International Journal of Mathematical, Engineering and Management Sciences
Vol. 7, No. 1, 49-67, 2022
https://doi.org/10.33889/IJMEMS.2022.7.1.004

A Novel S-LDA Features for Automatic Emotion Recognition from Speech using 1-D CNN

Pradeep Tiwari
Department of Electronics Engineering,
Sardar Vallabhbhai National Institute of Technology, Surat, Gujrat, India.
Department of Electronics and Telecommunication Engineering,
Mukesh Patel School of Technology Management and Engineering, NMIMS University, Mumbai, India.
Corresponding author: pradeep.tiwari@nmims.edu

A. D. Darji
Department of Electronics Engineering,
Sardar Vallabhbhai National Institute of Technology, Surat, India.
E-mail: add@eced.svnit.ac.in

(Received on May 21, 2021; Accepted on December 18, 2021)

Abstract
Emotions are explicit and serious mental activities, which find expression in speech, body gestures and facial features, etc. Speech is a fast, effective and the most convenient mode of human communication. Hence, speech has become the most researched modality in Automatic Emotion Recognition (AER). To extract the most discriminative and robust features from speech for Automatic Emotion Recognition (AER) recognition has yet remained a challenge. This paper, proposes a new algorithm named Shifted Linear Discriminant Analysis (S-LDA) to extract modified features from static low-level features like Mel-Frequency Cepstral Coefficients (MFCC) and Pitch. Further 1-D Convolution Neural Network (CNN) was applied to these modified features for extracting high-level features for AER. The performance evaluation of classification task for the proposed techniques has been carried out on the three standard databases: Berlin EMO-DB emotional speech database, Surrey Audio-Visual Expressed Emotion (SAVEE) database and eNTERFACE database. The proposed technique has shown to outperform the results obtained using state of the art techniques. The results shows that the best accuracy obtained for AER using the eNTERFACE database is 86.41%, on the Berlin database is 99.59% and with SAVEE database is 99.57%.

Keywords- Emotion recognition, LDA, MFCC, 1D-CNN, LDA.

1. Introduction

Emotion recognition includes analyzing an individual's facial expressions, non-verbal communication, or speech signals and grouping them as a particular emotion. It has stated that emotion recognition is critical for regular living and is fundamental while interacting with others (Chavhan et al., 2015). The event of advancement towards human and machine communication, for example, the interaction of a social robot with the human, the machine ought to have the option to perceive and react to the emotional state of the client (Cameron et al., 2015). Further, the medical fields like psychiatry and mental illness which deals with the understanding of the negative emotions of an individual are the recent applications of Emotion Recognition. Speech being a non-invasive and non-intrusive modality has attracted the researchers for determining AER from speech. The emotional state of the speaker could be recognized by extracting the paralinguistic information from the speech, which is speaker-independent and does not contain linguistic information. Speech signal varies under different emotions or stressed conditions as in (Hansen and Bou-Ghazale, 1995; Hansen and Womack, 1996; Ramamohan and Dandapat, 2006). Mental stress which is a common issue worldwide can be seen in human speech attributes like vocal jitter and

glottal flow spectrum (Ozdas et al., 2004; Vandyke, 2013). Stress can be known through seven universal emotions from humans like anger, fear, happy, disgust, contempt, surprise, and sad (Ekman and Friesen, 1978). Much research is going on in recognizing different emotions from speech modality, and over the last few decades, the Human-machine interface has significantly contributed to the field of medical assistance and psychiatry.

1.1 Motivation

The motivation behind this research work is to develop a system which can make humans comprehend their emotion level. Beside the tremendous development of psychiatry and medical science, stress has grown over years and poses as a modern epidemic to mankind. Negative emotions perceived as stress is a common issue worldwide that not only limits individual's capabilities, interest levels and mood but also causes physical and mental health problems. Modern youths are also facing the bane by this disease called stress. If a human can comprehend their mental status or stress level by an AER system, they can take an appropriate measure by consulting with a doctor to heal it before it causes great harm physically and psychologically. Such a technology would be extremely valuable especially during the times of a pandemic where social distancing is a must. Further, most of work done earlier by researchers are based on MFCC and pitch related features. These features are directly evaluated from static speech frames; thus, they do not consider the temporal varying information. Since, speech is quasi-stationary, shifted delta coefficient which considers all possible temporal varying information, may aid the performance of AER. The major contribution of the proposed work here, is to utilize the S-LDA feature along with 1D-CNN to enhance the accuracy for emotion recognition. Thus, the proposed algorithm considers the stationary (Pitch +MFCC), non-stationary (S-LDA) and high level (1D-CNN) features for achieving higher accuracy.

In AER, there are primarily two challenges, first is to identify the most effective interclass distinguishable features and second is to develop a robust classifier model (Li and Huang, 2014). The challenge in designing a speech-based AER system requires extracting features that uniquely distinguish different emotions and at the same time shows the least dependency on the speaker and the lexical content. As far as feature extraction from speech is concerned, various emotion dependent features have been extracted as in Akçay and Oğuz (2020) and Jassim et al. (2017) for AER applications. These features can be divided into three categories namely voice quality, prosodic, and spectrum-based features (Jiang et al., 2006; Busso et al., 2009; Chen et al., 2012). The voice quality, prosodic, and spectrum-based features are called as low-level hand-crafted features while the CNN based features are called as High-level Features. The key principle of deep learning networks like CNN is its capability to automatically learn the hidden features. To solve the problem of handcrafted feature extraction techniques that require features to be processed manually, CNN has a hierarchical network to extract the hierarchical information. The hierarchical model of CNN uses the derived S-LDA features as base features and extracted high-level features layer after layer from base features through the convolution process, max-pooling process, activation function modeling, and much more. Compared to conventional, 1D fully-connected neural networks, 1D-CNN can learn more robust features and improve its classification results. Various techniques are employed to modify the low-level features and obtain improved discriminative features (Akçay and Oğuz, 2020). The differential (or delta) features of MFCC provides significant cues for speech emotion recognition (Noroozi et al., 2017). Zhao et al. (2019) presented deep learning, 1D and 2D CNN-LSTM networks for extracting high-level features from speech. They also concluded that the speech log Mel spectrogram features with 1-D CNN gave gives better results than the 1-D audio sequences applied directly to 1-D CNN. Hence, the present

this paper proposes a new 'S-LDA' algorithm to calculate the modified differential or derived features from low-level features i.e., MFCC and pitch. The emotion-related information obtained using MFCC and pitch is calculated on consecutive speech frames gave static information of that particular frame. However, there may be dynamic information in the temporal domain which can be found by delta MFCC coefficients and pitch in the temporal domain. To get all possible derive all delta features and stack them one behind the others was achieved by using Shifted Delta Coefficients (SDC) algorithms. However, the stacking of the delta features would increase the feature dimension. Since there are multiple emotional classes, the increase in the feature dimension may decrease the AER recognition rate due to the overfitting problem. Thus, to combine the LDA with the SDC algorithm for feature dimensionality reduction would overcome overfitting issues and increase the AER performance. This combination of SDC with LDA called 'S-LDA' provides a novel approach to obtain modified differential features. Further 1-D CNN was applied to these derived features for extracting high-level features. The performance evaluation of the proposed techniques has been carried out on three standard databases: Berlin EMO-DB emotional speech database, SAVEE database, and eNTERFACE database. The proposed technique has shown to outperform the results obtained using state of the art techniques. The results shows that the best accuracy obtained for AER is 86.41% for eNTERFACE database (Martin et al., 2006), 99.59% for Berlin database (Burkhardt et al., 2005) and 99.57% for SAVEE database (Jackson and Haq, 2014).

The remaining paper is organized as follows. Literature review is discussed in section 2. Section 3 give details about the experiments conducted, while the results and its analysis are discussed in section 4. Section 5 represents conclusion and future scope of this work.

2. Literature Review

To implement a robust AER system, it is essential to have compelling interclass distinguishable features. Among the voice quality, prosodic, and spectrum-based features, spectrum features have gained more acceptance for AER. Nevertheless, the recent trend is to combine the above three feature categories to improve performance. El Ayadi et al. describes that pitch related features, formants, and energy-related features contribute to speech emotion recognition (El Ayadi et al., 2011). Ramamohan and Dandapat have used Sinusoidal features and concluded that the sinusoidal features are better in comparison to cepstral and linear prediction features (Ramamohan and Dandapat, 2006). AER was obtained on the Berlin EMO-DB speech emotion database by using the first 120 Fourier coefficients (Wang et al., 2015). Haq et al. (2015) has proposed feature selection algorithm for SAVEE database on audio features in the speaker-dependent scenario. Further, Gharavian et al. (2017) had used Fast Correlation-Based Filter algorithm for feature selection on spectral features like MFCC, Formants, and corresponding statistical features with fuzzy ARTMAP neural networks on SAVEE database. Significant improvement in AER was observed in Bozkurt et al. (2011) when weighted MFCC features were joint with spectral in addition to prosody features. Deb and Dandapat, proposed region switching technique for AER to switch between vowel-like regions and non-vowel like regions and gave an average accuracy of 85.1% on EMO-DB database by employing Extreme Learning Machine (ELM) classifier (Deb and Dandapat, 2017). Wissam et al. had combined neurogram features and traditional features to develop SVM model (Jassim et al., 2017). The best accuracy obtained in Jassim et al. (2017) on the eNTERFACE database was 77.27% and the EMO-DB berlin database was 84.68%. Sun et al. (2019a) had proposed AER based on decision tree SVM and fisher feature selection. Kerkeni et al. (2019) researchers have used the Empirical Mode Decomposition approach to enhance AER performance on Spanish and EMO-DB berlin databases. The hierarchical sparse coding framework was adopted in Torres-Boza et al. (2018) to extract features from speech automatically. Emotion-discriminative and domain-invariant

feature learning method were proposed in Mao et al. (2017) to learn the speech features by using emotion and domain label constraint. Experimental results based on a deep neural network - decision tree SVM model as proposed in Sun et al. (2019b) shows higher accuracy as compared to traditional SVM method.

The work done by other researchers presented in this section describes the voice quality, prosodic, and spectrum-based features which are known as low-level handcrafted features. From various low-level features used in literature previously, the MFCCs and Pitch was considered as starting features in the presented work here. These low-level features are insufficient in recognizing the emotions from speech. In addition, the above presented work by other researchers have used static features. However, the dynamic or differential features may be utilized for improving AER accuracy. Thus, this paper proposes a new 'S-LDA' algorithm to calculate the modified differential or derived features. Further 1-D CNN was applied to these derived features for extracting high-level features and classification.

3. AER System Set-Up

This section discusses the proposed AER system and the backgrounds of the AER system. The block diagram of the proposed AER system is shown in Figure 1 and has been explained in the following four steps. The speech signal was given as the input to the AER system. After preprocessing, the MFCC and pitch feature extraction algorithm was applied to get low-level features. Further, the proposed S-LDA algorithm has been used on low-level features to get modified and more discriminative, differential feature vectors. In the next step, the 1-D CNN model is employed to extract high-level features from differential features. Further, these high-level features were used for the classification of emotions using 1-D CNN. The description of these steps is further discussed in detail.

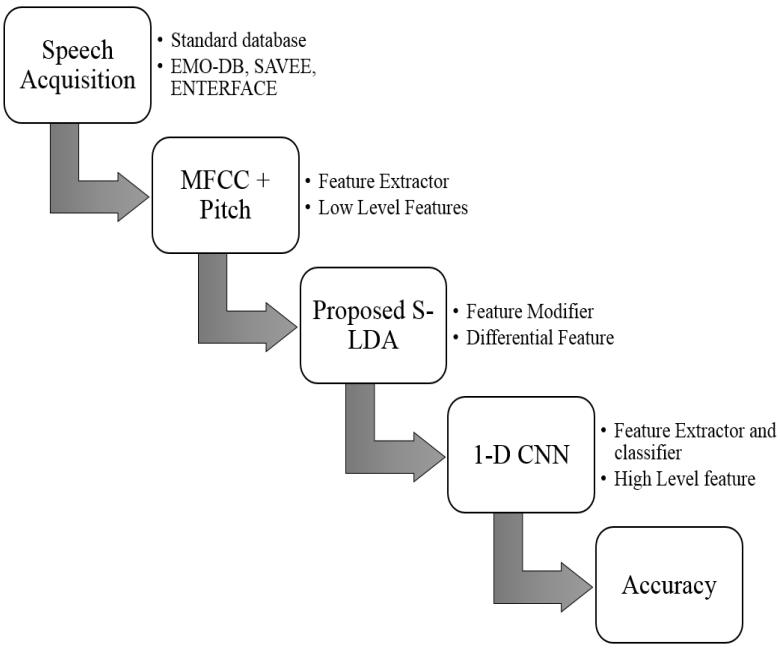

Figure 1. Proposed AER system.

3.1 Speech Signal Acquisition and Feature Extraction

The first step in AER is to acquire speech signal, which was taken from the three standard audio-visual databases: Berlin EMO-DB (Burkhardt et al., 2005), SAVEE (Jackson and Haq, 2014) and eNTERFACE (Martin et al., 2006) database. The second step in AER is to calculate the emotion related feature from the acquired speech signal. The accuracy of the AER system is generally governed by the features taken out from the acquired speech signal. The handcraft features which were taken as a base feature for modification in this work are (i) Pitch frequency and (ii) MFCC.

Pitch frequency or pitch (Noroozi et al., 2017) can be extracted in different ways such as time domain, frequency domain, and through the statistical techniques. The pitch frequency $P_0(s)$ of a speech signal 's', can be calculated from the mathematical formulation mentioned in equation (1).

$$P_0(s) = FT\{log|FT(s.w_n^H \parallel s \parallel)|\} \quad (1)$$

Mel frequency cepstral coefficients (MFCC) is extensively used speech feature extraction algorithm which is achieved by multiplying Mel filter bank with the Power spectrum of speech signal (Narayan and Alwan, 2000; Davis and Mermelstein, 1980). The block diagram for MFCC feature extraction is as shown in Figure 2.

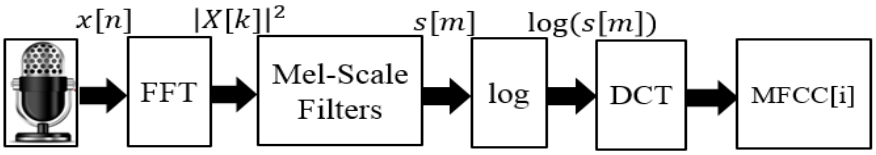

Figure 2. MFCC feature extraction.

To extract perception-based features, 128 triangular Mel filter banks are used. The short-time Fourier transform $X[k]$ with discrete frequency instances 'k' is obtained from the input signal $x[n]$ for discrete time instances 'n' for a frame of length N. Now, $X[k]^2$ which is called the Power spectrum is calculated by squaring the absolute value of fast Fourier transform of x[n]. If $X[k]^2$ is passed through Mel frequency filter bank $H_m[k]$ consists of 128 triangular filters, Mel Scaled power spectrum $S[m]$ is obtained using equation (2).

$$S[m] = \sum_{k=0}^{N-1} X[k]^2 \times H_m[k], \qquad 0 \le m \le M \quad (2)$$

MFCC is the log of Mel-scaled power spectrum $S[m]$ which is transformed again to the time domain by using a discrete cosine transform. The MFCC from $S[m]$ is calculated from equation (3).

$$MFCC[i] = \sum_{m=1}^{M} log(S[m]) \times \cos\left[i\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right], \qquad i = 1,2,\ldots L \quad (3)$$

In equation (3), 'L' indicates the number of MFCC coefficients calculated for each frame whereas 'M' represents the speech frames length. In the present work, AER 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 AER.

3.2 Proposed S-LDA Algorithm

The emotion-related information obtained using MFCC and pitch is called as low level or handcrafted features. These features were calculated for consecutive extracted frames to extract the static information of that particular frame. However, there may be dynamic information in the temporal domain which can be found by differentiating MFCC coefficients and pitch in the temporal domain. The first and second-order derivatives of MFCC coefficients called as delta MFCC and delta-delta MFCC respectively, are generally used to collect differential MFCC feature vectors. Torres-Carrasquillos proposed Shifted Delta Coefficients (SDC) which became the most 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 (Zhang et al., 2010). Thus, to get all possible differential features, SDC algorithms can play a very important role. SDC feature vectors are created by stacking delta MFCC (△ MFCC) computed across multiple speech frames of a speech sample (Wang et al., 2012). The △ MFCC known as differential coefficients extracts the higher-order variation in MFCC feature as represented in equation (4) (Noroozi et al., 2017). The MFCC feature vector conveys only the power spectral envelope of a single speech frame, but the nature of speech has information in the dynamics as well. The trajectories of the MFCC coefficients over time also carries speech information.

$$\triangle c(t) = \frac{\sum_{m=1}^{N} m(c(t+m)-c(t-m)}{2\sum_{1}^{N} m^2} \quad (4)$$

In equation (4), △ MFCC coefficient, $\triangle c(t)$ is obtained from MFCC frame $c(t+m)$ and $c(t-m)$ which is static coefficient, "t" is the analysis frame's index and "m" represent the amount of shift for delta computation. SDC coefficients are based on four parameters namely written as N-d-P-k where 'N' represents the number of MFCC coefficients i.e., the number of the column of MFCC. The amount of shift for delta computation is represented by 'd', 'P' represents the amount of shift for next frame whose deltas are to be computed and 'k' represents the number of a frame whose deltas are to be stacked. The value of N-d-P-k used for computing SDC is N=40, d=1, P=3, k=4, 8, 12. Researchers have found that when N=7, d=1, P=3, k=7, the SDC algorithm gives better results (Matejka et al., 2006). Since the number of MFCC coefficients were 40, N was kept 40. The values of d and P were not changed. Further the results of the AER performance were analyzed for different values of K such as k=4, 8, 12. The increased values of k gave better results. For a given time t, $\triangle c(t,i) = c(t+iP+d) - c(t+iP-d)$ for $i = 0,1,2,\ldots k-1$. Figure 3 shows the computation of S-LDA features. For calculating $\triangle c(t, i=0)$ keeping $d=1$, the difference of $c(t+1)$ and $(t-1)$ is obtained where $\triangle c(t,0)$ indicates that the difference between MFCC of 2^{nd} frame and MFCC of 0^{th} frame. Then with a shift of 'P' frame $\triangle c(t,1)$ is calculated. Similarly, till $\triangle c(t,7)$ is computed. Finally, for $k=8$, $\triangle c(t,i)$ for $i = 0,1,2\ldots7$ is stacked as shown in equation (5) (Wang et al., 2012). Thus total $N \times k$ i.e., 320 features are generated by SDC.

$$SDC(t) = [\triangle c(t,0) \quad \triangle c(t,1) \quad \triangle c(t,k-1)] \quad (5)$$

SDC provides all possible temporal, differentiating features representatives for a speech sample, however, the stacking of the delta features would increase the feature dimension. Since there are multiple emotional classes, the increase in the feature dimension may decrease the AER recognition rate due to the overfitting problem. Thus, in this paper, it is proposed to combine the LDA with Shifted Delta Coefficient (SDC) 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 equation (6) represents the

between-class scatter and S_W given in equation (7) represents the with-in class scatter.

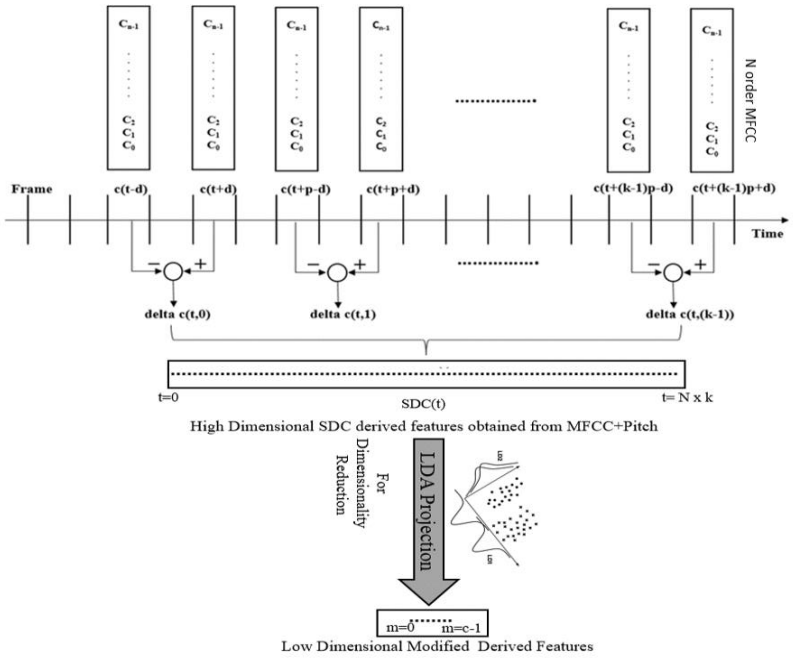

Figure 3. Proposed S-LDA algorithm.

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \quad (6)$$

In equation (6) and (7), '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).

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \quad (7)$$

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.

$$W_{opt} = argmax_W \frac{|W^T S_B W|}{|W^T S_W W|} \quad (8)$$

$\{w_i | i = 1,2,\ldots\ldots, w_m\}$ 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 into dimensionality reduction.

Thus, by combining SDC and LDA, a robust, discriminative feature vector can be obtained. The extracted modified features now need to be further applied to 1-D CNN for high-level feature extraction and classification.

3.3 Feature Importance Analysis of the S-LDA Feature

To analyze the importance and strength of the proposed S-LDA feature, we have evaluated the feature importance score from SAVEE database as shown in Figure 4 using Extra-Tree algorithm

(Pedregosa et al., 2011; Alsariera et al., 2020). Since, SAVEE database has 7 emotional classes, the S-LDA algorithm produces 6 dimensional features (F1, F2, F3, F4, F5, F6). The score depicts that F1 feature is more important than F2, F2 is more important than F3 and so on.

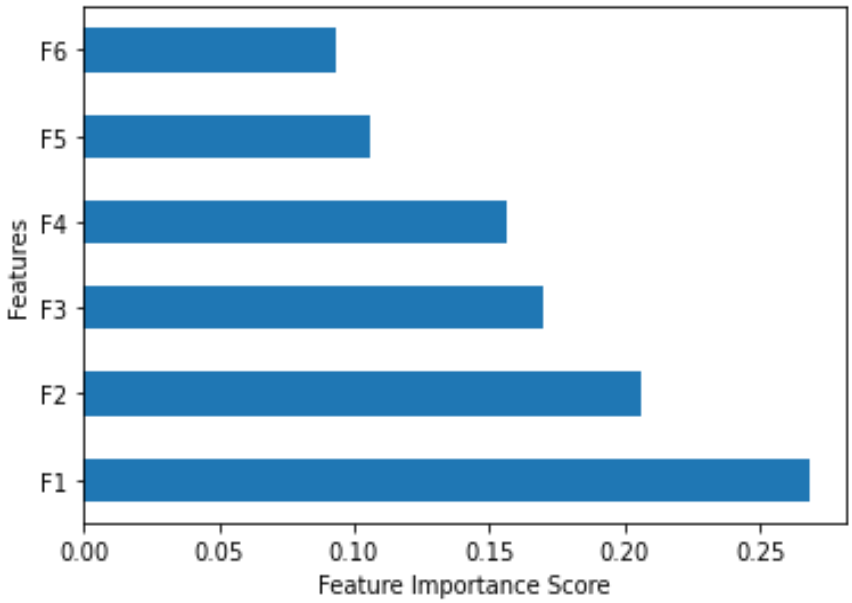

Figure 4. Feature importance score evaluation.

Further, the box-plot is obtained as shown in Figure 5 for F1-F6 features comparing the first quartile Q1, second quartile i.e. median Q2 and third quartile i.e. Q3 values of all 7 emotion classes. Figure 5(a) shows Box plot of F1 feature of all 7 emotion classes of SAVEE database. Similarly, Figure 5(b) represents F2 and so on.

The box plot indicates that the interclass difference in Q1, Q2 and Q3 values in most of the features i.e. F1-F6, however the Q2 values are close in F1, F3, F5 and F6 for neutral and sad emotions but F2 and F4 has given clear difference which would complementarily support the classifier.

A comparison between the T-SNE plot is given in Figure 6 to illustrate the strength of S-LDA algorithm over base features. Figure 6(a) represents the MFCC+Pitch feature T-SNE plot and the data points of different classes represented by labels (0.0 -Anger, 1.0- Disgust, 2.0-Fear, 3.0-Happy, 4.0-Neutral, 5.0-Sad, 6.0-Surprise) are clearly seen overlapping each other. Figure 6(b) represents the T-SNE plot of MFCC+Pitch + S-LDA(k=4) has lesser overlap of data point of different emotion classes while Figure 6(c) showing T-SNE plot of MFCC+Pitch + S-LDA (k=8) has least overlap and clear separation between the 7 emotion classes of the SAVEE database. This would increase the performance of AER in all the three databases considered.

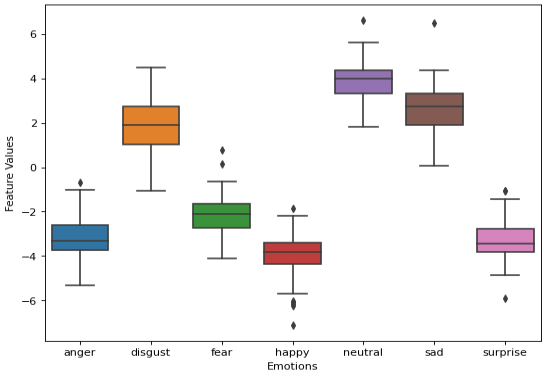

Figure 5(a) F1

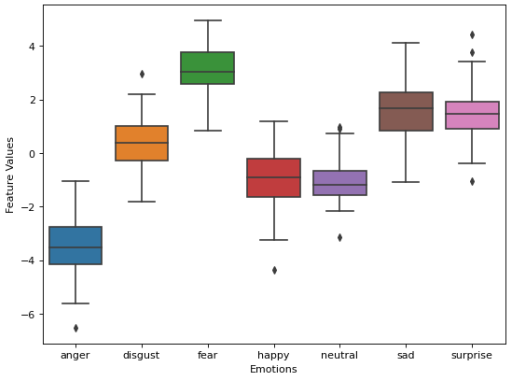

Figure 5(b) F2

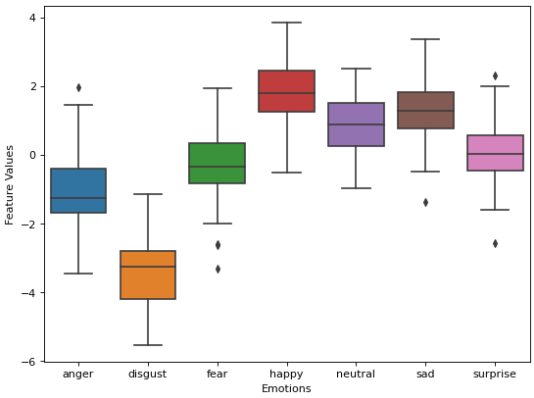

Figure 5(c) F3

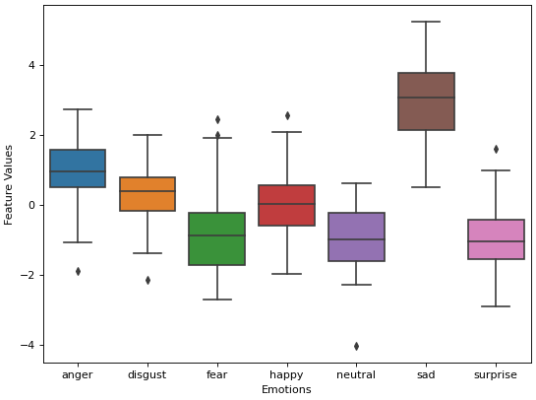

Figure 5(d) F4

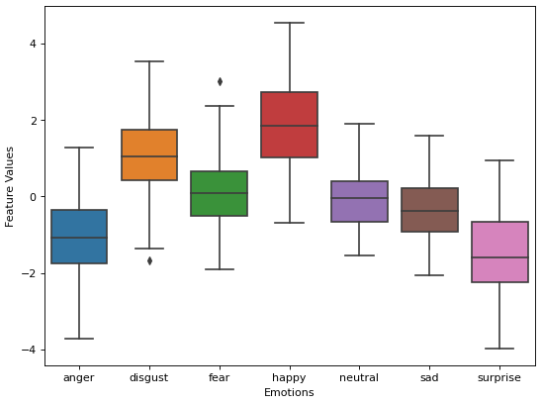

Figure 5(e) F5

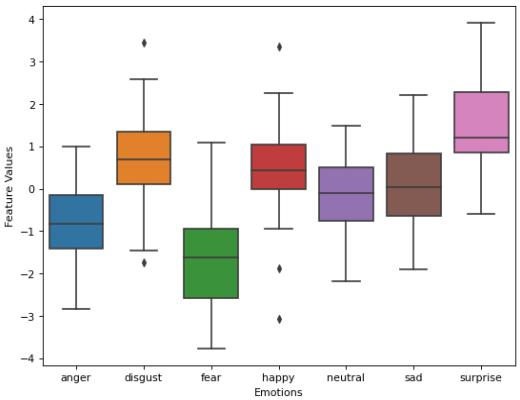

Figure 5(f) F6

Figure 5. Box-plot of S-LDA features for 7 emotions.

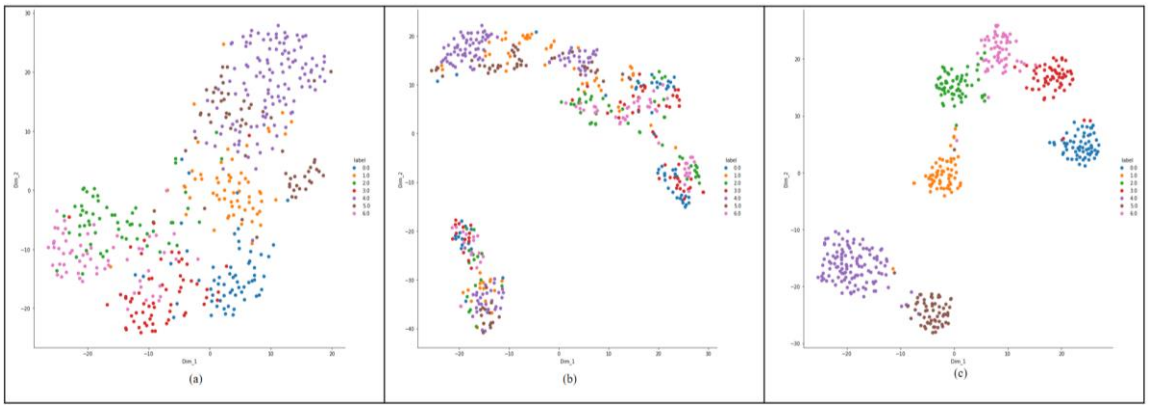

Figure 6. T-SNE plot of (a) MFCC+pitch (b) MFCC+pitch + S-LDA(k=4) (c) MFCC+pitch + S-LDA (k=8).

3.4 1-D CNN

CNNs are fundamentally one of the active models of Deep Neural Networks (DNN) (Husain et al., 2017). CNN essentially diminishes the network parameters by local connectivity and weights sharing utilizing convolutional layers. The main building blocks comprises a set of filters or kernels with a small receptive field. Each kernel moves over the input in a predetermined way performing the convolution operation. However, the kernel parameters don't vary to control the complete set of free parameters. These filters would help in extracting high-level features from features obtained by the S-LDA algorithm. Equation (9) indicates the computation of feature map $X_k^{(l)}$ in the l^{th}layer of a convolutional layer.

$$X_k^{(l)} = f\left(\sum_c W_k^{(l),c} * X^{(l-1),c} + B_k^l\right) \tag{9}$$

In equation (9), 'k' indicates the kernel number, the channel number of inputs $X^{(l-1)}$ is 'c', the k^{th} convolutional kernel is $W_k^{(l),c}$, the activation function is $f(.)$ and '$*$' represents the element wise multiplication. Parameters are further reduced by the insertion of pooling layers between consecutive layers. By selecting a maximum value in each kernel, subsampling functions are max-pooled. With multiple convolution layers, pooling layers, and fully connected layers, high-level feature extraction and classification can be achieved.

4. Results and Discussion

In this paper, three standard databases: eNTERFACE (Martin et al., 2006), SAVEE (Jackson and Haq, 2014), and Berlin EMO-DB (Burkhardt et al., 2005) database were used to evaluate the performance of the proposed method for AER. Python was used as a development tool on the system with 4GB RAM. From the databases considered for experimentation, the samples were randomly divided into two parts in the ratio of 3:1, 75% of the samples were considered for training, and 25% for testing purpose. The experimentation was repeated for 500 epochs and the accuracy was recorded. The final result would be the result obtained at 500^{th} epochs. The following two approaches for AER have been implemented and their accuracy was measured computed.

1. AER using Pitch+ MFCC + 1D-CNN: Pitch and MFCC features were taken as base features and high-level features were extracted on them using 1D-CNN. The further classification was also

done by using a softmax classifier in 1D-CNN.

2. AER using Pitch+ MFCC+ S-LDA +1D-CNN: Pitch and MFCC was taken as a base feature and modified derived features were calculated using the proposed S-LDA algorithm. Further, by applying 1D-CNN, high-level features were extracted on them. Furthermore, classification was also done by softmax classifier in 1D-CNN.

The implemented architecture of CNN for high-level feature extraction and classification is shown in Figure 7. The first layer in CNN architecture was the convolution layer which consists of 128 filters having dimension 5×1. Activation layer ReLU was added to convert the output of the convolution layer into nonlinear features. To reduce the overfitting a dropout of 0.1 was introduced. Max pooling of 4×1 was applied further for downsampling the non-linear features. The process was repeated but the max-pooling layer was not included. Lastly, the output of convolutional layer-2 was flattened to create a 1-D feature vector. Before applying the softmax classifier of 1D-CNN at the dense layer, the number of features obtained for 'AER using Pitch+ MFCC + 1D-CNN' was 640 and number of features obtained for 'AER using Pitch+ MFCC+ S-LDA +1D-CNN' was 128. The dense layer was then added to combine and converge the learned features into the number of emotion classes to be classified. The activation layer, Softmax then determined the output of the network. The Root Mean Square (RMS) optimizer further lead the learning curve. The work done on three databases are now discussed in detail.

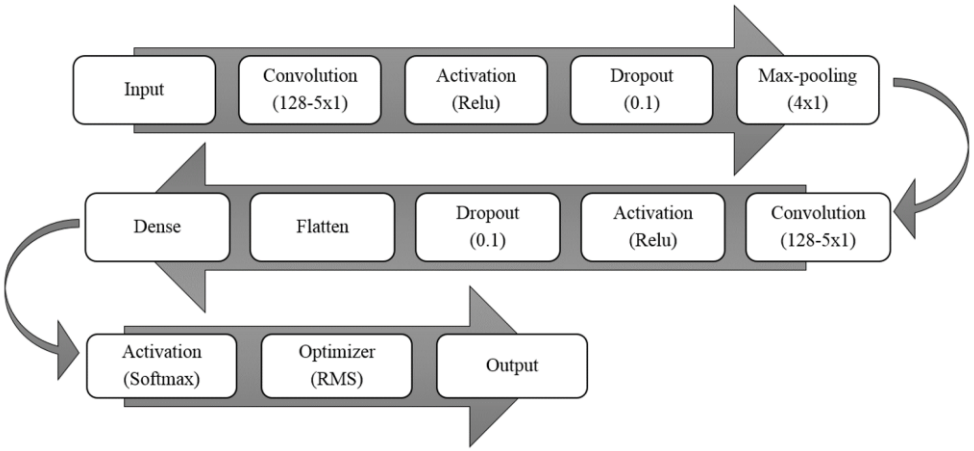

Figure 7. CNN Architecture for high-level feature extraction and classification.

4.1 AER using eNTERFACE Database

The eNTERFACE dataset is a standard emotional dataset that may be utilized for AER (Martin et al., 2006). It has considered 42 people speaking English language and acting out six emotions: anger, disgust, fear, happiness, sadness, and surprise. There are 5 video clips for each emotion, resulting in a total of $(42 \times 6 \times 5)$ i.e. 1260, videos. A sample rate of 48 kHz is used in the voice files retrieved from video files. The inter-class discriminative feature extracted from speech can describe emotion. For extracting the features from speech samples, pre-processing is done as explained in section 3.1. The recorded speech signals in the database are quasi-periodic and quasi-stationary. To obtain periodic and stationary speech signals from these recorded speech signals, frames of 2048 points are extracted along with the frameshift of 512 points.

In case I of Table 1, since eNTERFACE database has 1260 samples, with 40 MFCC features and 1 pitch feature, a training feature set of 945×41 was obtained. This feature set was the input to the CNN model, which is shown in Figure 8. The performance achieved for this feature set (case I of Table 1) after 500 epochs were 41.23%. The AER performance considering the 41-dimensional MFCC and pitch features is shown in Figure 8. The difference between the train and test curve shows that there is overfitting in the model. As the number of subjects and the size of the training feature set was increased, the performance of AER for eNTERFACE has decreased due to overfitting. Now, to improve the accuracy of AER, the proposed S-LDA feature was applied with 1-D CNN classifier. Three cases for AER have been carried on the eNTERFACE database. For the case II (a) of Table 1, the number of frames whose delta were stacked was k=4. By using S-LDA, the accuracy has been improved by 22.63% thus raising to 63.08%. Further, for the II(b) case of Table 2 with k=8 frame whose delta were stacked was considered. After applying S-LDA with SDC (k=8) the accuracy of 78.73% was obtained which marks an improvement of approximately 36.62%. Finally, when the II(c) case of SDC with k=12 as given in Table 1 was considered, after applying LDA with SDC(k=12) and MFCC and pitch feature the accuracy of 86.41% was obtained which marks the best accuracy for eNTERFACE with an improvement of approximately 45.38%. The results of AER for S-LDA with k=12 is shown in Figure 9. The difference in the linear curve between training and testing accuracy has decreased considerably, indicating that the S-LDA features performed better than MFCC and pitch features.

The Table 1 tabulates the average results of AER using eNTERFACE database. However, the Table 1 also indicates the individual accuracies computed for each of the 7 emotions considered here for experimentation purpose.

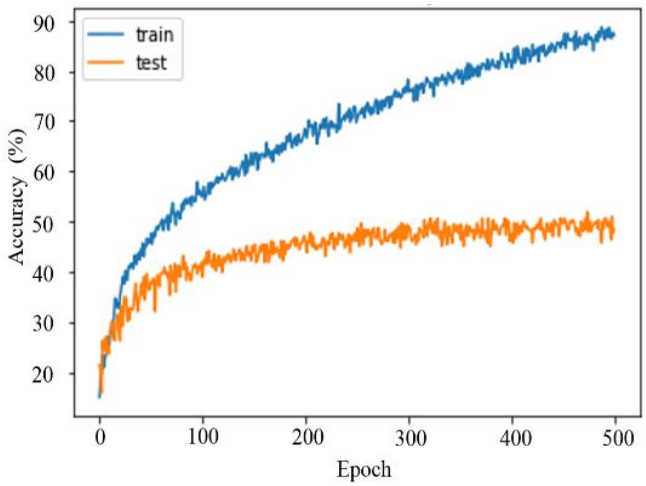

Figure 8. The train and test accuracy of 1-D CNN applying MFCC+pitch features (without S-LDA) eNTERFACE database.

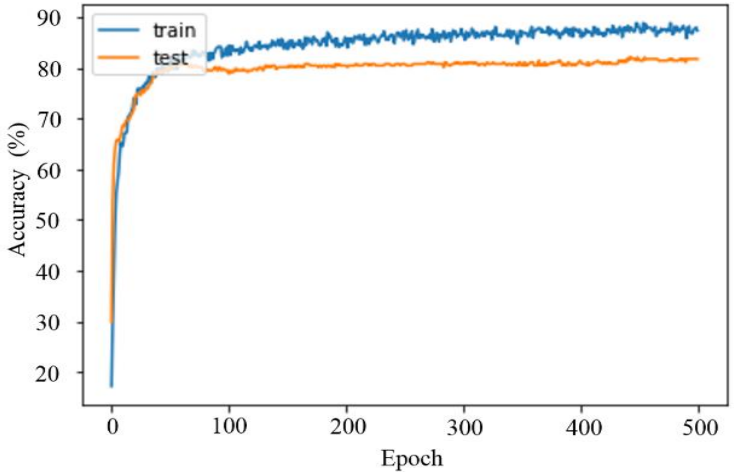

Figure 9. The train and test accuracy of 1-DCNN by applying MFCC+pitch features with proposed on S-LDA algorithm on eNTERFACE database.

Table 1. AER for eNTERFACE database using 1D-CNN classifier.

Case	Features	A	D	F	H	SA	SU	Avg. Accuracy
I	MFCC+Pitch	61.94	34.57	29.63	35.18	52.12	33.94	41.23
I(a)	MFCC+Pitch +S-LDA (k=4)	74.89	59.65	51.45	55.98	75.11	61.45	63.08
I(b)	MFCC+Pitch+S-LDA (k=8)	86.35	77.17	73.04	73.97	84.62	77.24	78.73
I(c)	MFCC+Pitch+S-LDA (k=12)	90.18	86.95	80.21	89.66	88.58	82.92	**86.41**

4.2 AER using Berlin EMO-DB Database

The Berlin database is a standard audio emotion database, which has been used here for evaluating the performance of AER algorithm (Burkhardt et al., 2005). It contains seven emotions: Anger, Boredom, Disgust, Fear, Happy, Sad and Neutral. The speech utterances were collected from ten actors comprising five females and five males. The sampling frequency of the recorded samples is 48 kHz. The total number of speech samples in the database is 535. In case I of Table 2, 40 MFCC features and 1 pitch feature were considered. Since, berlin database has 535 samples, a training feature set of 401×41 was obtained. The performance of AER achieved was 71.41% by using MFCC and pitch features. Further, the proposed S-LDA has applied to MFCC + Pitch features. S-LDA includes SDC features that are created by stacking delta MFCC features computed across multiple speech frames that carry dynamic temporal features. Two cases for AER have been done using S-LDA features and a 1D-CNN classifier.

For the case II (a) of Table 2, the number of frames whose delta were stacked was k=4. Hence, the feature vector length became 202 at the output of the SDC algorithm. Further, it was observed that the addition of features has increased the computation time and complexity. The linear transformation LDA transforms the high dimensional feature vector into a small feature dimension. The LDA techniques maximize the interclass separability by identifying a sub-space with a suitable direction. Since the number of emotion classes in the berlin database is seven, LDA reduces the feature dimension to six and increases the inter-class separability. For the MFCC feature (40) and pitch feature (1), the dimension of the training feature set becomes 401×6. Thus, by applying MFCC, pitch, and S-LDA (with k=4) features, the AER accuracy of 94.63% has been obtained which is almost 23% more than that observed for the same features before applying S-LDA. Further in case II (b) of Table 2, with S-LDA (k=8) applied on MFCC and pitch features, the accuracy of

99.59% was obtained which marks an improvement of approximately 28%. The emotion related information obtained using MFCC and pitch was calculated on consecutive speech frames which gave static information of that particular frame. Then, all possible dynamic information was calculated from MFCC and pitch, and stacked by applying Shifted Delta Coefficients (SDC) algorithms. However, the stacking of the delta features would increase the feature dimension. Since there are multiple emotional classes, the increase in the feature dimension increases the AER recognition rate due to overfitting problem. Thus, the LDA was combined with SDC algorithm for feature dimensionality reduction to overcome the overfitting issue and increase the AER performance. This combination of SDC with LDA called 'S-LDA' thus improved the AER performance using 1D-CNN classifier up to 99.59%.

Table 2. AER for EMO-DB database using 1D-CNN Classifier.

Case	Features	Anger	Boredom	Disgust	Fear	Happy	Sad	Neutral	Avg Accuracy
I	MFCC+ Pitch	85.73	71.43	73.56	60.81	60.02	86.3	62.07	71.41
II(a)	MFCC+Pitch+S-LDA(k=4)	98.43	90.55	92.98	94	92.38	98.33	95.78	94.63
II(b)	MFCC+Pitch+ S-LDA (k=8)	100	100	100	100	97.85	100	99.29	**99.59**

4.3 AER using SAVEE Database

The speech samples used in the analysis are taken from standard SAVEE database (Jackson and Haq, 2014). The database comprises of 480 British English utterances recorded from 4 male actors with 7 emotions (anger, disgust, fear, happy, neutral, sad and surprise). The sampling frequency of the recorded samples is 44100 Hz (16 bit). Initially in case I of Table 3, 40 MFCC features and 1 pitch feature were considered. Since, SAVEE database has 480 samples, a training feature set of 360×41 was obtained. A performance of AER achieved was 64.79%.

Table 3. AER for SAVEE database using 1D-CNN classifier.

Case	Features	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise	Avg Accuracy
I	MFCC+ Pitch	74.93	71.32	49.9	51.01	81.9	75.37	49.12	64.79
II(a)	MFCC+Pitch+ S-LDA (k=4)	93.83	94.54	88.14	93.54	95.42	87.47	83.95	90.98
II(b)	MFCC+Pitch+ S-LDA (k=8)	100	100	100	99.41	98.35	99.29	100	**99.57**

Among two cases of AER performed using S-LDA features, for the case-II(a) of Table 3 the number of frames whose delta were stacked was k=4. Hence, the feature vector length becomes 202, which is reduced to 6 and yields the accuracy of 90.98% which is almost 25% more than that observed for the same features before applying S-LDA. Further in case II(b) of Table 3, for k=8 frame whose delta were stacked, increased the length to 362 which after LDA becomes 6 gave the accuracy of 99.57% which marks an improvement of approximately 35%. The Table 3 tabulates the average results of AER using SAVEE database. However, the Table 3 also indicates the individual accuracies computed for each of the 7 emotions considered for the explanation. In the proposed AER system, features were extracted at three levels. First MFCCs and Pitch were selected as starting low-level handcrafted features. The next level of feature extraction was extracting the derived features using the proposed S-LDA algorithm on handcrafted features. Last level was high-level feature extraction using 1D-CNN. As it can be seen that eNTERFACE database which has 44 speakers is getting accuracy of 86.41% since the eNTERFACE database is difficult compare to other two databases i.e., EMO-DB with 10 speakers and SAVEE with 4 speakers giving approx. 99%. The results have shown that contribution of 1D-CNN is approximately 2-5% while S-LDA algorithm is approximately 20-40% considering different databases.

Further, cross data-set training and testing is also carried on as shown in the Table 4. The accuracy of cross dataset is low because the language of speech, number of speakers in the databases are different.

Table 4. AER performance for cross data-sets.

Training data-set	Testing data-set	Accuracy (%)
eNTERFACE	SAVEE	31.42
eNTERFACE	EMO-DB	41.78
EMO-DB	SAVEE	42.11
EMO-DB	eNTERFACE	28.32
SAVEE	eNTERFACE	24.68
SAVEE	EMO-DB	43.12

4.4 Comparison of the Performance of Proposed Work with the State-of-the-Art Methods

The performance of the proposed S-LDA feature extraction technique cannot be directly compared with the performance of various state-of-the-art techniques. The database used for performance comparison is kept the same while the features, number of features, training/testing vector subsets, and classifiers are variable. Nevertheless, it is useful and informative to analyze the results of the proposed and state-of-the-art techniques. Table 5, 6 and 7 shows the performance comparison of the proposed method for eNTERFACE, EMO-DB and SAVEE database respectively with state-of-the-art methods. The result shows that the best accuracy obtained for AER is 99.59% for the Berlin database which is better as compared to 99.57% for the SAVEE database, 86.41% for the eNTERFACE database using MFCC + SDC features and LDA feature selection.

Table 5. Performance comparison of present work with the state of the art methods for eNTERFACE database.

Literature	Feature	Classifier	Accuracy (%)
Noroozi et al. (2017)	MFCC, Prosody	SVM	41.32
Jassim et al. (2017)	Neurogram + Traditional features	SVM	77.27
Zhang et al. (2017)	Spectrogram	Anet	78.08
Proposed	Pitch + MFCC+ S-LDA (k=12)	1D CNN	**86.41**

Table 6. Performance comparison of present work with the state of the art methods for EMO-DB database.

Literature	Feature	Classifier	Accuracy (%)
Kerkeni et al. (2019)	EMD	SVM	86.22
Deb and Dandapat (2017)	MFCC + △ MFCC + △△ MFCC	ELM	85.1
Jassim et al. (2017)	Neurogram + Traditional features	SVM	84.68
Mao et al. (2014)	CNN	CNN	93.00
Meng et al. (2019)	3D log-mel Spectrogram	Dilated CNN	85.39
Proposed	Pitch + MFCC+ S- LDA (k=8)	1D CNN	**99.59**

Table 7. Performance comparison of present work with the state of the art methods for SAVEE database.

Literature	Feature	Classifier	Accuracy (%)
Noroozi et al. (2017)	MFCC, Prosody	SVM	48.81
Haq et al. (2015)	MFCC + energy + pitch	SVM	60.00
Gharavian et al. (2017)	MFCC, Formants	Fuzzy ARTMAP NN	53.00
Mao et al., (2014)	CNN	CNN	89.00
Proposed	Pitch + MFCC+ S-LDA (k=8)	1D CNN	**99.57**

The observations shows that the proposed method has higher AER rate in comparison with the state of the art techniques. The performance of AER for different databases is different because the language of speech, number of speakers in the databases are different. The number of subjects in SAVEE is '4' and the language is German, similarly, the number of subjects in EMO-DB is 10, and the language is German. These two databases have given a similar performance of approximately 99.5% for accuracy. The eNTERFACE database consists of 44 speakers and was recorded in English. Since the number of speakers is 44 which is too big than SAVEE and EMO-DB, it has shown the performance of 86.41% which is less in comparison with the other two databases. Most of the works with which the present work is compared as mentioned above belong to the low-level handcrafted features category. Noroozi et al. (2017), Haq et al. (2015) and Gharavian et al. (2017) have considered low level features like MFCC, Prosody, pitch, energy and formants. These low-level features are insufficient in distinguishing the different classes of emotions from speech and thus shows less accuracy of lower than 60%. Kerkeni et al. (2019) has used EMD features. Jassim et al. (2017) have considered neurogram features with traditional low-level features. Neurogram includes simulation of the response of an auditory-nerve fiber and a characteristic frequency to a speech signal is observed. Since the present work has considered dynamic features from MFCC, it gave better results than the above-mentioned work. Further, the above works have used the static features apart from (Deb and Dandapat, 2017) which has considered △△ MFCC differential features, however, the present work has considered all possible variation in the dynamic or differential features using SDC in S-LDA. The results clearly show that S-LDA has contributed to the improvement of AER accuracy. Finally, 1-D CNN has extracted high-level features and classified with increased accuracy. The Limitations of the proposed algorithm is that when number of subjects increases, the AER performance decreases. Also, the AER performance is low in cross dataset training and testing.

5. Conclusions

Human speech has received a lot of attention nowadays as a way to automatically detect accurate information about emotions. Normally AER were employed by using low-level handcrafted features. Further, these features were static and did not consider the dynamic or differential features. In this work, we have focused on the robust feature extraction techniques for AER from emotional speech. This paper has proposed and implemented a new algorithm S-LDA to extract differential features from low-level features such as MFCC and pitch. Further 1-D Convolution Neural Network (CNN) was applied on these derived features for extracting high-level features. The performance evaluation of the proposed techniques has been carried out on three standard databases: Berlin EMO-DB emotional speech database, SAVEE database and eNTERFACE database by employing 1D CNN classifier. The results show that the best accuracy obtained for AER is 99.59% for Berlin database, 99.57% for SAVEE database, 86.41% for eNTERFACE database.

This research found that differential features, in combination with high-level convolution features, provided adequate information for emotion classification. AER can be used in the field of medical assistance and psychiatry to identify the level of stress of a person. Since speech is non-invasive and nonintrusive, thus to acquire speech samples from a stressed person is painless and comfortable. In addition, the proposed speech-based AER work would be extended for audio-visual AER system.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments
We would like to thank the Department of Electronics Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, Gujrat, India and Department of Electronics and Telecommunication Engineering, Mukesh Patel School of Technology Management and Engineering, NMIMS University, Mumbai, India for providing the facilities to carry out the research work.

References

Akçay, M.B., & Oğuz, K. (2020). Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56-76.

Alsariera, Y.A., Adeyemo, V.E., Balogun, A.O., & Alazzawi, A.K. (2020). Ai meta-learners and extra-trees algorithm for the detection of phishing websites. IEEE Access, *8*, 142532-142542.

Bozkurt, E., Erzin, E., Erdem, C.E., & Erdem, A.T. (2011). Formant position based weighted spectral features for emotion recognition. *Speech Communication*, *53*(9-10), 1186-1197.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., & Weiss, B. (2005). A database of German emotional speech. In *2005 9th European Conference on Speech Communication and Technology Interspeech* (Vol. 5, pp. 1517-1520). Lisbon, Portugal.

Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 582-596.

Cameron, C.D., Lindquist, K.A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*, *19*(4), 371-394.

Chavhan, A., Chavan, S., Dahe, S., & Chibhade, S. (2015). A neural network approach for real time emotion recognition. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(3), 259-263.

Chen, L., Mao, X., Xue, Y., & Cheng, L.L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing*, *22*(6), 1154-1160.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357-366.

Deb, S., & Dandapat, S. (2017). Emotion classification using segmentation of vowel-like and non-vowel-like regions. *IEEE Transactions on Affective Computing*, *10*(3), 360-373.

Ekman, P., & Friesen, W.V. (1978). *Facial action coding system*. Palo Alto, CA: Consulting Psychologists Press.

El Ayadi, M., Kamel, M.S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572-587.

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 and Applications*, *76*(2), 2331-2352.

Hansen, J.H.L., & Bou-Ghazale, S.E. (1995). Robust speech recognition training via duration and spectral-based stress token generation. *IEEE Transactions on Speech and Audio Processing*, *3*(5), 415-421.

Hansen, J.H.L., & Womack, B.D. (1996). Feature analysis and neural network-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *4*(4), 307-313.

Haq, S., Jan, T., Jehangir, A., Asif, M., Ali, A., & Ahmad, N. (2015). Bimodal human emotion classification in the speaker-dependent scenario. *Pakistan Academy of Sciences*, *52*(1), 27-38.

Husain, F., Dellen, B., & Torras, C. (2017). Scene understanding using deep learning. In: Samui, P., Sekhar, S., Balas, V.E. (eds) *Handbook of Neural Computation.* Academic Press, pp. 373-382.

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

Jassim, W.A., Paramesran, R., & Harte, N. (2017). Speech emotion classification using combined neurogram and INTERSPEECH 2010 paralinguistic challenge features. *Institution of Engineering and Technology Signal Processing*, *11*(5), 587-595.

Ji, S., & Ye, J. (2008). Generalized linear discriminant analysis: a unified framework and efficient model selection. *IEEE Transactions on Neural Networks*, *19*(10), 1768-1782.

Jiang, X., Tian, L., & Cui, G. (2006). Statistical analysis of prosodic parameters and emotion recognition of multilingual speech. *Acta Acustica-Peking*, *31*(3), 217-221.

Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, M.A., & Cleder, C. (2019). Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Communication*, *114*, 22-35.

Li, Z., & Huang, C.W. (2014). Key technologies in practical speech emotion recognition. *Journal of Data Acquisition and Processing*, *29*(2), 157-170.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, *16*(8), 2203-2213.

Mao, Q., Xu, G., Xue, W., Gou, J., & Zhan, Y. (2017). Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. *Speech Communication*, *93*, 1-10.

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006, April). The eNTERFACE'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops* (pp. 8-8). IEEE. Atlanta, GA, USA.

Matejka, P., Burget, L., Schwarz, P., & Cernocky, J. (2006, June). Brno university of technology system for nist 2005 language recognition evaluation. In *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop* (pp. 1-7). IEEE. San Juan, PR, USA.

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access*, *7*, 125868-125881.

Narayanan, S., & Alwan, A. (2000). Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing*, *8*(3), 328-344.

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, *10*(1), 60-75.

Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K., & Wilkes, D.M. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, *51*(9), 1530-1540.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderpla, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Ramamohan, S., & Dandapat, S. (2006). Sinusoidal model-based analysis and classification of stressed speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(3), 737-746.

Sun, L., Fu, S., & Wang, F. (2019b). Decision tree SVM model with Fisher feature selection for speech emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, *2019*(1), 1-14.

Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019a). Speech emotion recognition based on DNN-decision tree SVM model. *Speech Communication*, *115*, 29-37.

Torres-Boza, D., Oveneke, M.C., Wang, F., Jiang, D., Verhelst, W., & Sahli, H. (2018). Hierarchical sparse coding framework for speech emotion recognition. *Speech Communication*, *99*, 80-89.

Torres-Carrasquillo, P.A., Reynolds, D.A., & Deller, J.R. (2002, May). Language identification using Gaussian mixture model tokenization. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. I-757). IEEE. Orlando, FL, USA.

Vandyke, D. (2013, September). Depression detection & emotion classification via data-driven glottal waveforms. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 642-647). IEEE. Geneva, Switzerland.

Wang, H., Leung, C.C., Lee, T., Ma, B., & Li, H. (2012). Shifted-delta mlp features for spoken language recognition. *IEEE Signal Processing Letters*, *20*(1), 15-18.

Wang, K., An, N., Li, B.N., Zhang, Y., & Li, L. (2015). Speech emotion recognition using Fourier parameters. *IEEE Transactions on Affective Computing*, *6*(1), 69-75.

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2017). Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030-3043.

Zhang, W.Q., He, L., Deng, Y., Liu, J., & Johnson, M.T. (2010). Time-frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(2), 266-276.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312-323.