 15s test video clips is observed to be better compared to 5s video clips. Therefore, video clips of 10s duration are used in our evaluation studies.

In this work, we have eight emotion recognition (ER) systems at the first level correspond to the feature vectors extracted from speech and video. Out of the eight systems, five systems are developed using acoustic features (namely spectral (MFCC), gross prosodic, duration patterns of the sequence of syllables, pitch contour and energy contour), and three systems are developed using the features extracted

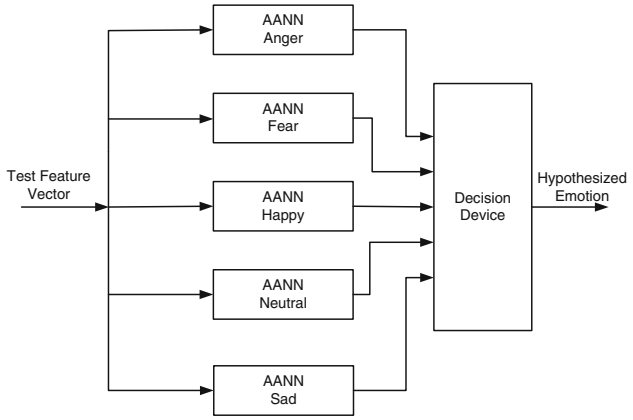

Fig. 4 Emotion recognition system (ERS) using AANN models

from left eye, right eye and mouth regions. Each ER system consists of 5 AANN models representing the five emotions: anger (A), fear (F), happy (H), neutral (N) and sad (S). The block diagram of the emotion recognition system using AANN models is shown in Fig. 4.

For evaluating the performance of the ER system (ERS), the feature vectors derived from the test video clips are given as input to five AANN models. The output of the each model is compared with the input to compute the normalized squared error. The normalized squared error (e) for the feature vector y is given by, $e = \frac{||y-o||^2}{||y||^2}$, where o is the output vector given by the model. The error e is transformed into a confidence score (c) using $c = \exp(-e)$. The average confidence score is calculated for each model. The category of the emotion is decided based on the highest confidence score.

In this work, first we analyzed the performance of the eight ER systems separately, and then they are combined in four phases.

(1) Phase 1:

   (a) Combining the ER systems developed using individual dynamic (local) prosodic features.
   (b) Combining the ER systems developed using the features extracted from left eye, right eye and mouth regions.

(2) Phase 2: Combining the ER system developed using gross prosodic features with the combined ER system developed using dynamic prosodic features in Phase 1(a).
(3) Phase 3: Combining the ER system developed using spectral features with the combined ER system obtained from Phase 2.
(4) Phase 4: Combining the fused ER system developed using facial features (obtained from Phase 1(b)) with the fused ER system developed using acoustic features (obtained from Phase 3).

All the four phases of combining the individual ER systems are shown in Fig. 5.

In this work, ER systems of different modalities are combined by summing the weighted confidence scores (evidences). The weighting rule for combining the confidence scores of individual modalities is as follows: $c^m = \frac{1}{m}\sum_{i=1}^{m} w_i c_i$, where c^m is the multimodal confidence score, w_i and c_i are weighting factor and confidence score of the ith modality, and m indicates the number of modalities used for combining the scores. In this study, each of the weights (w_i) is varied in steps of 0.01 from 0 to 1, and the sum of the weights should equal to unity ($\sum_{i=1}^{m} w_i = 1$). In this work, we have combined three modalities in Phase 1, and two modalities in Phases 2–4. By using the above weighting rule, for combining 2 modalities, 101 combinations of weights need to be explored, and for 3 modalities, 5151 combinations of weights have to be explored.

7.1 Performance of ERS using facial features

Performance of ER systems developed using the features extracted from eyes and mouth regions is given in Table 3. Columns 2–4 show the performance of ER systems developed by the features derived from left eye, right eye and mouth regions, respectively, and their average recognition performance is observed to be 76.30, 72.74 and 82.41 %. Column 5 shows the recognition performance by combining the confidence scores of the three ER systems developed using facial features (eyes and mouth). Here, the confidence scores of the individual systems are combined using the weighting rule mentioned above. It is observed that the best recognition performance is about 88.14 %, and the corresponding weighting factors associated with mouth-, left-eye- and right-eye-based ER systems are 0.62, 0.24 and 0.14, respectively. From the results, it is observed that the recognition performance has been improved in the combined system, compared to individual systems.

From the results, it is observed that average recognition accuracy using features extracted from left eye is 4 % higher than right eye. For analyzing the reasons to this variation, we have examined emotion recognition performance of all the subjects separately using features extracted from left eye and right eye. From the results, it is observed that for all the subjects, emotion recognition performance is better using the features extracted from left eye, compared to right eye. Among 20 subjects, about 17 subjects, the recognition accuracy is more than 3 % using features extracted from left eye, compared to right eye. To verify this fact, we have conducted a similar study on Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [43]. The recognition performance on IEMOCAP database using the features from left eye and right eye was observed to be similar. Among

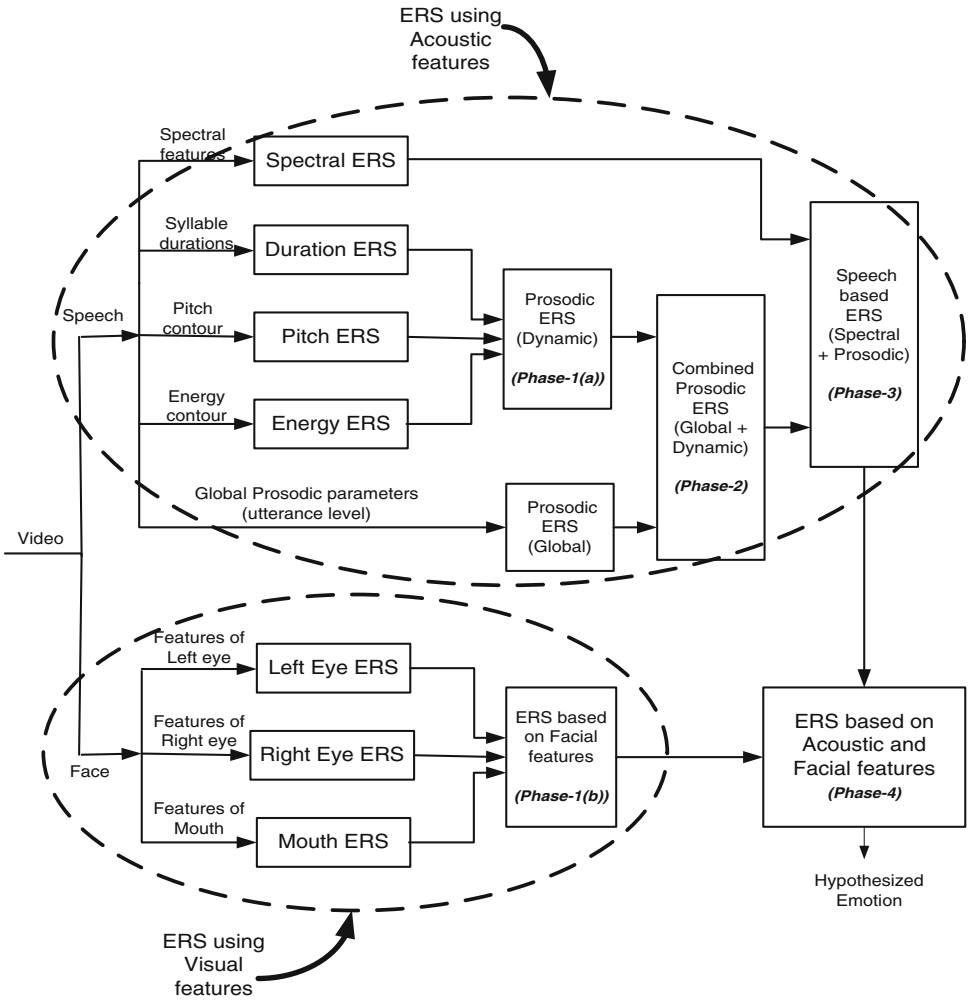

Fig. 5 Illustration of overall emotion recognition system (ERS) by combining the evidences at different phases

Table 3 Performance of the emotion recognition systems developed using facial features

Emotion	Emotion recognition performance (%)			
	Left eye (LE)	Right eye (RE)	Mouth (M)	LE+RE+M Phase-1(b)
Anger	81.16	76.07	78.73	86.72
Fear	73.69	71.34	82.87	87.91
Happy	78.41	72.35	80.62	92.78
Neutral	72.32	69.11	85.80	87.89
Sad	75.93	74.81	84.05	85.42

The weighting factors associated with the combined system are 0.14, 0.24 and 0.62 for the ER systems based on left eye, right eye and mouth features, respectively. Average emotion recognition performance of the combined system (Phase 1(b)) is 88.14 %

10 subjects of IEMOCAP database, for 8 subjects, the difference in the recognition performance is less than 2 %, for other 2 subjects the difference in the performance is found to be 2.27 % and 2.86 %. From these studies, we may conclude the variation in recognition accuracy between left eye and right eye is due to the presence of some recording artifacts in our database.

7.2 Performance of ERS using acoustic features

The following Sects. 7.2.1 and 7.2.2, discuss about the performance of ER systems developed using individual components of dynamic prosodic features (pitch, duration and energy contours), gross prosodic features, spectral features and their combinations.

7.2.1 Performance of ERS using dynamic prosodic features

The performance of ER systems developed using dynamic prosodic features extracted from the speech utterances is shown in Table 4. Columns 2–4 show the performance of ER systems developed by the features derived from duration contour, pitch contour and energy contour, respectively, and their average recognition performance is observed to be

Table 4 Performance of the emotion recognition systems developed using dynamic prosodic features

Emotion	Emotion recognition performance (%)			
	Duration (D)	Pitch (P)	Energy (E)	D+P+E (DP) Phase-1(a)
Anger	65.74	68.13	45.47	78.89
Fear	70.31	60.92	55.28	66.53
Happy	62.42	58.47	49.73	65.92
Neutral	76.17	73.28	57.88	84.27
Sad	71.83	67.35	54.16	80.58

The weighting factors associated with the combined system are 0.42, 0.35 and 0.23 for the ER systems developed using duration, pitch and energy contours, respectively. Average emotion recognition performance of the combined system (Phase 1(a)) is 75.24 %

Table 5 Performance of the emotion recognition systems developed using spectral and gross prosodic features

Emotion	Emotion recognition performance (%)	
	Spectral features (SP)	Gross prosodic features (GP)
Anger	72.38	72.16
Fear	76.83	64.95
Happy	68.91	65.29
Neutral	75.59	74.16
Sad	74.32	66.07

Average recognition performance of ER systems developed using spectral and gross prosodic features is 73.61 and 68.53 %, respectively

69.29, 65.63 and 52.50 %. Column 5 shows the recognition performance by combining the confidence scores of the three ER systems developed using duration, pitch and energy contours. It is observed that the best recognition performance of the combined system mentioned above is about 75.24 %, and the weighting factors associated with duration-, pitch- and energy-contour-based ER systems are 0.42, 0.35 and 0.23, respectively. From the results (see Column 5 in Table 4), it is observed that the recognition performance has improved for all the emotions, except for fear. The reason for decrease in recognition accuracy for fear may be due to the weights associated with the scores of individual ER systems. The weights may be optimal for achieving the highest overall recognition accuracy, but for fear they may not be optimal.

7.2.2 Performance of ERS using spectral and prosodic features

The performance of the ER systems developed using spectral and gross prosodic features is given in Table 5, and their average recognition performance is observed to be 73.61 and 68.53 %, respectively.

Table 6 Performance of combined emotion recognition systems developed using (i) gross and dynamic prosodic features (GP+DP: Phase 2) and (ii) spectral and prosodic features (SP+PR: Phase-3)

Emotion	Emotion recognition performance (%)	
	Prosodic (PR) Phase-2 (GP+DP)	Acoustic Phase-3 (SP+PR)
Anger	80.73	87.81
Fear	74.16	83.47
Happy	75.67	80.32
Neutral	86.42	91.65
Sad	82.26	85.28

Average recognition performances of ER systems based on overall prosodic (GP+DP: Phase-2) features and overall acoustic (SP+PR: Phase-3) features are 79.85 and 85.71 %, respectively

ER systems based on gross and dynamic prosodic features acquire the emotion-specific knowledge from the complementary features. One of the systems uses gross statistics of the prosodic parameters, and other one uses local as well as sequential knowledge of prosodic parameters. Therefore, combining the evidences of these two systems may improve the recognition performance. In this study, the evidences of gross- and dynamic-prosodic-feature-based ER systems is combined using the weighting rule. It is observed that the best recognition performance is about 79.85 %, and the weighting factors associated with ER systems developed using gross and dynamic prosodic features are 0.33 and 0.67, respectively. The recognition performance of the combined ER system developed using gross and dynamic prosodic features is shown in column 2 of Table 6. The results show that the recognition performance has been improved for all emotions by combining the gross and dynamic prosodic features.

The evidences of ER systems based on spectral and prosodic features can be viewed as complementary to each other. This is because, spectral features (i.e., MFCCs) are extracted by processing the speech frames of size 20 ms, whereas prosodic features are extracted from the whole speech utterance. Another basic difference between these two features is that spectral features characterize the emotions in the form of distinct shapes of the vocal tract, whereas prosodic features characterize the emotions in terms of distinct duration, intonation and energy patterns. Hence, combining the evidences of these two systems may enhance the recognition performance further. Evidences of the spectral and prosodic systems are combined using the weighting rule. The best performance is observed to be about 85.71 % for the weighting factors of 0.43 and 0.57 to spectral- and prosodic-feature-based ER systems, respectively. The recognition performance of combined ER system developed by using spectral and prosodic features is shown in column 3 of Table 6.

Table 7 Performance of the combined emotion recognition system by combining the evidences of individual ER systems developed using acoustic and facial features

Emotion	Emotion recognition performance (%)		
	Acoustic (Phase-3)	Facial (Phase-1(b))	Acoustic + Facial (Phase-4)
Anger	87.81	86.72	95.78
Fear	83.47	87.91	91.14
Happy	80.32	92.78	93.26
Neutral	91.65	87.89	94.09
Sad	85.28	85.42	93.88

Average emotion recognition performance of the combined (acoustic+facial) emotion recognition system is 93.62 %

From the results, it is observed that the recognition performance is enhanced for each emotion in the combined system.

7.3 Performance of ERS using acoustic and facial features

So far we have evaluated the recognition performance separately, for the ER systems developed using acoustic and facial features. Its known that the representation of these two modalities is entirely different. The speech signal in its basic form can be viewed as acoustic pressure variations with respect to time. A visual signal can be viewed as the intensity variations of the pixels with respect to time. The basic degradations in these two modalities are also independent. The variations in scaling, illumination, pose and motion may cause the degradations in visual information. For speech, the basic degradation is due to background noise and reverberation. By keenly observing these degradations, it can be noted that they are independent and non-overlapping. In other words, the degradation in speech will not show any effect in video. Similarly, the degradation of video does not affect the speech signal. Therefore, the combination of evidences of ER systems developed using acoustic and facial features will certainly improve the recognition performance and ensure the robust recognition performance even in the presence of degradations. Evidences of acoustic and facial ER systems are combined using weighting rule. The best performance is observed to be about 93.62 %, and the weighting factors associated with ER systems developed using acoustic and facial features are 0.38 and 0.62, respectively. Table 7 shows the recognition performance of the ER systems developed using acoustic features, facial features and the combination of the evidences of the ER systems developed using acoustic and facial features. The recognition performance of each of the emotions has been improved in the combined system compared to individual systems.

A summary of the recognition performance of individual and combined ER systems is shown in Table 8. From the results, it is observed that the evidences provided by the different feature sets are complimentary. Therefore, in each phase, the combination of evidences resulted in an improvement in the recognition performance. Within the modality also, the combination of various evidences (i.e., eyes and mouth in case of face; duration, pitch and energy in case of prosody; prosody and spectral in case of speech) provided enhancement in the performance. The improvement in recognition performance for various combinations mentioned in Table 8 was analyzed using hypothesis testing. The level of confidence was observed to be high (>99 %) for all cases.

Table 8 Summary of the performance of (i) individual ER systems and (ii) combined ER systems at different phases

Features used for developing ERS	Recognition performance (%)
Left eye	76.30
Right eye	72.74
Mouth	82.41
Duration contour	69.29
Pitch contour	65.63
Energy contour	52.50
Gross prosody	68.53
Spectral	73.61
Phase 1	
(i) Left eye + Right eye + Mouth (Facial features)	88.14
(ii) Duration contour + Pitch contour + Energy contour (Dynamic prosodic features)	75.24
Phase 2	
Dynamic prosodic features + Gross prosody (Prosodic features)	79.85
Phase 3	
Prosodic features + Spectral features (Acoustic features)	85.71
Phase 4	
Acoustic features + Facial features	93.62

Emotion recognition performance by the proposed AANN models is also compared with other popular classifiers such as hidden Markov models (HMMs), Gaussian mixture models (GMMs) and support vector machines (SVMs). In this study, HMM models are developed using 16 states, and each state is represented by 16 mixtures. GMMs used in this study are developed using 256 mixtures. In case of SVM models, Gaussian kernel with $\sigma = 0.035$ and C (penalty factor) = 25 values are used. Recognition performance using various classifiers considered in this work is given in Table 9. Among the four classifiers considered, performance of AANN models is observed to be superior over other models. Recognition performance using Gaussian mixture models is observed to be close to AANN models. The ability of capturing the dis-

Table 9 Performance of emotion recognition systems developed using autoassociative neural network (AANN) models, hidden Markov models (HMMs), Gaussian mixture models (GMMs) and support vector machines (SVMs)

Emotion	Emotion recognition performance (%)		
	Acoustic	Facial	Acoustic + Facial
AANN	85.71	88.14	93.62
HMM	74.36	83.18	87.29
GMM	82.76	87.95	92.38
SVM	82.08	85.32	89.15

Table 10 Statistical significance of comparative performance of AANN models with HMM, SVM and GMM models

Significance level (%)	Confidence intervals in terms of 10^{-4}		
	HMM	SVM	GMM
90	549–717	364–524	51–197
95	533–733	348–540	37–211
99	501–765	318–570	9–239
99.5	489–777	307–581	−1.3–249

tribution using GMM depends on the number of mixtures considered and amount of data used for modeling. The performance of HMM is observed to be inferior among the models considered. If we use other types of HMMs such as suprasegmental HMM, multistream HMM and discriminative HMM, the recognition performance may be comparable to the proposed AANN models. The performance of SVM models is slightly inferior compared to the GMM and the AANN models, because of non-optimal feature extraction methods. With GMM-based super vectors, SVMs may perform as good as GMMs and AANNs. The statistical significance of comparative performance of AANN models with other models is given in Table 10, in the form of confidence intervals (CIs). Here, CIs are mentioned in the orders of 10^{-4}. If the confidence interval includes zero, then the difference in recognition accuracy between the models is not significant. For analyzing the significance of difference in recognition accuracy between different pairs of models, the relative proximity of zero with respect to CIs is used. From the CIs and their associated significance levels, it is observed that the difference in recognition performance between GMM and AANN is not significant. On the other hand, difference in recognition performance between HMM and AANN, and SVM and AANN is found to be significant.

The proposed acoustic and facial features are also compared with some existing features for recognizing emotions. In this work, we have considered eigen vectors derived from eye and mouth regions for representing facial features, and linear prediction cepstral coefficients (LPCCs) derived from speech signal for representing spectral features for developing emotion recognition systems. Recognition performance using facial features represented by eigen vectors is observed to be around 84.62 %. Performance using spectral features represented by LPCCs is found to be 71.08 %. From the results, it is observed that recognition performance is superior in case of the proposed facial and spectral features, compared to the facial features represented by eigen vectors and spectral features represented by LPCCs. In both the cases, the improvement in performance using proposed features is statistically significant.

In this work, the proposed fusion technique is compared with one of the existing fusion techniques. The proposed fusion technique is based on combining the evidences from different modalities using appropriate weighting factors. The fusion technique considered for comparison is based on combining evidences from multiple modalities using highest score among the modalities. Recognition performance of the fusion technique based on highest score among the modalities is observed to be 83.92, 81.46 and 87.75 % while combining facial features, acoustic features and facial plus acoustic features, respectively. From this result, it is observed that the proposed fusion technique based on weighted combination of evidences of multiple modalities is better for improving the recognition performance than the fusion technique mentioned above.

7.4 Performance of ERS for subject-independent case

For studying subject-independent emotion recognition performance, ER systems are developed using the video clips of 15 subjects (7 males and 8 females). The developed ER systems are evaluated using the video clips of 5 subjects, which are not used for developing the models. The models are validated based on leave-one-speaker-out strategy. The recognition performance is evaluated on ER systems developed using (i) acoustic features, (ii) facial features and (iii) acoustic plus facial features. Table 11 shows the recognition performance of ER systems developed using acoustic features, facial features and acoustic plus facial features, where the speakers of the test video clips are different from the speakers of the video clips used for developing the models. The average recognition performance is observed to be 76.32, 83.50 and 89.56 % for the ER systems developed using acoustic features, facial features and acoustic plus facial features, respectively. The weighting factors of 0.37 and 0.63 are found to be optimal for combining the evidences of ER systems developed using acoustic and facial features, respectively. This result indicates that facial features are robust to subject variations for recognizing the emotions. On the whole, the performance of the subject-independent ER system is comparable to subject-dependent ER system. This indicates that the

Table 11 Performance of subject-independent emotion recognition systems developed using (i) acoustic features, (ii) facial features and (iii) acoustic and facial features

Emotion	Emotion recognition performance (%)		
	Acoustic	Facial	Acoustic + Facial
Anger	81.26	85.17	90.98
Fear	73.82	80.92	86.59
Happy	78.36	80.26	89.12
Neutral	72.18	89.35	92.83
Sad	75.97	81.78	88.26

Average emotion recognition performance using acoustic features is 76.32 %, facial features is 83.50 % and acoustic plus facial features is 89.56 %

Table 12 Performance of emotion recognition systems developed using (i) acoustic features, (ii) facial features and (iii) acoustic and facial features from IEMOCAP database

Emotion	Emotion recognition performance (%)		
	Acoustic	Facial	Acoustic + Facial
Anger	78.38	72.48	81.58
Happy	69.86	83.79	84.73
Neutral	71.92	69.73	78.93
Sad	73.68	86.78	88.74

Average emotion recognition performance using acoustic features is 73.46 %, facial features is 78.12 % and acoustic plus facial features is 83.50 %

proposed AANN models capture the emotion-specific information, independent of the speaker identity.

7.5 Evaluation of proposed features and models on Interactive Emotional Dyadic Motion Capture database (IEMOCAP)

In this work, the proposed features and models are evaluated on IEMOCAP database for analyzing their ability in recognition of real-life emotions. The IEMOCAP database consists of 12 h of audiovisual data collected from five male and five female actors [43]. In this database, the actors were asked to perform by use of plays (scripted sessions) and improvisation-based hypothetical scenarios (spontaneous sessions). The database was recorded in dyadic sessions in order to facilitate a more natural interaction and expression of the targeted emotion. The emotions present in data were evaluated by three evaluators. Each sentence has been tagged with the most appropriate emotional tag (e.g., happiness, sadness), according to the overall audiovisual impression of the sentence. The final emotional label of the sentence is decided by majority voting among evaluators. The present study examines sentences that are classified in four emotional states: angry, happy, neutral and sad. Since the number of sentences with fear emotion is very less in the database, we have not considered fear in this study. The recognition experiments were conducted in similar way as our earlier studies. Recognition models were developed using autoassociative neural networks. The recognition performance is evaluated on ER systems developed using (i) acoustic features, (ii) facial features and (iii) acoustic plus facial features. Table 12 shows the recognition performance of ER systems developed using aforementioned features. The average recognition performance is observed to be 73.46, 78.12 and 83.50 % for the ER systems developed using acoustic features, facial features and acoustic plus facial features, respectively. From the results, it is observed that the overall recognition accuracy by the proposed acoustic and facial features is about 6 % less with IEMOCAP database compared to our database used in this work. But the recognition trend seems to be similar to our database. The cause for lower recognition accuracy in case of IEMOCAP database may be due to the presence of realistic emotions, which may have some overlapping characteristics and difficult to discriminate from each other.

For comparing the emotion recognition accuracy of the proposed facial and acoustic features on IEMOCAP database, some of the existing works on the same database are given below. In [44], facial features extracted from six regions (right cheek, left cheek, right eye, left eye, chin, forehead) and acoustic features represented by Mel frequency cepstral coefficients and its derivatives are used for discriminating four emotions present in IEMOCAP database. In this work, emotion recognition is carried out in two stages. At the first stage, emotion recognition is performed separately from each of the seven modalities using Gaussian mixture models. At the second stage, these seven modalities (six face models and one acoustic model) are combined by using support vector machine classifier. The overall recognition accuracy is observed to be around 76 %. Emily Mower et al. [45] have proposed an emotion classification paradigm, based on emotion profiles. In this study, emotion profiles provide an assessment of the emotion content of an utterance in terms of a set of simple categorical emotions: anger; happiness; neutrality; and sadness. In this study, visual features are derived from motion capture markers placed on different facial regions, and audio features are represented by Mel frequency coefficients. The overall emotion recognition system consists four binary support vector machine classifiers followed by emotion profile building system and final emotion label decision-making system. The recognition accuracy on IEMOCAP database by this method is found to be 68.2 %. In [46], authors have explored context-sensitive schemes for emotion recognition. Here, the term context-sensitive refers to the emotional content of past and future observations

with respect to present observation. In this work, bidirectional Long Short-Term Memory (BLSTM) neural networks, hierarchical hidden Markov Model classifiers (HMMs) and hybrid HMM/BLSTM classifiers are explored for modeling emotion evolution within an utterance and between utterances. Overall experimental results have indicated that incorporating long-term temporal context is beneficial for improving the accuracy of emotion recognition systems. Among various systems, BLSTM neural-network-based system has achieved the highest recognition accuracy of about 78 % on IEMOCAP database. From the existing works on IEMOCAP database, it is observed that our proposed features and models have shown slightly better performance. The improvement in the performance may be due to effectiveness of proposed acoustic features and the capability of autoassociative neural networks for effective capturing of emotion-specific information.

8 Discussion

In this work, we have explored various facial and acoustic features for investigating the presence of emotion-specific information. Here, facial features are derived from only eyes and mouth regions. Feature extraction is not involved with any of the complex transformations. Simple morphological operations are used for the detection of eye and mouth regions and extraction of features from the detected regions. From the results, it is observed that the recognition performance is better for the features extracted from the mouth region compared to eye regions. This result also concurs with our intuition that during the expression of emotions, compared to eye regions, mouth region shows more specific clues with respect to emotions. From the classification performance of individual systems, it is observed that anger, happy and fear form one group and neutral and sad form the other group during classification by eye features. Here, group means classification/misclassification is limited to a particular set of emotions. Whereas in case of mouth features, we have not observed any grouping of emotions during classification. This observation confirms the presence of some non-overlapping emotion-specific information in the regions of eyes and mouth. Therefore, while combining these evidences, we have observed the improvement in the recognition performance. From the literature, it is observed that the facial features are extracted from the facial action units, spacial Gabor energy filters and active appearance models. All the above feature extraction methods involve complex nonlinear transformations, which are more computationally intensive. The performance of the proposed simple features is comparable to some of the existing works. The advantages of the proposed facial features are simple in nature, easy to extract, and the performance is also reasonable.

The acoustic features proposed in this work are based on spectral and prosodic characteristics of speech. Here, spectral features represent the unique sequence of vocal tract shapes specific to each of the emotions. For deriving this information, spectral features are extracted from the speech frame of size 20 ms with an overlap of 10 ms between the successive frames, whereas prosodic features usually represent long-term characteristics of speech such as duration, intonation and intensity patterns. Within this prosodic activity, we have investigated the global (gross)- and local (dynamic)-level prosody for discriminating the emotions. In general, prosodic features are extracted from the utterance or sentence levels. From the feature extraction point of view, spectral features represent the emotion-specific information at the frame level, and the prosodic features represent the emotion-specific information at longer speech segments such as words and sentences. This will ensure that the emotion-specific information captured by spectral and prosodic features is non-overlapping in nature, and hence, we will consider these features as complementary to each other.

Among the prosodic features, the global and the local prosodic features may capture different aspects of emotion. Here, the terms global and local refer to sentence and syllable or word levels, respectively. In this work, the local variations in prosody are termed as dynamic prosody. The three components of dynamic prosody (duration, pitch and energy contours) are also observed to be distinct to each other. Hence, their combination has enhanced the recognition performance. Among the three components of dynamic prosody, recognition performance using energy contour seems to be lowest. This is mainly due to similarity of energy contours present in broad groups of emotions such as active and passive. The energy contours carry discriminative information between active and passive groups of emotions. But, within a group (active or passive), the amount of discrimination across the emotions seems to be less.

Further, the combinations of (i) global and dynamic prosodic features and (ii) spectral and prosodic features have been achieved improved performance, compared to individual features. This is mainly due to non-overlapping emotion-specific information captured by individual features. In the literature, mostly, prosodic features are used as acoustic features for emotion recognition, and few attempts were made toward spectral features. In this work, we have proposed global and dynamic prosodic features for exploiting their complementary evidences to enhance the accuracy of emotion recognition. Here, complementary refers to presence of non-overlapping information offered by individual features. Additionally, we have combined spectral and prosodic features for exploiting their complementary characteristics to improve the emotion discrimination. In the literature, there is no systematic study on acoustic features in view of emotion recognition. Therefore, the study carried out in this paper will

provide clarity on contribution of various speech features in the context of emotion recognition.

For online and real-time applications, the proposed emotion recognition system may not be directly suitable, due to high computational time. Most of the time is consumed for extraction of acoustic features at different phases and sequence of score-level combination schemes. Among dynamic and gross prosodic features, dynamic prosodic features seem to be more promising compared to gross prosodic features. Dynamic prosodic features include both frame-level information and temporal information. For reducing the computational time, we may consider only dynamic features for representing emotion-specific prosodic information. By omitting gross prosodic features, recognition accuracy may not decrease significantly due to the fact that all frame-level features are preserved in dynamic prosodic features. Similarly, among facial features, either left or right eye features alone may be explored for reducing the computational time and make it suitable for real-time applications.

Finally, the combination of facial and acoustic features has shown significant improvement in the performance. This is due to the complementary nature of facial and acoustic features in the context of emotions. In this work, we have considered five basic emotions, which are collected in constrained studio environments. In real life, emotions are not limited to the basic emotions considered in this work. They are highly complex and influenced by several factors such as inherent attitude of the speaker, language, culture and society. The assumption of constant background and noise-free environment is hypothetical. Under the above situations, the characterization of the emotions from video is a challenging task.

9 Conclusion

From the results, it was clearly observed that the individual feature sets chosen in both modalities (speech and face) have indeed captured the emotion-specific information. Therefore, the average recognition performance has observed to be more than 50 % for all individual feature sets. The proposed feature sets from face (eyes and mouth) or speech (prosodic and spectral) were found to be complimentary within their respective modalities. Hence, by combining their evidences, the recognition performance was improved in all cases. Finally, by observing the combination of evidences from speech and face, we could conclude that the chosen feature sets from face and speech are complimentary to each other. Emotion recognition performance of the proposed AANN models is observed to be slightly better compared to GMM, HMM and SVM models. The proposed emotion recognition systems also demonstrated that their recognition performance was independent of speaker characteristics. The proposed features and models are also evaluated on real-life emotional database IEMOCAP (Interactive Emotional Dyadic Motion Capture). From the results, it is observed that the proposed features have followed similar trend in discriminating emotions in both databases. The recognition accuracy is slightly low in case of IEMOCAP compared to emotional database collected in this study. This may be due to difficulty in disambiguating the realistic emotions present in IEMOCAP database.

One can extend this basic work by considering the facial features from other regions such as eyebrows, head movement and movement of cheeks. Features extracted from the excitation source of speech signal can be explored for discriminating the emotions. Hybrid models can be explored at different phases for improving the recognition performance. The database can be further improved with respect to the number of subjects, the duration of recording in each session and relaxing the constraints for the suitability of real-time applications.

References

1. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. **13**, 293–303 (March 2005)
2. Schroder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S.: Acoustic correlates of emotion dimensions in view of speech synthesis. In: EUROSPEECH, Aalborg, Denmark (2001)
3. Pantic, M., Bartlett, M.: Machine analysis of facial expressions. In: Delac, K., Grgic, M. (eds.) Face Recognition, Vienna, pp. 377–416. I-Tech Education (2007)
4. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. J. Synth. Emot. **1**(1), 68–99 (2010)
5. Douglas-Cowie, R., Tsapatsoulis, E., Votsis, N., Kollias, G., Fellenz, S., Fellinge, W., Taylor, J.: Emotion recognition in human computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)
6. Turk, O., Schroder, M.: Evaluation of expressive speech synthesis with voice conversion and copy re-synthesis techniques. IEEE Trans. Speech Audio Process. **18**(5), 965–973 (2010)
7. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system: the manual (666 Malibu Drive, Salt Lake City UT 84107). A Human Face (2002)
8. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 39–58 (Jan 2009)
9. Yegnanarayana, B., Kishore, S.P.: AANN an alternative to GMM for pattern recognition. Neural Netw. **15**, 459–469 (Apr. 2002)
10. Rao, K.S.: Voice conversion by mapping the speaker-specific features using pitch synchronous approach. Comput. Speech Lang. **24**(1), 474–494 (2010)
11. Pantic, M., Patras, I.: Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans. Syst. Man Cybern. Part B **36**(2), 433–449 (2006)
12. Bartlett, M., Littlewort, G., Vural, E., Lee, K., Cetin, M., Ercil, A., Movellan1, J.: Data mining spontaneous facial behavior with automatic expression coding. In: Lecture Notes in Computer Science, pp. 1–21 (2008)

13. Petridis, S., Pantic, M.: Fusion of audio and visual cues for laughter detection. In: Proceedings of CIVR, pp. 329–338 (2008)
14. Gunes, H., Piccardi, M.: Automatic temporal segment detection and affect recognition from face and body display. IEEE Trans. Syst. Man Cybern. Part B Special Issue Hum. Comput. **39**, 64–84 (Feb. 2009)
15. Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Prkachin, K., Solomon, P.: The painful face: pain expression recognition using active appearance models. In: Proceedings of ICMI, pp. 1788–1796 (2007)
16. Rudovic, O., Patras, I., Pantic, M.: Coupled Gaussian process regression for pose-invariant facial expression recognition. In: Proceedings of 11th European Conference on Computer Vision (ECCV) (2010)
17. Wu, T., Bartlett, M.S., Movellan, J.R.: Facial expression recognition using Gabor motion energy filters. In: IEEE CVPR Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (2010)
18. Lajevardi, S.M., Lech, M.: Facial expression recognition using neural networks and log-Gabor filters. In: Digital Image Computing: Techniques and Applications (DICTA), pp. 77–83 (2008)
19. Filko, D., Martinovic, G.: Emotion recognition system by a neural network based facial expression analysis. AUTOMATIKA **54**, 263–272 (Aug. 2013)
20. Ioannou, S.V., Raouzaiou, A.T., Tzouvaras, V.A., Mailis, T.P., Karpouzis, K.C., Kollias, S.D.: Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural Netw. **18**(2), 423–435 (2005)
21. Busso, C., Lee, S., Narayanan, S.: Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Speech Audio Process. **17**, 582–596 (2009)
22. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S.: Approaching automatic recognition of emotion from voice: A rough benchmark. In: ISCA Workshop on Speech and Emotion, Belfast (2000)
23. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion recognition in speech using neural networks. In: 6th International Conference on Neural Information Processing (ICONIP), pp. 495–501 (1999)
24. Nwe, T.L., Foo, S.W., Silva, L.C.D.: Speech emotion recognition using hidden Markov models. Speech Commun. **41**, 603–623 (Nov. 2003)
25. Kwon, O., Chan, K., Hao, J., Lee, T.: Emotion recognition by speech signals. In: Eurospeech, Geneva, pp. 125–128 (2003)
26. Wang, Y., Guan, L.: An investigation of speech-based human emotion recognition. In: IEEE 6th Workshop on Multimedia Signal Processing, pp. 15–18 (2004)
27. Iliev, A.I., Scordilis, M.S., Papa, J.P., Falco, A.X.: Spoken emotion recognition through optimum-path forest classification using glottal features. Comput. Speech Lang. **24**(3), 445–460 (2010)
28. Wu, S., Falk, T.H., Chan, W.-Y.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2011)
29. Lee, C.M., Narayanan, S.: Toward detecting emotions in spoken dialogs. IEEE Trans. Audio Speech Lang. Process. **13**, 293–303 (March 2005)
30. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion recognition in speech using neural networks. Neural Comput. Appl. **9**, 290–296 (2000)
31. Rao, K.S., Koolagudi, S.G.: Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. **16**, 181–201 (2013)
32. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. Int. J. Speech Technol. **15**, 495–511 (2012)
33. Kanluan, I., Grimm, M., Kroschel, K.: Audio-visual emotion recognition using an emotion space concept. In: Proceedings of EUSIPCO (2008)
34. Datcu, D., Rothkrantz, L.J.M.: Semantic audio-visual data fusion for automatic emotion recognition. In: Proceedings of Euromedia (2008)
35. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human–computer interaction. In: Proceedings of the IEEE, vol. 91, pp. 1370–1390 (2003)
36. Yoshitomi, Y., Kim, S.I., Kawano, T., Kitazoe, T.: Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In: 9th International Workshop on Robot and Human Interactive Communication, pp. 178–184 (2000)
37. Huang, T.S., Chen, L.S., Tao, H., Miyasato, T., Nakatsu, R.: Bimodal emotion recognition by man and machine. In: ATR Workshop on Virtual communication environments, Kyoto, Japan (1998)
38. Zeng, Z., Tu, J., Pianfetti, B., Huang, T.S.: Audiovisual affective expression recognition through multistream fused HMM. IEEE Trans. Multimed. **10**(4), 570–577 (2008)
39. Palanivel, S., Yegnanarayana, B.: Multimodal person authentication using speech, face and visual speech. Comput. Vis. Image Underst. **109**, 44–55 (2008)
40. Vuppala, A.K., Yadav, J., Chakrabarti, S., Rao, K.S.: Vowel onset point detection for low bit rate coded speech. IEEE Trans. Audio Speech Lang. Process. **20**, 1894–1903 (Aug. 2012)
41. Prasanna, S.R.M., Yegnanarayana, B.: Extraction of pitch in adverse conditions. In: Proceedings of IEEE International Conference Acoustics, Speech, Signal Processing, Montreal, Canada (2004)
42. Ikbal, M.S., Misra, H., Yegnanarayana, B.: Analysis of autoassociative mapping neural networks. In: International Joint Conference Neural Networks, USA, pp. 854–858 (1999)
43. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: interactive emotional dyadic motion capture database. J. Lang. Resour. Eval. **42**, 335–359 (Dec. 2008)
44. Metallinou, A., Lee, S., Narayanan, S.S.: Audio-visual emotion recognition using Gaussian mixture models for face and voice. In: IEEE International Symposium on Multimedia (ISM), USA, Berkeley (2008)
45. Mower, E., Mataric, M.J., Narayanan, S.S.: A framework for automatic human emotion classification using emotional profiles. IEEE Trans. Audio Speech Lang. Process. **19**, 1057–1070 (May 2011)
46. Metallinou, A., Woellmer, M., Katsamanis, A., Eyben, F., Schuller, B., Narayanan, S.: Context-sensitive learning for enhanced audiovisual emotion classification. IEEE Trans. Affect. Comput. (TAC) **3**, 184–198 (April 2012)