Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-03529-7

ORIGINAL RESEARCH

Multimodal emotion recognition using SDA-LDA algorithm in video clips

Pradeep Tiwari[1,2] · Harshil Rathod[1] · Sakshee Thakkar[1] · A. D. Darji[2]

Received: 12 January 2021 / Accepted: 23 September 2021

Abstract
This paper focuses on Multimodal Emotion Recognition (MER) which can be conceptually perceived as the superset of Speech Emotion Recognition (SER) and Facial Emotion Recognition (FER). The challenges faced in designing the MER system are the extraction of the discriminative features. The features are strategically selected from speech and visual (facial) modalities and subsequently fused together. The base features extracted from the speech segment was Mel-Frequency Cepstral Coefficients (MFCC) while the Facial Landmarks Distances (FLD) was calculated from every frame of a video sequence as visual base features. Since these base features are static in nature, they hinder the performance of emotion recognition process. Hence, a novel algorithm of Shifted Delta Acceleration Linear Discriminant Analysis (SDA-LDA) has been proposed here to extract the most discriminative, dynamic and robust features of speech and video sequence. The Support Vector Classifier (SVM) was used for the classification of the emotions from fused multimodality features. The MER experiment was performed on the eNTERFACE, Surrey Audio-Visual Expressed Emotion (SAVEE), spontaneous BAUM-1s, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) databases, which resulted into accuracy of 95.65%, 100%, 100% and 93.3% respectively. The accuracies obtained for MER for all the 4 databases considered have outperformed the accuracies achieved for SER, FER and state-of-the-art techniques.

1 Introduction

Multimodal Human Machine Interface (HMI) has numerous applications in engineering, automobile, processing and other fields, such as: food processing, pharmaceutical production HMI has sought for decades to bestow machines with emotion identifying capabilities to provide more interactive and natural experiences (Turk 2014). Human emotions can be recognized from various modalities such as a person's facial expressions (Guo et al. 2018), speech (Wang et al. 2021; Tautkute et al. 2018), eye blinking (Soleymani et al. 2011) and posture (Yan et al. 2018), etc. Multimodal Emotion Recognition (MER) aims to design a system which can automatically recognize, understand and reflect human emotions. Since, MER is an interdisciplinary research area in the field of computer science, neuroscience, psychology and cognitive science (Huang et al. 2019), the exploration in this area becomes challenging. It is easy for human beings to perceive the world through comprehensive information provided by multiple sensory organs (Guo et al. 2019), nevertheless to endow machines with analogous cognitive capabilities is still an open question. Single-Modality deals with utilising one feature at a time which is insufficient to provide nonpareil emotion recognition results (Wu and Cao 2005). All subjects cannot accurately reflect all their emotions by single modality, for example a patient who is suffering from facial neuritis would fail to express

Harshil Rathod, Sakshee Thakkar and A. D. Darji contributed equally to this work.

✉ Pradeep Tiwari
pradeep.tiwari@nmims.edu

Harshil Rathod
harshilrathod14@gmail.com

Sakshee Thakkar
saksheethakkar@gmail.com

A. D. Darji
add@eced.svnit.ac.in

[1] Department of Electronics and Telecommunication Engineering, Mukesh Patel School of Technology Management and Engineering, NMIMS University, BS Marg, Mumbai 400056, Maharashtra, India

[2] Department of Electronics Engineering, Sardar Vallabhai National Institute of Technology, Ichchhanath Surat-Dumas, Road, Surat 395007, Gujrat, India

Published online: 04 October 2021

positive emotions through facial features (Petrantonakis and Hadjileontiadis 2011). Moreover, the presence of occlusions and changes might hinder the Facial Emotion Recognition (FER) method. Similarly, if the emotion needs to be recognized from speech, the ambient noise and the differences in the speeches of various subjects are significant factors that could perhaps affect Speech Emotion Recognition (SER) performance. MER is the technique which is used to identify the emotions through Facial and Speech Modalities, overcoming the limitations of using single modality. Thus, MER is deployed in the presented work.There are primarily three challenges in developing a robust MER system: identifying the most effective interclass distinguishable features from different modalities, fusing the features obtained from different modalities and finally develop a robust classifier model. The speech-based SER system aims at extracting features which show the least dependency on the speaker and the lexical content. The speech segment was used here to calculate Mel- Frequency Cepstral Coefficients (MFCC) as the base feature. The facial expression-based FER aims to determine the geometric features consisting of the shape of eyes, nose, lips etc. The facial landmarks were obtained from an image frame and the key points capable of depicting facial expressions were selected. Further, Facial Landmark Distances (FLD) features were computed using key points which would serve as base feature of an image (Noroozi et al. 2017). The information extracted up till now was static in nature and to improve the performance it is required to ameliorate its dynamic features. To extract dynamic features, the proposed methodology avails the Shifted Delta Acceleration (SDA) algorithm. SDA will help in retrieving all possible hidden dynamic features. It is a technique in which the blend of Shifted Acceleration Coefficient (SAC) and Shifted Delta Coefficients (SDC) is performed. These techniques were deployed to extract all possible delta and acceleration features and to stack them one behind the other. The real quandary takes place when stacking of the features increase the feature dimension. Since there is a fusion of speech and visual modalities the increase of dimensions may decrease the performance of MER due to overfitting. To overcome this drawback, the combination of SDA with Linear Discriminant Analysis (LDA) named here as SDA-LDA algorithm is proposed for obtaining modified differential features. To show the contribution of SDA-LDA feature extraction algorithm for fusion simple concatenation technique was adopted. Further Support Vector Machine (SVM) was employed, as it offers a principled approach to problems because of its mathematical foundation in statistical learning theory (Zhang et al. 2020). The databases used in the implementation were eNTERFACE, SAVEE, BAUM-1s, RAVDESS giving a sublime accuracy of 95.65%, 100%, 100%, 93.3% respectively (Martin et al. 2006; Jackson and Haq 2014; Zhalehpour et al. 2016; Livingstone and Russo 2018).

The remaining paper is organized as follows. In the Sect. 2, literature review is discussed. Section 3 explores the methodology used in the implementation of the MER system. Section 4 describes the implementation and the results of the proposed algorithm utilized for MER performance. Further, the paper is concluded in Sect. 5.

2 Literature review

In modern world, HMI certainly plays a significant role in our day-to-day life. In multidisciplinary research areas, automatic study and understanding of human emotions have drawn growing interest from researchers to achieve HMI (Wu et al. 2014). Further, recognised emotions could also play an important part in a wide range of applications such as affective computing, psychiatry, human computer interaction, educational software etc. The crucial steps in multimodal emotion recognition consist of two steps: feature extraction and multimodality fusion. The extensively used speech affective features can be classified into prosody features, voice quality features and spectral features (Li et al. 2015). The pitch period, speech rate, amplitude energy, etc. are classified as prosodic features, while the glottis parameters and formant frequency are known as voice quality features. The linear predictor coefficient, MFCC is called spectrum-based features. MFCC is the effective spectral feature since it can be used to model the human auditory perception system (Majeed et al. 2015). Similarly, FER features can also be categorized into two groups based on whether it uses static frame or video sequence. First, static frame-based FER which actually depends on static facial features collected from selected peak expression frames of images sequences by extracting handcrafted features. Second, spatio-temporal characteristics are used by dynamic video-based FER to record the patterns of expression in facial expression sequences (Kim et al. 2017). Further, multimodal data fusion was categorized into feature level fusion, score-level fusion, decision-level fusion etc. (Wang et al. 2012). Based on the eNTERFACE'05 and BAUM-1s database, Zhang et al. (2017), deployed the technique of producing audio-visual segment features with Convolutional Neural Networks (CNN) and 3D-CNN . The audio-visual segment features were fused in a Deep Belief Networks (DBNs) to obtain MER performance. Based on eNTERFACE and SAVEE, Noroozi et al. (2017) utilized MFCC, Filter Bank Energies for audio and facial landmarks for visual cues. The audio-video channels were fused by employing CNN to increase the performance of MER. Avots et al. (2019) utilized the eNTERFACE and SAVEE database for audio-visual emotion recognition. A cross-corpus evaluation was

made using MFCC for the audio part and faces in key frames were extracted using the Viola-Jones face recognition algorithm (Avots et al. 2019). Wu et al. (2020) proposed a two-stage fuzzy fusion based-CNN for dynamic emotion recognition. The two-stage fuzzy fusion strategy was developed by integrating canonical correlation analysis and fuzzy broad learning system (Wu et al. 2020). Haq et al. (2011) used the audio features such as pitch, energy, duration and MFCC features, whereas the visual features were related to positions of the 2D marker coordinates. Then Gaussian Classifier with Principal Component Analysis was also applied. Gera et al. (2014) surveyed the effects of changing the features and methods on the eNTERFACE'05 database. A 41-dimensional feature vector was computed from Pitch, Energy, the first thirteen MFCCs and their first and second derivatives for audio-based emotion recognition (Gera and Bhattacharya 2014). The face tracking and geometric features for facial emotion recognition was considered in the visual channel. Then the application of a multiclass SVM for classification was discussed. Wang et al. (2012) applied multimodal recognition system using the pitch, intensity, and the first 13 MFCC features on audio feature extraction tasks. For the facial feature extraction, a Gabor filter bank of 5 scales and 8 orientations was adopted to extract high-dimensional Gabor coefficients from each facial images. In recent years, CNN was used to extract facial features from static facial images. Venkataraman et al. (2019) used the Log-Mel Spectrogram, MFCC, pitch and energy features for audio modality with RAVDESS database. The significance of these features were compared by using Long Short Term Memory (LSTM), CNNs, Hidden Markov Models and Deep Neural Networks. Based on SAVEE database, Kim et al. (2017) proposed an Informed Segmentation and Labelling Approach (ISLA) that uses speech signals to alter the dynamics of the lower and upper face regions. Gharavian et al. (2017), recognized emotions by employing SAVEE database based on fuzzy ARTMAP Neural Network (FAMNN) and extracting features like MFCC, pitch, Zero Crossing Rate (ZCR) for the audio channel. Based on BAUM-1s database, Zhalehpour et al. (2016) extracted MFCC, and Relative Spectral Features (RASTA) from speech while for the facial expressions, Local Phase Quantization (LPQ) features and Patterns of Oriented Edge Magnitude (POEM) features were used. Livingstone et al. (2018) utilized the RAVDESS database in audio modality for testing Cohens and Kappa score. De Silva et al. (2000) applied a rule-based method for decision level fusion of speech and visual based systems with SAVEE database. The pitch was extracted and used in the near neighbour classification method in speech and the Hidden Markov Model (HMM) was trained as the classifier model for video. Issa et al. (2020) employed the MFCC, Chromagram, Mel-Scale spectrogram, Tonnetz representation, and spectral contrast features in the audio channel and 1-D CNN for identification of emotions. Thus, it is seen that various examiners have used different feature extraction techniques on numerous standard databases for MER. Hence, to exalt them this paper has focused on improving the performance of MER system in terms of accuracy by focusing on Facial Landmarks Distance (FLD) as the base feature for the visual cues and MFCC as the base feature of speech modality. Since these FLD and MFCC features are static and handcrafted features, thus to get dynamic and discriminative features, the paper proposed and implemented a novel SDA-LDA algorithm. From the above discussions it is observed that CNN is a widely applied algorithm for emotion recognition however, due to high complexity of facial expression images, CNN model usually requires high convolutional layers in order to extract desired set of features (Bengio 2009). The major drawback of increased network depth is the complexity of the network and the training time which can grow adequately with the addition of each layer. Hence, to avoid this difficulty a simple concatenation technique was employed. The SVM classifier was utilized for the classification. In this paper an attempt has been made and delivered to show significant performance improvement on eNTERFACE, SAVEE, BAUM-1s, RAVDESS databases defining the state-of-the-art.

3 Proposed MER system set-up

This section emphasizes on making the process as limpid as possible and hence, intricate attributes of the proposed MER system. The block schematic of the proposed MER system as shown in Fig. 1 illustrate the methodology of the system. The data acquisition, the extraction of facial and speech features, the proposed SDA-LDA algorithm and SVM classifier is described further in subsections.

3.1 Data acquisition

The first step in MER is Video (speech + video sequence) signal acquisition which is accomplished from the standard Audio-Visual database. The proposed MER system utilizes following databases: eNTERFACE, SAVEE, BAUM-1s and RAVDESS (Martin et al. 2006; Jackson and Haq 2014; Zhalehpour et al. 2016; Livingstone and Russo 2018) correspondingly. From the audio-visual database, the video sequences and speech content were segregated. The acquired video sequences comprise of frames at the rate of 25 frames per second, i.e. for example a video file of 3 seconds consists of 75 frames. These frames are utilized for extracting base visual FLD features that are static in nature. While the acquired speech signal may comprise of undesired information like silence, surrounding noise and Direct Current (DC) offset values introduced by the microphone while recording process. These unwanted information content in the acquired

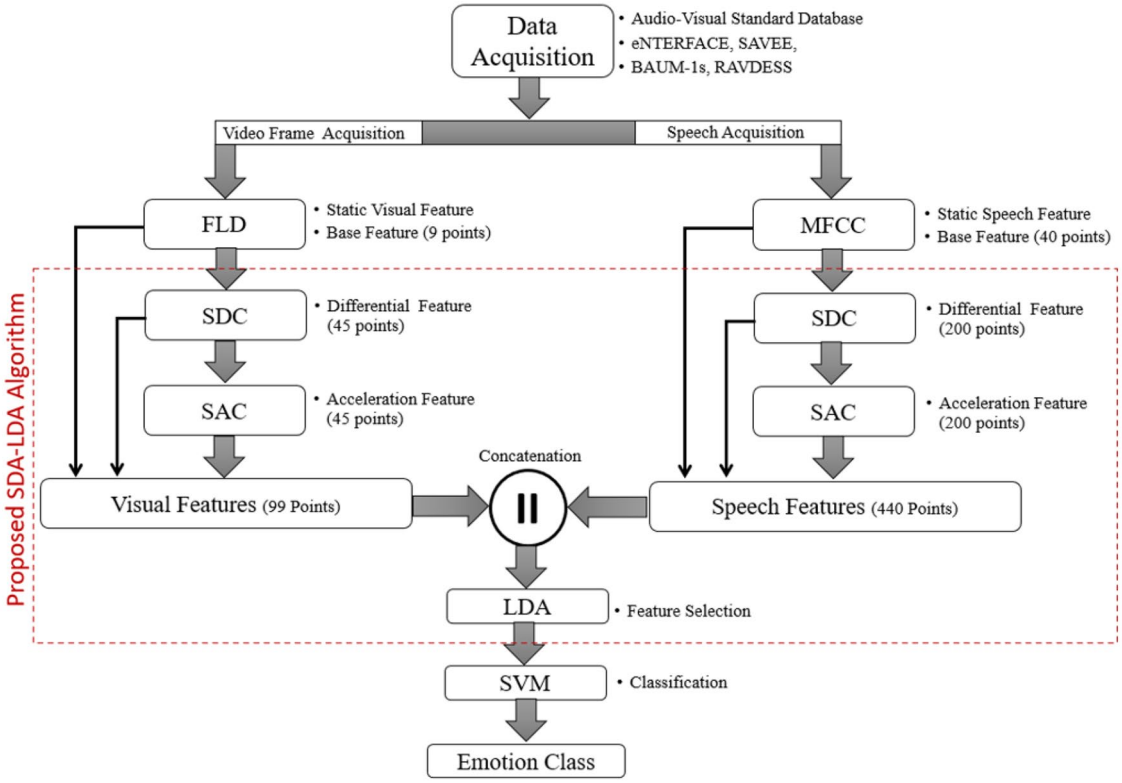


Fig. 1 The proposed MER system

speech signal is removed by applying pre-processing techniques such as normalization and pre-emphasis (Ibrahim et al. 2017). This filtered speech sequence was used further for MFCC feature extraction.

3.2 Base feature extraction

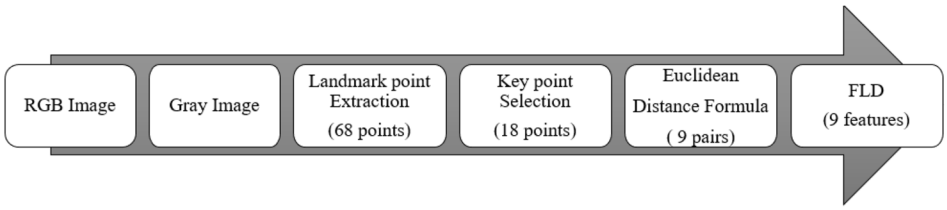
The images retrieved from videos consists of pixels arranged in two dimensional rows and columns, while the acquired speech is one-dimensional discrete values. Further, in visual and speech modalities, FLD and MFCC features were considered as base features respectively.

3.2.1 Visual FLD feature

The procedure followed for FLD feature extraction is as shown in Fig. 2. The frames acquired from the video file comprises of colour images, but the colour information doesn't contribute to the motive of emotion extraction hence, first the RGB images were converted into grey scale.

The next step is the assortment of features from the geometrical attributes such as the nose, the movement of the eyes, the displacement in the position of the lips, cheeks etc. The main categories of facial features are geometric and appearance-based features. The distances between two determined facial landmarks known as FLD features are used as the base feature of the visual modality (Ghimire and Lee 2013). Total 68-landmark points were obtained from an image frame as indicated in Fig. 3. These landmarks are extracted from a Dlib library which consists of an algorithm for the detection and extraction of facial landmarks (King 2009). The difficulty arises when similar facial expressions appear within the same video which represent

Fig. 2 FLD feature extraction



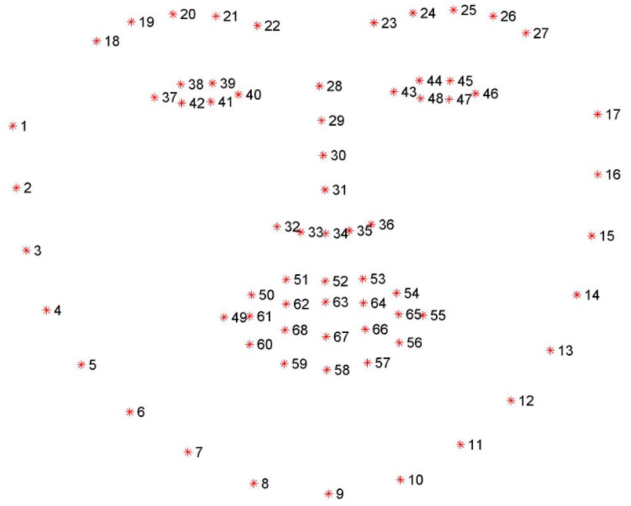


Fig. 3 Facial landmarks (68 points)

Table 1 Key points and their feature

Sr No	Feature	Landmark point pair
a	Lip and Eye	49–37
b	Lip and Eye	55–46
c	Right Eye	48–44
d	Left Eye	42–38
e	Outer Lip	55–49
f	Inner Lip	67–63
g	Lip and Nose	52–34
h	Eye to Eyebrow	38–20
i	Eye to Eyebrow	44–24

the same emotion hence, to overcome this limitation a set of key points were selected.

The 18 key points (9 pair) out of 68 facial landmark points were selected and the Euclidian's formula as given in Eq. (1) is applied to get 9 features strategically so that when displacement occurs the correct emotion is identified.

$$d(p_i, p_{i+1}) = \sqrt{\left(p_{i+1,x} - p_{i,x}\right)^2 + \left(p_{i+1,y} - p_{i,y}\right)^2} \tag{1}$$

In Eq. (1), $(p_{i,x}, p_{i,y})$ and $((p_{i+1,x}, (p_{i+1,y})$ represents the x and y coordinates of landmark points in pairs which gives the FLD $d(p_i, p_{i+1})$. Further on, Table 1 represent the selected landmark points in pair and corresponding features which were chosen on the basis of their movements and their effect on the emotion. For example, when a subject experience the disgust emotion, the tip of his nose tends to move a little and his lips tend to twitch. In Table 1, the movement of landmark points pair and the related features are tabulated. As all are aware, humans react uniquely to different emotions such as anger, disgust, surprise etc. Therefore, the points are selected in such a manner that, their displacement provides complete information about emotions that the subject might be experiencing. The points selected are shown in the Table 1 and the Fig. 4 gives us the pictorial depiction of the distances between the points. In Fig. 4, the part 4a is illustrating the FLD between the corner of the left side of the lip and the corner of the right side of the eye i.e. landmark point 49 and 37. Similarly, the part 4b is representing the FLD between the corner of the right side of the lip and the corner of the right side of the eye i.e. landmark point 55 and 46. Likewise, the part 4c is showing the FLD between the landmark 44 and 48 i.e. top and bottom of the right eye. The part 4d is displaying the FLD between the landmark 38 and 42 i.e. top and bottom of the left eye. Correspondingly part 4e is illustrating the FLD between the movement of the outer lip top and bottom lip i.e. landmark point 49 and 55. Identically the part 4f is showing the FLD between the inner lip top and bottom i.e. landmark points 67 and 63. Similarly, part 4g is depicting the lip and nose i.e. landmark points 52 and 34. The part 4h is indicating the FLD between left eye to left eyebrow i.e. landmark points 38 and 20. Finally, part 4i is representing the FLD between the right eye and right eyebrow i.e. landmark points 44 and 24. Thus, the above set of 9-FLD features were calculated.

3.2.2 Speech MFCC features

This is a widely used speech feature obtained with the multiplication of Mel filter bank with the Power spectrum of speech signal (Noroozi et al. 2017). The block diagram for MFCC feature extraction is as shown in Fig. 5.

Mel-scale (S_k) can be calculated for the given frequency f in Hz, with Eq. (2).

$$S_k = Mel(f) = 2595 \times (log_{10}(1 + \frac{f}{700})) \tag{2}$$

To extract perception based features, 128 triangular Mel filter banks are obtained using Eq. (2). Power Spectrum is the square of the absolute value of the discrete Fourier transform of the discrete-time speech input signal $x[n]$. If $x[n]$ is the input signal for discrete time instances n, then the short-time Fourier transform $X[k]$ with discrete frequency instances k for a frame of length N. Now, $X[k]^2$ is called the Power spectrum. If it is passed through 128 triangular filters of Mel frequency filter bank $H_m[k]$, Mel Scaled power spectrum $S[m]$ is obtained using Eq. (3).

$$S[m] = \sum_{k=0}^{N-1} X[k]^2 H_m[k], \quad 0 \le m \le M \tag{3}$$

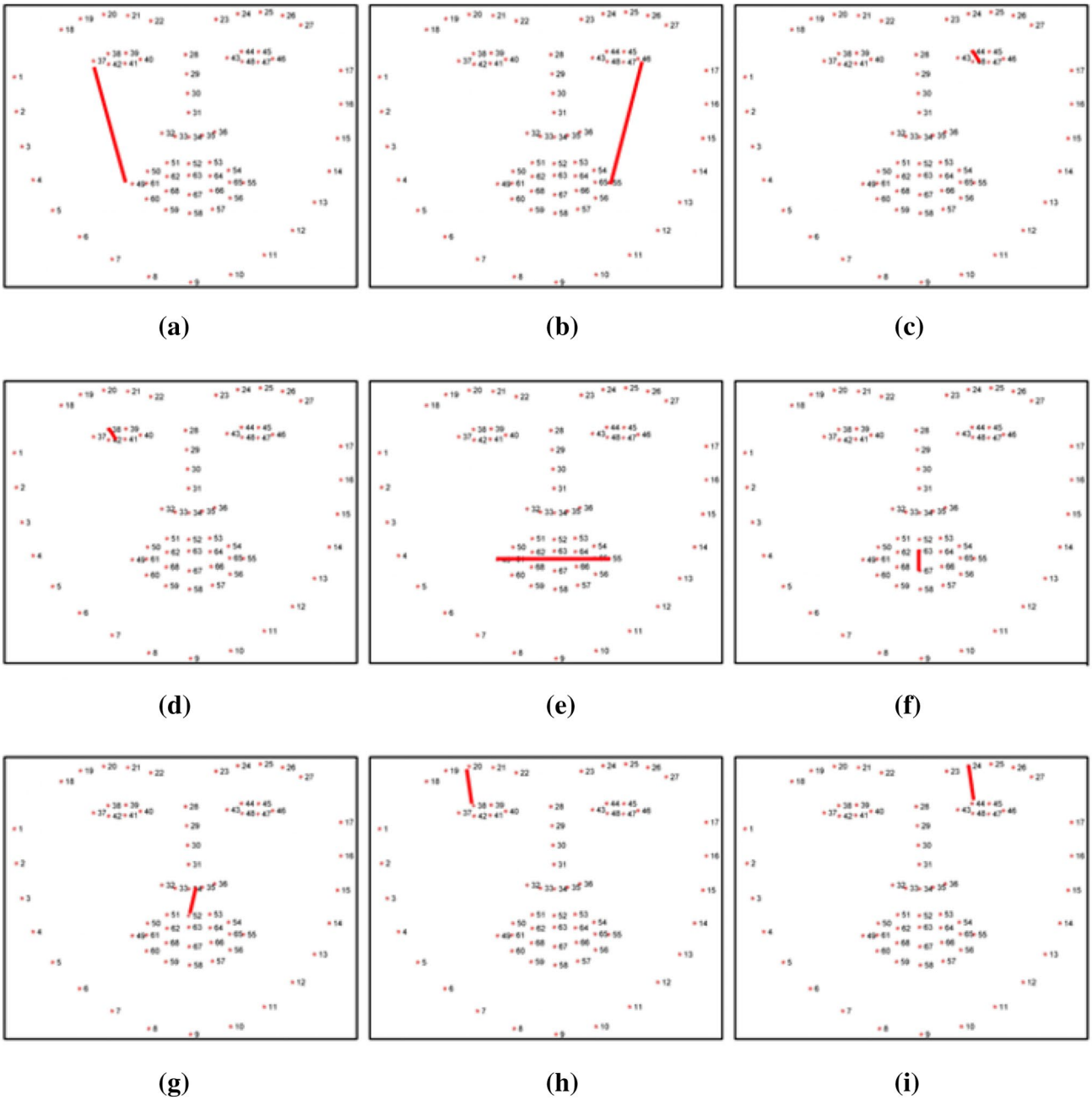


Fig. 4 Distance between the selected points

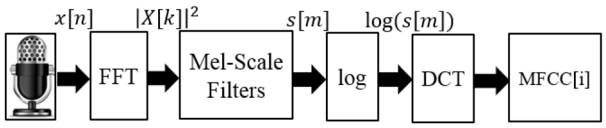


Fig. 5 MFCC feature extraction

MFCC is the log of mel power spectrum output $S[m]$ transformed back to the time domain by using a discrete cosine transform. The MFCC from $S[m]$ is given in Eq. (4).

$$MFCC[i] = \sum_{m=1}^{M} log(S[m]) \cos\left[i\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right] \quad i = 1, 2, \ldots, L \tag{4}$$

The value of 'L' represents the order of MFCC or MFCC coefficients for each frame whereas 'M' refers to the length

of the speech frames. In the present work, SER performance was analyzed for MFCC with L=40 because emotion-related information can also be found in the high frequencies. MFCC and pitch being the low-level features, it is desirable to modify the features further and improve the accuracy of SER.

3.2.3 Proposed SDA-LDA algorithm

The emotion-related information obtained using MFCC of speech modality and FLD of image modality are called base features. Since these features were calculated for consecutive speech and image frames thus, they gave static information of that particular frame (Wu and Cao 2005). However, the extra information about the temporal dynamics of the signal may be obtained by computing first and second derivatives of base features (Ekman 1977; Li and Huang 2014; Akçay and Oğuz 2020). The first and second-order derivatives of MFCC coefficients known as delta MFCC and acceleration MFCC respectively are generally used to get the dynamic MFCC feature vectors. Torres-Carrasquillo's proposed SDC which became a commonly used derived feature vector (Torres-Carrasquillo et al. 2002). The SDC function has far broader significance than the delta MFCC features because it collects additional differential information (Chittaragi and Koolagudi 2020). Thus, to get al.l possible differential features, SDC algorithms can play a very important role. SDC feature vectors are created by stacking delta MFCC (d_i) computed across multiple speech frames of a speech sample (Zhang et al. 2010). Further, this paper proposed the delta of SDC named as Shifted Acceleration Coefficients (SAC). Furthermore, SDC and SAC features are obtained from video sequence to collect their temporal dynamics. Let 'c_i' be the MFCC of *i th* frame of a speech file then the first-order derivative of MFCC referred as the delta MFCC coefficients can be expressed by d_i as given in Eq. (5) (Zhang et al. 2010).

$$d_i = \sum_{\tau=1}^{T} c_(i+\tau) - c_(i-\tau) \tag{5}$$

In Eq. (5), 'τ' is the frame delay and 'T' is the number of frames in a speech file. The second-order derivative of MFCC referred as the acceleration or delta-delta MFCC expressed as 'a_i'is defined in a similar way except that the input is the delta MFCC and the output is the acceleration MFCC coefficients.

The SDC which is expressed by δ_i is a stack of k-frames of the delta MFCC (d_i) as illustrated in Eq. (6) where 'k' is the number of frames being stacked with P amount of frame shift Majeed et al. (2015).

$$\delta_i = \begin{bmatrix} d_i \\ d_{i+P} \\ \cdots \\ d_{i+(k-1)P} \end{bmatrix} \tag{6}$$

Similarly, the proposed SAC which is expressed by 'α_i' is a stack of k-frames of the acceleration MFCC i.e. 'a_i', as mentioned in Eq. (7) where 'k' is the number of frames being stacked and 'P' is the amount of frame shift.

$$\alpha_i = \begin{bmatrix} a_i \\ a_{i+P} \\ \cdots \\ a_{i+(k-1)P} \end{bmatrix} \tag{7}$$

To calculate the SDC and SAC coefficients, the considered values for parameters are $\tau = 1, P = 3, k = 5$. Finally, the MFCC, SDC and SAC were stacked one behind another for obtaining the combined static, differential and acceleration information from MFCC named as Shifted Delta Acceleration SDA as shown in Eq. (8.

$$SDA = \begin{bmatrix} c_i \\ \delta_i \\ \alpha_i \end{bmatrix} \tag{8}$$

Since, MFCC resulted into 40 coefficients, SDC and SAC gave $40 \times 5 = 200$ coefficients each, i.e. in total a speech file gave 440 coefficients. Similarly, FLD features provided 9 coefficients, SDC and SAC yielded $9 \times 5 = 45$ coefficients individually, i.e. FLD computed 99 features from a video sequence. Further, after fusing speech and visual features using concatenation technique, total audio-visual feature set dimension contributed to 539 features. Here, 'll' indicates the concatenation of the features from speech and visual modality.

SDA provides all possible temporal, differentiating feature representatives for a speech and video sequence, however, the stacking of the delta and acceleration features would increase the feature dimension. Since there are multiple emotional classes, the increase in the feature dimension may decrease the MER recognition rate due to the overfitting problem. Thus, it is proposed here to combine the LDA with SDA algorithm for feature dimensionality reduction and overcome overfitting issues. LDA identifies a sub-space with suitable direction from the higher dimension of the data to maximize the inter-class separability. This is accomplished by finding a weight matrix W which maximizes the ratio of S_B and S_W where S_B given in Eq. (9) represents the between-class scatter and S_W given in equation (10) represents the with-in class scatter.

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{9}$$

In Eqs. (9) and (10), 'μ_i' represents the mean of class 'X_i', 'N_i' represents the number of samples in class X_i and 'c' represents number of classes (Ji and Ye 2008).

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \tag{10}$$

The 'W' is to be selected such that it maximizes the determinant of the ratio of $W^T S_B W$ and $W^T S_W W$. This W is called optimal matrix of W.

$$W_{opt} = \underset{x}{\text{argmax}} \frac{|W^T S_B W|}{|W^T S_W W|} \tag{11}$$

$w_i \mid i = 1, 2, \ldots, w_m$ is the set of eigenvectors of : $S_W^{-1} S_B$.

There are maximum $(c - 1)$ non-zero eigenvalues, so the upper bound of m is (c-1). Thus, it results in dimensionality reduction. Hence by combining SDA and LDA, a robust, discriminative feature vector SDA-LDA would be obtained. The Support Vector Classifier is then used as an emotion classifier.

3.2.4 SVM classifier

A pattern classifier called SVM has been used to classify the emotion class of the utterance (Li and Huang 2014). SVM has utilized linear kernel for generating feature representative models based on training vectors. The model developed is used for the recognition of emotions from the test speech.

4 Implementation and results

In this section, results of the implemented MER system are presented. Four databases were utilized for the implementation of the MER system namely eNTERFACE database, SAVEE database, BAUM-1s database, and RAVDESS database (Martin et al. 2006; Jackson and Haq 2014; Zhalehpour et al. 2016; Livingstone and Russo 2018). The execution of the MER system was divided into three elemental steps as follows:

1. Facial Expression based Emotion Recognition (FER)
2. Speech based Emotion Recognition (SER)
3. Multimodal Emotion Recognition (MER)

The FLD features forms a base in FER while, the MFCC features form a base in SER. For MER the FLD and MFCC features are combined. Then, to extract the dynamic features a novel SDA-LDA algorithm was proposed and implemented and the SVM classifier was applied. In essence, the brief expounding of each steps give rise to 3 cases which are *Case A*: The extraction of base feature and combining it with LDA. *Case B*: The extraction of base feature and combining it with SDC-LDA algorithm and *Case C*: The extraction of base feature and combining it with the proposed SDA-LDA algorithm. The implementation of MER system using these steps and the cases aroused as mentioned above is described in the following sections.

4.1 eNTERFACE database

The eNTERFACE database is a standard audio-visual emotion database which can be used for audio-visual emotion recognition (Martin et al. 2006). It contains six emotions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise which were acted by 42 English speaking subjects. For each emotion, there are five video samples provided in this database, hence the total number of video samples become $42 \times 6 \times 5$ i.e. 1260 video samples as shown in Table 2. In Table 2, column 2 is representing numerous classes of emotions, column 3 is highlighting the number of video files utilised in all classes of emotions and last column is giving information about abbreviations assigned to each classes of emotions. The speech files were extracted at the sampling rate of 48 kHz from video files. The video is recorded at a frame rate of 25 Frames Per Second (FPS), hence for a video sequence of 3-4 seconds will have approximately 75 to 100 frames per video sample to be processed. The recognition of emotion from this database is simplified into three steps, FER, SER and MER.

4.1.1 FER on eNTERFACE database

In the FER channel, if a video sequence with 'n' image frames exists, it will produce $n \times 9$ FLD features. Then the mean of 9 FLD features from each video sequence would form the base feature of FER channel. The FLD features are then combined with LDA, SDC and SDA to give the 3 cases as tabulated in Table 3. In Table 3, first column represents the different cases for evaluating the FER performance and second column shows the various features in these

Table 2 eNTERFACE database with 6 emotion classes

SR No	Emotion	No of video files	Abbreviations
1	Anger	210	A
2	Disgust	210	D
3	Fear	210	F
4	Happiness	210	H
5	Sadness	210	Sa
6	Surprise	210	Su

cases. Column three to column eight depicts the classes of emotions and the last column is giving the details about the achieved accuracy. According to Case A combining FLD+LDA has given an accuracy of 45.36%. When SDC was combined with FLD it showed almost 7% increase in the accuracy as represented in Case B with accuracy of 52.98%. When the proposed SDA-LDA algorithm was applied to the base feature of video sequence, around 16% increment in the accuracy of FER was noticed, as shown in Case C of Table 3 with total accuracy of 61.58%. Thus, it is observed that Case C using FLD + SDA-LDA gave the maximum accuracy of 52.98% as compared with Case A and Case B.

4.1.2 SER on eNTERFACE database

In Speech channel, initially the base features were extracted from speech frame, then the mean of all the frames were computed. If a speech file consists of 'k' frames it would give $k \times 40$ dimensions MFCC feature vectors. The mean of all frames gave 40 MFCC features from each speech file. The MFCC features are then combined with LDA, SDC and SDA and classified by using SVM classifier which yielded into three cases as represented in Table 4. In Table 4, first column represents the different cases for evaluating the SER performance and second column shows the various features in these cases. Column three to column eight depicts the classes of emotions and the last column is giving the details about the achieved accuracy. According to Case A, a linear transformation dimensionality reduction algorithm called LDA was applied on 40 MFCC features, that resulted into 5 dimensional features, thus combining MFCC+LDA provided an accuracy of 52.64%. When SDC was combined with MFCC and LDA, it showed an increase in the accuracy with 72.84% as shown in Case B of Table 4. When the proposed SDA-LDA feature was added to the base feature of speech sequence an increment in the accuracy was noticed, as also indicated in Case C of Table 4 with accuracy of 83.77%. Thus, it is observed that Case C using MFCC + SDA-LDA gave the maximum accuracy of 83.77% as compared with Case A and Case B.

4.1.3 MER on eNTERFACE database

In MER, initially the base features (MFCC || FLD) were extracted from speech frame and image frame, then the mean of all the frames were computed. Further, a concatenation technique was used to combine audio and visual base features to derive multimodal features. Furthermore, SVM was applied on these derived multimodal features for determining the accuracy in MER, as tabulated in Table 5. In Table 5, first column represents the different cases for evaluating the MER performance and second column shows the various features in these cases. Column three to column eight depicts the classes of emotions and the last column is giving the details about the achieved accuracy. In Case A as shown in Table 5, the base feature MFCC was combined with LDA which gave an accuracy of 65.23%. In Case B of Table 5, when MFCC and FLD was combined with SDC and LDA, a considerable increase of 19% in accuracy was observed. In the Case C, the base feature was combined with the proposed SDA-LDA algorithm provided an accuracy of 95.65%. Thus, it is observed that Case C using (MFCC || FLD) + SDA-LDA provided the maximum accuracy of 95.65% as compared with Case A and Case B.

Table 3 FER performance with eNTERFACE database

CASE	Features	A	D	F	H	SA	SU	Accuracy
Case A	FLD-LDA	48	42	40	55	41	41	45.36
Case B	FLD+SDC-LDA	48	64	44	66	45	47	52.98
Case C	FLD+ SDA-LDA	60	69	45	75	58	58	61.58

Table 4 SER performance with eNTERFACE database

CASE	Features	A	D	F	H	SA	SU	Accuracy
Case A	MFCC-LDA	59	59	49	51	54	44	52.64
Case B	MFCC+SDC-LDA	76	80	55	73	82	71	72.84
Case C	MFCC+ SDA-LDA	84	83	79	86	85	84	83.77

Table 5 MER performance with eNTERFACE database

CASE	Features	A	D	F	H	SA	SU	Accuracy
Case A	(MFCCFLD)-LDA	64	76	50	68	74	60	65.23
Case B	(MFCCFLD) +SDC-LDA	98	75	82	91	90	83	86.09
Case C	(MFCCFLD)+ SDA-LDA	97	100	96	92	90	96	95.65

Table 6 SAVEE database with 7 emotion classes

SR No	Emotion	No of video files	Abbreviations
1	Anger	60	A
2	Disgust	60	D
3	Fear	60	F
4	Happiness	60	H
5	Neutral	120	N
6	Sadness	60	Sa
7	Surprise	60	Su

From the above discussion, it is clear that the SDC-LDA has improved the performance of FER, SER and MER over the base features with LDA. The SDA-LDA algorithm has shown further increase in the accuracy in all the modalities. Also, it was observed that the performance of MER is higher than the performance of FER and SER. Thus, the comparison of all the 3 steps of modalities justify that the emotion recognition using multimodality is better than unimodality in eNTERFACE database.

4.2 SAVEE database

The SAVEE data-base comprises of 480 British English utterances recorded from 4 male actors with 7 emotions shown in Table 6 along with the file count for each emotion Jackson and Haq (2014). In table 6, column 2 is representing numerous classes of emotions, column 3 is highlighting the number of video files utilised in all classes of emotions and last column is giving information about abbreviations assigned to each classes of emotions.

4.2.1 FER on SAVEE database

Similar steps that were carried out on the eNTERFACE database were performed on the SAVEE database also. In the FER channel, a video sequence with 75 image frames had produced 75×9 FLD features. The mean of this feature set gave 9 FLD features from each video. These FLD features form the base feature of FER channel. The FLD features are then combined with LDA, SDC-LDA and SDA-LDA to give the 3 cases shown in Table 7. In Table 7, first column represents the different cases for evaluating the FER performance and second column shows the various features in these cases. Column three to column eight depicts the classes of emotions and the last column is giving the details about the achieved accuracy. According to Case A combining FLD+LDA has given an accuracy of 85.83%. When SDC was combined with FLD it showed an increase in the accuracy shown in Case B, giving the accuracy of 89.16%. When the proposed SDA-LDA component was added as given in Case C to FLD+SDC drastic increment in the accuracy was noticed with accuracy of 94.16%. Thus, it is observed that Case C using FLD + SDA-LDA contributed the maximum accuracy of 94.16% as compared with Case A and Case B.

4.2.2 SER on SAVEE database

In SER channel, initially the MFCC features were extracted from speech frame, then the mean of all the features for all the frames were taken. A speech file consists of 200 frames would give 200×40 dimension MFCC feature vector and the mean of 200 frames would give 40 MFCC features from each speech file. The MFCC features are then applied to LDA, SDC-LDA and SDA-LDA to give the 3 cases as represented in Table 8. According to Case A combining MFCC+LDA yields an accuracy of 86.66%. When SDC was combined with MFCC and LDA it showed an increase in the accuracy as shown in Case B, giving the accuracy of 96.66%. When the proposed SDA-LDA feature was applied to MFCC feature of speech sequence an increment in the accuracy was noticed, as shown in Case C with an accuracy of 99.18%. Thus, it is observed that Case C using MFCC + SDA-LDA presented the maximum accuracy of 99% as compared with Case A and Case B.

Table 7 FER performance for SAVEE database

CASE	Features	A	D	F	H	N	SA	SU	Accuracy
Case A	FLD-LDA	81	80	75	89	97	90	73	85.83
Case B	FLD+SDC-LDA	90	88	82	95	96	76	92	89.16
Case C	FLD+ SDA-LDA	95	92	83	100	97	100	86	94.16

Table 8 SER performance for SAVEE database

CASE	Features	A	D	F	H	N	SA	SU	Accuracy
Case A	MFCC-LDA	92	100	93	75	90	81	89	86.66
Case B	MFCC+SDC-LDA	90	100	100	92	100	90	100	96.66
Case C	MFCC +SDA-LDA	98	100	100	98	100	98	100	99

Table 9 MER performance for SAVEE database

CASE	Features	A	D	F	H	N	SA	SU	Accuracy
Case A	(MFCCFLD)-LDA	100	100	86	100	100	95	93	96.66
Case B	(MFCCFLD) +SDC-LDA	99	100	99	100	99	100	99	99.42
Case C	(MFCCFLD)+ SDA-LDA	100	100	100	100	100	100	100	100

Table 10 BAUM 1-s database with 6 emotion classes

SR No	Emotion	No. of Video files	Abbreviations
1	Anger	56	A
2	Disgust	79	D
3	Fear	37	F
4	Happiness	170	H
5	Sadness	134	Sa
6	Surprise	41	Su

4.2.3 MER on SAVEE database

In MER channel, initially the base features (MFCC||FLD) were extracted from speech frame and image frame, then the mean of all the frames were taken. Futher, a concatenation technique is used to give multimodal features. Furthermore, SVM was applied and the resultant MER accuracy of 96.66% was obtained as shown in Case A of Table 9. A considerable increase in the accuracy was observed when MFCC and FLD was combined with SDC-LDA resulting in an accuracy of 99.5% as mentioned in Case B. In the Case C, when the base feature was combined with SDA-LDA gave an accuracy of 100%. Thus, it is observed that Case C using MFCC + SDA-LDA provided the maximum accuracy of 100% as compared with Case A and Case B.

From the above obtained results its evident that just by using basic features is not sufficient to obtain high accuracies. An improvement is seen in the performance of the SDC-LDA algorithm over base features with LDA. Finally, it can be understood that the proposed algorithm SDA-LDA outperformed the base features and SDC algorithms performance in the multimodality scenario for SAVEE database.

4.3 BAUM-1s database

BAUM-1s is a Turkish database consists of a total of 517 video samples with 6 emotion classes: Anger, Disgust, Fear, Happiness, Sadness, Surprise as mentioned in Table 10 (Zhalehpour et al. 2016). In Table 10, column 2 is representing numerous classes of emotions, column 3 is highlighting the number of video files utilised in all classes of emotions and last column is giving information about abbreviations assigned to each classes of emotions. This database has been enacted by a total of 31 subjects of which 13 are female and 18 are male.

4.3.1 FER on BAUM-1s database

Similar steps that were carried out on the eNTERFACE and SAVEE database for emotion classes were performed on the BAUM-1s database. The FLD features form the base feature of FER channel. The FLD features were combined with LDA, SDC-LDA and SDA-LDA to give the three cases as shown in Table 11. According to Case A, the FLD+LDA gave an accuracy of 54.33%. When SDC-LDA was applied on FLD it showed an increase in the accuracy as described in Case B, giving the accuracy of 65.35%. When the proposed SDA-LDA algorithm was applied on FLD features, drastic increment in the accuracy was noticed, as mentioned in Case C of 11 with FER accuracy of 77.95%. Thus, it is observed that Case C using FLD + SDA-LDA provided the maximum accuracy of 100% as compared with Case A and Case B.

4.3.2 SER on BAUM-1s database

Here, initially the base features were extracted from speech frame, then the mean of the features extracted from the frames were calculated. If a speech file consists of 200 frames, it would give 200×40 dimension MFCC feature vector. Now the mean of all frame would give 40 MFCC feature from each speech file. These MFCC features were combined with LDA, SDC-LDA and SDA-LDA to give the 3 cases as shown in Table 12. According to Case A combining MFCC+LDA yields an accuracy of 59.05%. When SDC was combined with MFCC and LDA it showed an increase in the accuracy as shown in Case B, giving the accuracy

Table 11 FER performance for BAUM 1-s database

CASE	Features	A	D	F	H	SA	SU	Accuracy
Case A	FLD-LDA	57	38	40	63	55	35	54.33
Case B	FLD+SDC-LDA	43	40	86	75	62	83	65.35
Case C	FLD+ SDA-LDA	82	80	73	90	61	100	77.95

Table 12 SER performance for BAUM 1-s database

CASE	Features	A	D	F	H	SA	SU	Accuracy
Case A	MFCC-LDA	67	47	67	61	64	35	59.05
Case B	MFCC+SDC-LDA	100	94	100	76	93	80	87.40
Case C	MFCC +SDA-LDA	100	100	100	94	100	100	97.63

Table 13 MER performance for BAUM 1-s database

CASE	Features	A	D	F	H	SA	SU	Accuracy
Case A	(MFCCFLD)-LDA	31	50	50	85	86	62	69.29
Case B	(MFCCFLD) +SDC-LDA	100	89	88	93	89	100	92.91
Case C	(MFCCFLD)+SDA-LDA	100	100	100	100	100	100	100

of 87.40%. When the proposed SDA-LDA algorithm was applied to MFCC, base feature of speech sequence it caused an increment in the accuracy of SER to 97.63% as shown in Case C. Thus, it is observed that Case C using MFCC + SDA-LDA provided the maximum accuracy of 100% as compared with Case A and Case B.

4.3.3 MER on BAUM-1s database

For MER, initially the base features were extracted from speech frame and image frame, then the mean of the features obtained from all the frames were considered. Then a concatenation technique is used to give multimodal features. Furthermore, SVM was applied and the resultant MER accuracy of 69.29% was obtained as shown in Case A of Table 13. The table shows a considerable increase in the accuracy when SDC-LDA was applied on base features (MFCC||FLD) resulting in an accuracy of 92.91% as in Case B. In the Case C the base feature (MFCC||FLD) was combined with SDA-LDA features produced an accuracy of 100%.

From the above discussion, it is evident that the base features (MFCC and FLD) are insufficient to yield high accuracies. Similar to eNTERFACE and SAVEE database, proposed SDA-LDA algorithm has given highest MER accuracy for BAUM 1-s database also.

4.4 RAVDESS database

This database comprises a total of 1440 videos enacted by 24 subject (12 male and 12 female) and consists of 8 emotions: Angry, Calm, Disgust, Fearful, Happy, Neutral, Sad, Surprised Livingstone and Russo (2018). Each emotion consists of 192 videos except for neutral which has 96 videos as described in 14. In Table 14, column 2 is representing numerous classes of emotions, column 3 is highlighting the number of video files utilised in all classes of emotions and last column is providing information about abbreviations assigned to each classes of emotions.

Table 14 RAVDESS database with 8 emotion classes

SR no	Emotion	No. of video files	Abbreviations
1	Angry	192	A
2	Calm	192	C
3	Disgust	192	D
4	Fearful	192	F
5	Happy	192	H
6	Neutral	96	N
7	Sad	192	Sa
8	Surprise	192	Su

4.4.1 FER on RAVDESS database

Similar steps that were carried out on the eNTERFACE, SAVEE and BAUAM-1s databases were performed on the BAUM-1s database. 9 FLD features are extracted which form the base feature of FER channel. The FLD features are then combined with LDA, SDC and SDA to give the 3 cases shown in Table 15. According to CASE A combining FLD+LDA gives an accuracy of 57.22%. When SDA-LDA was applied with FLD an increase in the accuracy was observed as shown in CASE B, giving the accuracy of 64.16%. When the proposed SDA-LDA component was added to FLD drastic increment in the accuracy was noticed, as shown in CASE C with an accuracy of 68.05%.

4.4.2 SER on RAVDESS database

Here, 40 MFCC features were extracted from each speech file. The MFCC features were combined with LDA, SDC-LDA and SDA-LDA to give the 3 cases shown in 16. According to CASE A combining MFCC+LDA yields an accuracy of 53.05%. When SDC-LDA was combined with MFCC, it showed an increase in the accuracy as shown in CASE B, giving the accuracy of 64.72%. When the proposed SDA-LDA feature was added to base feature MFCC of

Table 15 FER performance for RAVDESS database

CASE	Features	A	C	D	F	H	N	SA	SU	Accuracy
Case A	FLD-LDA	75	44	73	54	67	65	43	44	57.22
Case B	FLD+SDC-LDA	79	67	59	52	75	71	44	77	64.16
Case C	FLD +SDA-LDA	71	71	67	70	76	77	57	62	68.05

Table 16 SER performance for RAVDESS database

CASE	Features	A	C	D	F	H	N	SA	SU	Accuracy
Case A	MFCC-LDA	63	52	52	61	42	54	39	60	53.05
Case B	MFCC+SDC-LDA	72	65	57	71	64	74	56	66	64.72
Case C	MFCC +SDA-LDA	83	74	76	76	88	74	81	68	77.77

Table 17 MER performance for RAVDESS database

CASE	Features	A	C	D	F	H	N	SA	SU	Accuracy
Case A	(MFCCFLD)-LDA	83	90	90	76	83	62	68	59	78.05
Case B	(MFCCFLD) +SDC-LDA	96	84	92	91	90	94	85	79	88.88
Case C	(MFCCFLD)+SDA-LDA	92	94	95	87	96	100	96	89	93.33

Table 18 Summarized FER, SER and MER performance using proposed SDA-LDA algorithm

Database	FER	SER	MER
eNTERFACE (Haq and Jackson 2011)	61.58	83.77	95.65
SAVEE (Hossain and Muhammad 2016)	94.16	99.00	100
BAUM-1s (Huang et al. 2019)	77.95	97.63	100
RAVDESS (Ibrahim et al. 2017)	68.05	77.77	93.33

speech sequence an increment in the accuracy was noticed, as shown in CASE C with total accuracy of 77.77%.

4.4.3 MER on RAVDESS database

Initially, the base features were extracted from speech frame and image frame, then the mean of all the frames were taken. Then a concatenation technique is used to give multimodal features. Furthermore, SVM was applied and the resultant MER accuracy was of 78.05% as shown in CASE A of Table 17. The table shows a considerable increase in the accuracy when MFCC and FLD was combined with SDC and LDA resulting in an accuracy of 88.88%. In the CASE C the base feature was combined with SDA-LDA giving a combined accuracy of 93.33%.

Similar to eNTERFACE, SAVEE and BAUM-1s database, proposed SDA-LDA algorithm has given highest MER accuracy for RAVDESS database also. The Table 18 illustrates the performance of FER, SER and MER in terms of its accuracies (in percentage) after the proposed SDA-LDA algorithm was applied.

4.4.4 Comparisons of performance of the present work and the state-of-the-art techniques

llowing required packages to your document preamble:

The results obtained in the previous section conveys that the proposed SDA-LDA algorithm outperforms SDC-LDA and LDA as well as that using multimodal features improves emotion classification. To evaluate the exact effectiveness of proposed method for all the emotion classes, the confusion matrix is obtained along with accuracy as described from Tables 19, 20, 21, 22, 23 , 24, 25, 26, 27, 28, 29 and 30. Since the databases considered for emotion recognition are imbalanced F_1-Score is also included in the last column

Table 19 MER using (MFCCFLD)-LDA features on eNTERFACE database

LDA	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Accuracy	F_1-Score
Anger	64	9	2	7	4	15	64	68
Disgust	5	76	3	10	2	3	76	75
Fear	9	9	50	7	16	9	50	53
Happiness	0	8	12	68	5	8	68	61
Sadness	0	4	10	4	77	6	74	74
Surprise	11	2	8	13	8	58	60	60

Table 20 MER using (MFCCFLD)+SDC-LDA features on eNTERFACE Database

SDC-LDA	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Accuracy	F_1-Score
Anger	98	0	2	0	0	0	98	89
Disgust	10	75	12	4	0	0	75	82
Fear	7	4	82	4	2	0	82	77
Happiness	2	0	7	91	0	0	91	90
Sadness	0	2	4	0	90	4	90	89
Surprise	3	2	3	2	8	83	83	89

Table 21 MER using (MFCCFLD)+SDA-LDA features on eNTERFACE database

SDA-LDA	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Accuracy	F_1-Score
Anger	97	3	0	0	0	0	97	97
Disgust	0	100	0	0	0	0	100	93
Fear	0	0	96	0	2	2	96	90
Happiness	4	0	2	92	0	2	92	100
Sadness	0	2	6	0	90	2	90	96
Surprise	0	0	2	0	2	96	96	95

Table 22 MER using (MFCCFLD)-LDA features on SAVEE database

LDA	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Accuracy	F_1-Score
Anger	100	0	0	0	0	0	0	100	100
Disgust	0	100	0	0	0	0	0	100	100
Fear	0	0	86	0	0	0	14	86	86
Happiness	0	0	0	100	0	0	0	100	100
Neutral	0	0	0	0	100	0	0	100	100
Sadness	0	0	0	0	5	95	0	95	95
Surprise	0	0	0	7	0	0	93	93	93

Table 23 MER using (MFCCFLD)+SDC-LDA features on SAVEE database

SDC-LDA	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Accuracy	F_1-Score
Anger	99	1	0	0	0	0	0	99	99
Disgust	0	100	0	0	0	0	0	100	100
Fear	0	0	99	0	0	0	1	99	99
Happiness	0	0	0	100	0	0	0	100	100
Neutral	0	0	0	0	99	1	0	99	99
Sadness	0	0	0	0	0	100	0	100	100
Surprise	0	0	1	0	0	0	99	99	99

Table 24 MER using (MFCCFLD)+SDA-LDA features on SAVEE database

SDC-LDA	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Accuracy	F_1-Score
Anger	100	0	0	0	0	0	0	100	100
Disgust	0	100	0	0	0	0	0	100	100
Fear	0	0	100	0	0	0	0	100	100
Happiness	0	0	0	100	0	0	0	100	100
Neutral	0	0	0	0	100	0	0	100	100
Sadness	0	0	0	0	0	100	0	100	100
Surprise	0	0	0	0	0	0	100	100	100

Table 25 MER using (MFCCFLD)-LDA features on BAUM-1s database

LDA	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Accuracy	F_1-Score
Anger	31	0	0	31	38	0	31	43
Disgust	10	50	10	5	25	0	50	50
Fear	0	0	50	11.2	38	0	50	57
Happiness	0	12	0	85	2	0	85	83
Sadness	0	11	0	0	86	8	86	74
Surprise	0	12	0	12	12	62	62	71

Table 26 MER using (MFCCFLD)+SDC-LDA features on BAUM-1s database

SDC-LDA	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Accuracy	F_1-Score
Anger	100	0	0	0	0	0	100	100
Disgust	0	89	0	6.6	6	0	89	89
Fear	0	0	88	0	0	12	88	88
Happiness	0	7	0	93	0	0	93	93
Sadness	0	3	0	5	89	3	89	89
Surprise	0	0	0	0	0	100	100	100

Table 27 MER using (MFCCFLD)+SDA-LDA features on BAUM-1s database

SDA-LDA	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Accuracy	F_1-Score
Anger	100	0	0	0	0	0	100	100
Disgust	0	100	0	0	0	0	100	100
Fear	0	0	100	0	0	0	100	100
Happiness	0	0	0	100	0	0	100	100
Sadness	0	0	0	0	100	0	100	100
Surprise	0	0	0	0	0	100	100	100

Table 28 MER using (MFCCFLD)-LDA features on RAVDESS database

LDA	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprise	Accuracy	F_1-Score
Angry	83	0	2	4	4	0	2	6	83	83
Calm	0	90	0	2	5	0	3	0	90	85
Disgust	4	0	90	0	2	0	0	4	90	87
Fearful	5	0	0	76	0	0	16	3	76	74
Happy	2	4	2	0	83	2	4	2	83	81
Neutral	0	23	0	0	0	62	8	8	62	71
Sad	2	6	4	8	4	4	68	4	68	69
Surprise	8	5	10	10	5	0	3	59	59	63

Table 29 MER using (MFCCFLD)+SDC-LDA features on RAVDESS database

SDC-LDA	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprise	Accuracy	F_1-Score
Angry	96	0	0	2	0	0	2	0	96	91
Calm	0	84	0	0	6	6	2	2	84	88
Disgust	2	0	92	2	0	0	4	0	92	92
Fearful	2	0	0	91	0	0	5	2	91	91
Happy	0	5	0	0	90	0	0	5	90	87
Neutral	0	0	0	0	6	94	0	0	94	86
Sad	4	4	2	2	2	2	85	0	85	85
Surprise	5	0	5	2	3	0	5	79	79	85

Table 30 MER using (MFCCFLD)+SDA-LDA features on RAVDESS database

SDA-LDA	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprise	Accuracy	F_1-Score
Angry	92	0	0	0	2	0	2	4	92	90
Calm	0	94	0	0	2	2	2	0	94	96
Disgust	5	0	95	0	0	0	0	0	95	96
Fearful	4	2	0	87	0	2	0	6	87	88
Happy	0	0	0	2	96	0	2	0	96	96
Neutral	0	0	0	0	0	100	0	0	100	95
Sad	2	0	0	2	0	0	96	0	96	95
Surprise	2	0	2	7	0	0	0	89	89	89

Table 31 Comparison of performance on eNTERFACE database using state-of-art-technique

Literature	Speech	Video sequence	Multimodal
Noroozi et al. (2017)	41.32	30.38	94.92
Zhang et al. (2017)	78.08	54.35	84.97
Zhalehpour et al. (2016)	72.95	42.26	75.78
Cornejo and Pedrini (2019)	62.00	68.33	85.00
Nguyen et al. (2017)	82.83	83.34	89.39
Hossain and Muhammad (2016)	64.04	58.38	85.06
Wu et al. (2020)	73.27	76.38	90.82
Present work	83.77	61.58	95.65

Table 32 Comparison of performance on SAVEE database using state-of-art-technique

Literature	Speech	Video sequence	Multimodal
Noroozi et al. (2017)	48.81	36.10	98.10
Gharavian et al. (2017)	63.10	93.75	97.92
Kim and Provost (2017)	74.32	81.61	86.01
Wu et al. (2020)	62.54	97.71	99.79
Present work	99.00	94.16	100

Table 33 Comparison of performance on BAUM 1-s database using state-of-art-technique

Literature	Speech	Video Sequence	Multimodal
Zhang et al. (2017)	42.26	50.11	54.57
Zhalehpour et al. (2016)	29.41	45.04	50.18
Cornejo and Pedrini (2019)	46.76	59.52	59.70
Present work	97.63	77.95	100

Table 34 Comparison of performance on RAVDESS database using state-of-art-technique

Literature	Speech	Video sequence	Multimodal
Livingstone and Russo (2018)	67	75	83
Siddiqui and Javaid (2020)	73.28	71.19	86.36
Present work	77.77	68.05	93.33

of the Tables 19, 20, 21, 22, 23 , 24, 25, 26, 27, 28, 29 and 30. Tables 19, 20 and 21 describes the confusion matrix, accuracy and F_1-Score on eNTERFACE database. Tables 22, 23 and 24 illustrates the the confusion matrix, accuracy and F_1-Score on SAVEE database. Further, Tables 25, 26 and 27 depicts the confusion matrix, accuracy and F_1-Score on BAUM-1s database, however Tables 28, 29 and 30 mentions the confusion matrix, accuracy and F_1-Score on RAVDESS database.

Thus, it is clear that the proposal and implementation of SDA-LDA in the different databases used has resulted in opulent accuracies defining the state-of-the-art techniques. The results of the previous esteemed literatures and the present work on the 4 databases used have been compared and tabulated further. The comparison is done on the basis of the features avail in each modality, the fusion techniques used and the classifier employed. These attributes are examined as they play a crucial role in the determination of the MER accuracy. The Table 31 shows the comparison of the eNTERFACE database. Table 32 encompasses the comparison for SAVEE database. Table 33 consists of the comparison of BAUM-1s database. Table 34 incorporates the RAVDESS database and its comparison with previous literatures.

In the work by Noroozi et al. (2017) and Kim et al. (2017) low-level handcrafted key features like MFCC, prosody and phoneme for speech and facial landmark features and upper and lower face landmarks have been used for video sequence these features are inadequate to distinguish between the different classes from speech and video frames hence the resultant accuracy is comparatively low. In the work by Zhang et al. (2017) and Zhalehpour et al. (2016) use Mel-Spectrogram, Relative Spectral Features (RASTA) based on Perpetual Linear Prediction have been used for speech channel and for video channel 3D-CNN technique and IntraFace+LPQ techniques have been

used respectively. Jadisha Yarif et al. (2019) proposed a hybrid deep CNN to extract audio and visual features along with PCA & LDA as feature reduction algorithm. Siddiqui et al. (2020) has utilized CNN for feature extraction and classification and the hybrid fusion approach. Nguyen et al. (2017) introduced a 3-dimensional CNN to extract the spatio-temporal information and deep-belief networks (DBNs). But these techniques presented in Cornejo and Pedrini (2019), Siddiqui and Javaid (2020) and Hossain and Muhammad (2016) has not focussed directly on extracting dynamic features. Hossain et al. (2016) have used derivative features by implementing multi-directional regression (MDR) features from speech and ridgelet transform based features from images. face image features. However, not used the acceleration features. Min Wu et al. (2020) presented the two-stage fuzzy fusion based-CNN for extracting dynamic and discriminative features but implemented on non-spontaneous databases. In the work by Venkataramanan et al. (2019) a 2D CNN technique with average pooling has been used for speech channel. However, the features extracted were static in nature and the present work is entirely focused on extracting and utilizing multilevel dynamic features extraction technique proposed SDA-LDA algorithm is applied to the base features of speech and video sequence. The comparison is accurate in most of the cases as the database and classifier are kept the same. As far as fusion techniques are considered simple concatenation technique has been applied to emphasise on the contribution of SDA-LDA dynamic feature extraction algorithm. The accuracy of SER modality has been comparatively high compared to the previous works considered, the values of the accuracies were 83.77% for eNTERFACE database, 100% for SAVEE database, 97.63% for BAUM-1s database and 77.77% for RAVDESS database. Similarly, the accuracies for FER in all the databases were also better with the figures of 61.58% for eNTERFACE database, 94.16% for SAVEE database, 77.95% for BAUM-1s database and 68.05% for RAVDESS database Hence, it has been persistently noticed that the accuracies of the present work are considerably high with the figures of 95.65% for MER modality in eNTERFACE database, 100% for MER modality in SAVEE database, 100% for MER modality in BAUM-1s database and 93.33% in RAVDESS database compared to the previous literatures. Thus, irrespective of the database and modality the SDA-LDA algorithm has given exemplary results in recognizing the different classes of emotion. Hence, it can be said with full confidence that the present work exalts the previous works that's are considered.

5 Conclusion

Multitudinous studies in the field of Human-Machine Interface (HMI), psychiatry and biomedical engineering have given a rise to recognition of emotions using numerous modalities. The Multimodal stratagem was used in the implementation of this paper as it was observed that multimodality is better than unimodality as it integrates complementary information of the unimodality cues. It was also realized that dynamic features played a crucial role in the enhancement of the accuracies. In this paper eNTERFACE, SAVEE, BAUM-1s and RAVDESS databases were used and the base features extracted were the FLD features from the video channel and MFCC features from the speech channel. Further, to procure high-level discriminative and dynamic features SDA-LDA algorithm was proposed and implemented. Hereafter, concatenation technique was employed to fuse the speech and video channels and SVM classifier was deployed for the classification of the emotions. The MER accuracies obtained for each database were 95.65% for eNTERFACE database, 100% for SAVEE database, 100% for BAUM-1s database and 93.33% for RAVDESS database. Hence, to conclude, the proposed SDA-LDA algorithm is exceptionally advantageous in defying the state-of-the-art techniques irrespective of the database and modalities used. Furthermore, the Electroencephalogram signals and various Deep Neural Methods can be amalgamated in the future work. DNN based algorithm for big data analytics in real time systems for MER can be analyzed. Numerous modalities like gestors and haptics can be fused with the present work.

References

Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76

Avots E, Sapiński T, Bachmann M et al (2019) Audiovisual emotion recognition in wild. Mach Vis Appl 30(5):975–985

Bengio Y (2009) Learning deep architectures for AI. Now Publishers Inc, Norwell

Chittaragi NB, Koolagudi SG (2020) Automatic dialect identification system for kannada language using single and ensemble svm algorithms. Lang Resour Eval 54(2):553–585

Cornejo JYR, Pedrini H (2019) Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform. In: 2019 IEEE international conference on systems, man and cybernetics (SMC), IEEE, pp 3396–3402

De Silva LC, Ng PC (2000) Bimodal emotion recognition. In: Proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), IEEE, pp 332–335

Ekman P (1977) Facial action coding system

Gera A, Bhattacharya A (2014) Emotion recognition from audio and visual data using f-score based fusion. In: Proceedings of the 1st IKDD conference on data sciences, pp 1–10

Gharavian D, Bejani M, Sheikhan M (2017) Audio-visual emotion recognition using fcbf feature selection method and particle swarm optimization for fuzzy artmap neural networks. Multimedia Tools Appl 76(2):2331–2352

Ghimire D, Lee J (2013) Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors 13(6):7714–7734

Guo J, Lei Z, Wan J et al (2018) Dominant and complementary emotion recognition from still images of faces. IEEE Access 6:26391–26403

Guo W, Wang J, Wang S (2019) Deep multimodal representation learning: a survey. IEEE Access 7:63373–63394

Haq S, Jackson PJ (2011) Multimodal emotion recognition. In: Machine audition: principles, algorithms and systems. IGI Global, p 398–423

Hossain MS, Muhammad G (2016) Audio-visual emotion recognition using multi-directional regression and ridgelet transform. J Multimodal User Interfaces 10(4):325–333

Huang H, Hu Z, Wang W et al (2019) Multimodal emotion recognition based on ensemble convolutional neural network. IEEE Access 8:3265–3271

Ibrahim YA, Odiketa JC, Ibiyemi TS (2017) Preprocessing technique in automatic speech recognition for human computer interaction: an overview. Ann Comput Sci Ser 15(1):186–191

Issa D, Demirci MF, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control 59(101):894

Jackson P, Haq S (2014) Surrey audio-visual expressed emotion (savee) database. University of Surrey, Guildford

Ji S, Ye J (2008) Generalized linear discriminant analysis: a unified framework and efficient model selection. IEEE Trans Neural Netw 19(10):1768–1782

Kim DH, Baddar WJ, Jang J et al (2017) Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans Affect Comput 10(2):223–236

Kim Y, Provost EM (2017) Isla: temporal segmentation and labeling for audio-visual emotion recognition. IEEE Trans Affect Comput 10(2):196–208

King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10:1755–1758

Li Y, Chao L, Liu Y, et al (2015) From simulated speech to natural speech, what are the robust features for emotion recognition? In: 2015 international conference on affective computing and intelligent interaction (ACII), IEEE, pp 368–373

Li Z, Huang C (2014) Key technologies in practical speech emotion recognition. J Data Acquisit Process 29(2):157–170

Livingstone SR, Russo FA (2018) The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north american english. PloS One 13(5):e0196391

Majeed SA, Husain H, Samad SA et al (2015) Mel frequency cepstral coefficients (mfcc) feature extraction enhancement in the application of speech recognition: a comparison study. J Theor Appl Inf Technol 79(1):38

Martin O, Kotsia I, Macq B et al (2006) The enterface'05 audio-visual emotion database. In: 22nd international conference on data engineering workshops (ICDEW'06), IEEE, pp 8

Nguyen D, Nguyen K, Sridharan S et al (2017) Deep spatio-temporal features for multimodal emotion recognition. In: 2017 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1215–1223

Noroozi F, Marjanovic M, Njegus A et al (2017) Audio-visual emotion recognition in video clips. IEEE Trans Affect Comput 10(1):60–75

Petrantonakis PC, Hadjileontiadis LJ (2011) A novel emotion elicitation index using frontal brain asymmetry for enhanced eeg-based emotion recognition. IEEE Trans Inf Technol Biomed 15(5):737–746

Siddiqui MFH, Javaid AY (2020) A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images. Multimodal Technol Interaction 4(3):46

Soleymani M, Pantic M, Pun T (2011) Multimodal emotion recognition in response to videos. IEEE Trans Affect Comput 3(2):211–223

Tautkute I, Trzcinski T, Bielski A (2018) I know how you feel: Emotion recognition with facial landmarks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1878–1880

Torres-Carrasquillo PA, Singer E, Kohler MA et al (2002) Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In: Seventh international conference on spoken language processing

Turk M (2014) Multimodal interaction: a review. Pattern Recogn Lett 36:189–195

Venkataramanan K, Rajamohan HR (2019) Emotion recognition from speech. arXiv preprinthttp://arxiv.org/abs/1912.10458

Wang W, Chen J, Zhang Y et al (2021) A multi-graph convolutional network framework for tourist flow prediction. ACM Trans Internet Technol (TOIT) 21(4):1–13

Wang Y, Guan L, Venetsanopoulos AN (2012) Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Trans Multimedia 14(3):597–607

Wu CH, Lin JC, Wei WL (2014) Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. In: APSIPA transactions on signal and information processing, p 3

Wu M, Su W, Chen L et al (2020) Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition. IEEE Trans Affect Comput

Wu Z, Cao Z (2005) Improved mfcc-based feature for robust speaker identification. Tsinghua Sci Technol 10(2):158–161

Yan J, Lu G, Bai X et al (2018) A novel supervised bimodal emotion recognition approach based on facial expression and body gesture. IEICE Trans Fundam Electron Commun Comput Sci 101(11):2003–2006

Zhalehpour S, Onder O, Akhtar Z et al (2016) Baum-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans Affect Comput 8(3):300–313

Zhang S, Zhang S, Huang T et al (2017) Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Trans Circ Syst Video Technol 28(10):3030–3043

Zhang S, Chen A, Guo W et al (2020) Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition. IEEE Access 8:23496–23505

Zhang WQ, He L, Deng Y et al (2010) Time-frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition. IEEE Trans Audio Speech Lang Process 19(2):266–276